التقرير الفني لمشروع :

إطار لتعدين وسائل الإعلام العربية واستخلاص المعلومات باستخدام التمثيل الدلالي المتصل للكلمات

Technical report of the project:

A Framework for Arabic Media Mining and Information Extraction Using Continuous Domain Semantic Representation

Table of Contents

1. Letter of Financial Support from the Deanship of Scientific Research

جامعة الملك سعود (0134)
هاتف +966 11 467 33 55
فاكس +966 11 467 33 66

المملكة العربية السعودية
ص.ب 2454 الرياض 11451
www.ksu.edu.sa

جامعة الملك سعود
King Saud University

عمادة البحث العلمي
مكتب العميد

إفادة لمن يهمه الأمر

تفيد عمادة البحث العلمي في جامعة الملك سعود بأن سعادة الأستاذ الدكتور/ منصور بن محمد السليمان الأستاذ بكلية علوم الحاسب والمعلومات هو الباحث الرئيس للمشروع البحثي رقم -DRI KSU-1292 بعنوان: **"إطار لتعدين وسائل الإعلام العربية واستخلاص المعلومات باستخدام التمثيل الدلالي المتصل للكلمات"** ضمن مبادرة التعاون الدولي (منحة القدرات البحثية) والمدعوم من وكالة البحث والابتكار بوزارة التعليم، حيث بدأ المشروع بتاريخ ١٤٤١/٦/٢٩ هـ الموافق ٢٠٢٠/٢/٢٣م.

وقد أعطيت هذه الإفادة للباحث بناء على طلبه لتقديمها إلى من يهمه الأمر، دون أدنى مسؤولية على عمادة البحث العلمي.

والله الموفق ،،

عميد البحث العلمي

أ.د. محمد بن إبراهيم الوابل

2. Summary of the project in Arabic

مع تزايد وتنوع المواد الاعلامية المنشورة باللغة العربية (المقروءة والمسموعة والمرئية)، تبرز الحاجة الى تسخير التقنيات الرقمية لتمكين استيعاب تلك المواد ومعالجتها وتصنيفها. وكذلك تتأكد الحاجة (مع المتغيرات العديدة المحيطة بنا) لاستنباط معلومات حول تلك المواد والرأي العام المصاحب لها. وهنا تأتي المهام المتعلقة بميكنة أعمال المتابعة والمعالجة والتصنيف الآلي للمواد الاعلامية التي تبُث وتنشر على مدار الساعة. وتعد دراسات وابحاث التنقيب في الفيديوهات من المواضيع الحية بحثيا، خاصة ما يتعلق بمعالجة المواد المنشورة باللغات العالمية كالإنجليزية، اذ ان معظم اعمال البحث والاسترجاع التقليدية تُبنى على فهارس مسبقة البناء وفق مصطلحات مقننة ومحددة. وتعود أسباب محدودية آليات البحث والاسترجاع الى ما يسمى فجوة المعنى، والتي تشير الى الفجوة الكبيرة بين ما يفهم من اللغات الطبيعية وسياقاتها من قبل الانسان مقابل ما يفهم من قبل الحواسيب الآلية.

وتبرز أهداف هذا المشروع البحثي في استكشاف وتطوير نموذج عمل موحد لفهم المواد الاعلامية المنشورة باللغة العربية. ويتطلب ذلك دراسة طرق الربط المفاهيم المرئية والمسموعة والنصية التي تظهر بشكل غير منتظم في الوسائل الاعلامية الدارجة. كذلك سيتم النظر الى التطبيقات ذات العلاقة بالأمن السيبراني لهذه الدراسات والابحاث.

وسعيا لتحقيق الأهداف البحثية المرجوة لفهم المعاني والمغازي من محتويات الفيديو المعروض، فإننا بحاجة الى بناء عدة أنظمة بشكل متوازي. أول هذه الأنظمة هو نظام التعرف على المواد المسموعة باللغة العربية (واللهجة السعودية). وثاني هذه الأنظمة هو نظام اكتشاف والتعرف على النص المعروض ضمن المواد المرئية. وثالث هذه الأنظمة هو نظام التعرف على المفاهيم والأشياء والأفعال التي تحتويها المواد المرئية. وسترسل مخرجات كل نظام من هذه الأنظمة الثلاثة الى نظام رابع مخصص لتحليل

المعنى واستنتاج تمثيل مناسب للمعارف ذات الأهمية. إن هذا التمثيل لتوصيف وفرز المواد المرئية (الفيديو) سيستخدم في البحث والاسترجاع للمواد المرئية بعد اكتمال عناصر المشروع البحثي.

وتأتي الأعمال المتضمنة في هذا التقرير لتشمل مراجعة شاملة وتحليل للأبحاث المماثلة التي تطرقت لمثل هذه المواضيع بمساراتها الأربعة. وقد ركزنا في ذلك على أحدث الأدوات والتقنيات والتطبيقات للتعلم العميق (deep learning). كما أجرينا دراسة لقواعد البيانات التي بُنيت سابقا في مجالات المشروع البحثي، وعقدنا المقارنات بالإيجابيات والسلبيات لكل منها وحللنا أبرز المزايا التي ستساعدنا في المراحل البحثية المتقدمة.

بدأنا بعد ذلك ببناء أولي للأنظمة الثلاثة للتعرف على الأجزاء المسموعة، والمقروءة، والمرئية ضمن المواد الإعلامية، وقد وصلنا الى نتائج ممتازة في مسار التعرف على الصوت ومسار التعرف على النص، وسيتم استعراض ذلك خلال التقرير.

ومن أهم الانجازات خلال العام الأول من المشروع البحثي هو بناء قاعدة بيانات للنصوص العربية الفصحى الحديثة ونصوص باللهجة السعودية. حيث استهدف المشروع بناء قاعدة لا تقل عن 100 مليون كلمة، ومع ذلك من الجدير بالذكر أننا تمكنا من تجاوز أكثر مما كان مستهدف بالفعل لإنشاء أحدث وأكبر معجم للغة العامية السعودية، متضمنًا أكثر من بليون كلمة. فيحتوى المعجم على +8 مليون كلمة فريدة و+ 146 مليون جملة وما يقارب بليون كلمة من اللغة العربية الفصحى الحديثة. كما يحتوي على + 6 مليون كلمة فريدة و+ 14 مليون جملة و+150 مليون كلمة من اللغة العامية السعودية؛ والتي تم جمعها من خمسة مصادر مختلفة، تشمل المعاجم الموجودة مسبقًا، والمواقع الإلكترونية، ومنصات وسائط التواصل الاجتماعي. وسيكون هذا المعجم -مستقبلا- الحجر الأساس للتعرف على المعاني المتضمنة في المواد المرئية والفيديوهات.

3. Summary of the project in English

With the ever increasing quantity of online multimedia information, there is a pressing demand for technologies that facilitate accessibility and exploitation of the mostly unstructured knowledge contained in that media. The requirement for tracing and classifying the popping topics in media and the requirement for the automatic collection of peoples' opinions about these topics has gained much attention recently. The idea of automatic understanding of events in videos is not new. Extensive research has targeted the problem in the last two decades, in what is known as content-based video archival and retrieval. However, the problem has proven very challenging, and the state of the art is still very limited. Most of today's commercial video retrieval systems mostly rely on meta-data for indexing video content. This is mainly due to what is called the "semantic gap", which denotes the difference between the sophisticated and context-dependent human-level semantics, as in natural language, and the computational representations achieved by state-of-the-art artificial intelligence methodologies, such as computer vision and speech recognition.

The main goal of this research project is to investigate and develop a unified framework for understanding Arabic multimedia content. To achieve this goal, we are investigating methods to map different visual, auditory, and textual information in unstructured videos to a common semantic space. We are also looking into the important application of this work to cyber security.

In order to achieve our research target in extracting semantic information solely from the video content, we need to develop three independent systems. The first one is a recognition system for Arabic speech in the Saudi dialect. The second system is for Arabic text extraction from video content. The third one is a recognition system for the visual content in the video. Outputs from each independent system will be sent to a semantic module that will analyze and deduce effective representation of the video content. This representation will be later used for sorting videos and also for querying videos based on similar content.

To achieve these objectives, we performed a comprehensive review to study and analyze the available state of the art models related to these three systems in the literature. We paid more attention to the deep learning-based systems. We also analyzed and compared many of the

available datasets in the literature to know the pros and cons of each dataset and the implications of their use in the project. This knowledge helped us to select the most suitable datasets for our preliminary experiments, and it might be helpful later in constructing our own dataset.

Then we started developing the first versions of the three systems to recognize speech, text, and video content and evaluated them on some of the available datasets. For the speech and text recognition systems, we got excellent results as will be described in this report and its appendices.

Another achievement for the first year is the massive collection of Modern Standard Arabic (MSA) and Saudi Dialect (SD) texts. Our goal was to build a database of at least 100 million words; yet we were able to exceed such a goal and created the newest and largest SD corpora to date. The size of the MSA text is +8M unique words, +146M sentences, and ~1B words, while the size of the SD and mixed texts is +6M unique words, +14M sentences, and +150M words. It was collected from five resources, including pre-existing corpora, websites, and different social media platforms such as comments in YouTube videos. In the future, this corpus will serve as a backbone for our video semantic processing system.

4. Stages of the project completion:

A description of the stages of completion during the project period (also add the completion chart to the project phases, you can use the form below)

Stage	Project duration in months											
	1	2	3	4	5	6	7	8	9	10	11	12
1.1 Acoustic model training												
1.2 Language model training												
1.3 Large vocabulary ASR decoder finite state compilation												
1.5 Data annotation for 300 hours of calls recordings												
1.6 Data collection and cleaning of 100 Million words of MSA and Saudi Colloquial Arabic												
1.7 Creation for a pronunciation dictionary for MSA and Saudi colloquial												
2.1 Developing Arabic text detection algorithm in videos												
2.2 Developing Arabic text extraction algorithm in videos												
3.1 Arabic text data collection and pre-processing												
3.2 Arabic Word vectors estimation												
4.1 Recognizing video concepts												
4.2 Mapping video concepts												
		Work is in progress					Completed				Planned	

5. Main Outputs (Publication)

quarter			Search Status			Publishing authority	search title
Q3	Q2	Q1	Posted for publication	Accepted for publication	Published		
		X			x	Sensors, MDPI	Sensor-based Human Activity Recognition with Spatio-Temporal Deep Learning
	X				x	IEEE Access	Building a Large Contemporary Saudi Corpus: An Incremental Approach to Corpus Design and Construction
							Arabic text detection and recognition from streaming videos using deep learning techniques (**in process**)

6. Other research outputs

Notes	Output type
	Authored books
	Translation of books
	Patented
	Other outputs (specify)
We contribute to the literature of Saudi Dialect corpora by creating the largest Saudi corpus – the King Saud University Saudi Corpus (KSUSC) – with +1Billion total words, including 126M Saudi Dialect words. The KSUSC not only is the newest and largest Saudi Dialect corpus but is also diverse, covering 26 domains.	**Dataset**: the King Saud University Saudi Corpus (KSUSC) – with +1B total words, including 126M Saudi Dialect words
We developed two datasets of text images: 1. A manual human labelled dataset from video frames. (21402 Characters including 2728 spaces) 2. A synthetic Arabic OCR dataset with diverse texts and fonts over various backgrounds.	**Dataset**: Arabic OCR database (real and Synthetic database)
We propose to develop an Arabic OCR to detect Arabic texts in video frames and still images	**System**: Arabic Automatic OCR.
We are developing our own Arabic ASR engine to recognize the speech from Arabic videos, for both MSA and Saudi Dialect.	**System**: Arabic Automatic speech recognition engine (Arabic ASR)
Pre-processing system to clean and normalize the semantic text corpus	**System**: Incremental pre-processing system

7. Technical details

1. Phase 1: Developing Arabic speech recognition for MSA and Saudi dialect

We present the results of the Arabic ASR system which was built using the KALDI toolbox. We used the whole version of the Arabic GALE corpus to train the model. Then, we tested the trained model using continuous speech from streaming videos, where we got excellent results. We tested the performance of the trained model using the Saudi dialect. We note that the performance of a Saudi dialect recognition system requires enhancement, by using more speech data. Hence, for future work, we plan to collect Saudi dialect speech, other than what was available in GALE, to fine-tune the trained model for better performance to achieve the objectives of the project. In addition, we aim to investigate an end-to-end speech recognition system such as a transformer for an Arabic ASR system. We anticipate that the end-to-end system will be more efficient and will have better performance.

For literature review and detailed technical information, please see the report in section "**Phase 1 Developing Arabic speech recognition for MSA and Saudi dialect".**

2. Phase 2: Improving Arabic text extraction from videos:

Beside retrieving objects, actions and speech from videos in the process of video mining, it is crucial to extract the texts that are within each frame. The complexity of this task includes randomness in text position, text size, font type and text content, etc. All these factors make the text extraction a very tedious task, and requires very advanced methods in machine learning. The problem is then formulated as two folds, detect the text positions as boundary boxes images then extract the content from the small text images. We investigated a single deep learning method that accomplishes the above two steps in one pass. The experiments of this proposed method on the public dataset ALIF have given better results compared to the state of the art.

Additionally, we are preparing a new dataset for Arabic OCR recognition that includes other variants of texts that are related to the Saudi Arabian culture. This dataset will be used for our Video Text OCR and to train our deep models. A paper describing the proposed technique and its related results, is being prepared for publication in a couple of weeks.

For literature review and detailed technical information, please see the report in section "**Phase 2 Improving Arabic text extraction from videos**".

3. Phase 3: Learning a Semantic Representation for Arabic.

In the following we report the progress of Phase#3 (learning a semantic representation for Arabic). In the first year plan, phase#3 included two main objectives: "Arabic text data collection and preprocessing" (phase#3.1) and starting with "Arab word vectors estimation" (phase#3.2).

In order to accomplish phase#3.1, we **first** conducted a comprehensive survey of 33 Arabic corpora to help researchers understand the progress of Arabic corpora and the current limitations of Saudi Dialect corpora. **Second**, we intensively collected text from different sources and diverse domains that contains over 1.2B words. During the collection process, pre-exiting corpora, Facebook, YouTube, and Twitter as well as other websites were used to collect text discussing recent events. **Thirds**, we built a new incremental preprocessing system to create Saudi Dialect lexicons and use them to clean and normalize the text. **Fourth,** we accomplished the targeted KPI by building a contemporary linguistic corpus for the Saudi language, which is named the KSUSC corpus. To our knowledge, this corpus is the newest and largest Saudi Dialect corpora to date, with +1B words, +184M sentences, and +26M unique words, covering 26 different domains. Moreover, the collection process for building the KSUSC is discussed in details, and the challenges in collecting SD text with respect to each platform are highlighted in secion2.

When it comes to phase#3.2, it was found that new models and representation have emerged recently and, particularly, sentence embeddings (BERT-Like models) outperformed the word vector estimation. BERT-Like models are generally self-supervised machine learning techniques that make use of the huge amounts of unlabeled text data available on the internet. Thus, we **first** surveyed BERT-Like models to help researchers understand these models and how they can improve the semantic representation of Dialect text. **Second**, we have empirically tested the most powerful BERT-Like models with different datasets over semantic sentence similarity search task. **Third**, we conducted another empirical test to evaluate different BERT-Like models on a dataset of Arabic pairs of sentences. From the preliminary results, it was

possible to conclude that SBERT-paraphrase model had the best performance and, thus, further test will be conducted in the future to generate a sentence embeddings model for Saudi dialect.

For literature review and detailed technical information, please see the report in section **"Phase 3 Learning a Semantic Representation for Arabic"**.

4. Phase 4: Developing video content recognition system

Video content recognition is one of the most challenging and comprehensive problem being addressed in this project. We are working to have an effective video content recognition system in which multiple components such as static scenes, objects and dynamic events and actions should be addressed in integrated ways.

Video understanding and analysis, action recognition is challenging due to the extremely high variation of the representative video features, which compose both the spatial and temporal aspects

In the context of video understanding and analysis, action recognition is the most challenging part due to the extremely high variation of the representative video features, which compose both the spatial and temporal aspects. A robust modelling for such complex features is the key point for retrieving the actions in the video. Such robust modelling is nearly impossible without investigating the powerful deep learning techniques.

In this work we address the problem of video understanding in two ways. In one hand we try to enhance the performance of action recognition techniques on well-known benchmarks such as Kinetics dataset via two approaches. The first approach is based on employing a variational feature learning on top of action recognition models for better discriminate inter/intra action categories. The second approach is based on feature fusion of the action recognition with the classification features of the corresponding frames while on the other hand, we try to generalize the solution toward a comprehensive video understanding.

We also investigated the performance of some state-of-the-art deep learning concepts, such as the inflated 3D convolution neural (I3D) networks and the slow-fast frameworks with the powerful ResNet architecture as a backbone. Some modifications were investigated to improve the performance of the mentioned architectures. With careful data pre-processing, both architectures obtain comparable results in multiclass classification. The performance of SlowFast in multi-label classification is slightly low. This might be attributed to the lack of balance in the

dataset action tags. Various experiments with different hyper-parameters were conducted partially on two datasets, Kinetics-400 and holistic video understanding (HVU). Overall the experimental results show encouraging performances in many scenarios.

For literature review and detailed technical information, please see the report in section "**Phase 4 Developing video content recognition system"**

Phase#1

Developing Arabic speech recognition for MSA and Saudi dialect

Rajab 1442 – Feb 2021

Table of Contents

1. Literature Review

When we need to extract information form any video, the video consists of audio, visual, and text content. Because our project is based on Arabic language, we need to perform automatic Arabic speech recognition (AASR) to extract the audio information from the streaming videos. In this section, we review the state-of-the-art in the AASR filed. The review is divided into three parts: speech corpora, techniques, and tools.

Form the database point of view, we can notice that the lack of availability of a large-scale Arabic speech corpora compared to other languages. Most of existing databases were recorded from TV broadcast. One of the largest corpus is GALE Arabic, which was developed by linguistic data consortium (LDC). There are many versions from GALE corpus. The corpus consists a recording of broadcast conversation and broadcast news from Arabic TV channels [1]. Another important database is KSU speech database, which was produced by speech processing group at King Saud University [2]. This corpus was designed to fulfil the requirement of Arabic speaker/speech recognition systems. It contains the recording of large number of speakers, about 257 in 3 sessions, from different nationalities (Saudi, Arab, Non-Arab) [2]. MGB-2, is another important database and stand for Multi-Genre Broadcast and consists of 1200 hours of recording form TV programs [3].

From techniques point of view, deep learning has become the dominant technique in AASR, hence we focus herein only on research based on deep learning. A. Ali et al. developed a broadcast news system based on 200 hours from GALE phase 2. They used a conventional acoustic model and deep neural network acoustic model. The best results were achieved by DNN-MPE model with recorded 15.8% 32.21%, and 26.95% word error rate (WER), for reports, conversational, combined sets, respectively [1]. Long short-term memory (LSTM) and gated recurrent unit (GRU) are used for Arabic speech recognition in [4]. They tested the proposed system for 10 spoken Arabic digit recognition task and 10 spoken command TV task [4]. Deep auto-encoder was used for speech enhancement in [5], for remote Arabic speech recognition system. The authors used isolated words Arabic speech database for their experiments, where the database contained only recording of 20 words.

In terms of available toolboxes for ASR, with the era of deep learning, the availability of huge amount of training data, and the fast growth in computation devices resulted in the release of lot of ASR toolkit such as Baidu's Deep Speech from Mozilla [6], wav2letter from Facebook [7], PyTorch-Kaldi [8], openseq2seq from Nvidia [9], and ESPnet [10]. This is in addition to the old ASR toolkit such as HTK and Sphinx. Each of these toolkits has advantages and disadvantages. In our project, we are currently using Kaldi because it supports an available recipe of the Gale database, which we intend to use in the project. Moreover, we will investigate another state-of-art end-to-end ASR such as wav2letter in future.

As a conclusion, we notice that AASR needs more study and enhancement. Databases containing Saudi dialect is also required. Moreover, applying End-to-End (E2E) ASR for Arabic speech recognition system still needs more investigation.

2. Initial Arabic Speech Recognition system

In this section, we show the process of building an Arabic ASR system using KALDI tool. We follow the chronological steps of KALDI to build the system, which covers the tasks (1.1, 1.2, 1.3, 1.4 and 1.7) of phase 1 in the proposal. The aim of this project is to build an Arabic video mining system, in which ASR is an important component to recognize the speech from the audio part of the video clips. To be able to compare the performance of our techniques and system with the state of the art results, we opted to use an available public domain speech database rather than building the database ourselves.

2.1. Database used

From searching for Arabic speech corpus, we selected the Arabic GALE corpus, which is very large and is the most suitable for our video mining project because it was collected from broadcast news and conversations. Moreover, there are several existing research in the literature using the GALE

corpus. Gale has the following characteristics: (i) is publicly available, (ii) has a large collection of vocabulary and a very good amount of annotated data, and (iii) has published research using it. Gale has many phases and each phase has conservation and news parts. In the first six months of the project, we used the parts of Gale initially available to us: GALE Phase 2 Arabic Broadcast Conversation Speech Part 1 and Part 2 and GALE Phase 2 Arabic Broadcast News Speech Part 1, as shown in Figure 1 .

GALE Arabic speech database		
Phase Name	Total hours	Total of Saudi hours
GALE Phase 2 Arabic Broadcast Conversation Speech Part 1	123	7
GALE Phase 2 Arabic Broadcast Conversation Speech Part 2	128	5.3
GALE Phase 2 Arabic Broadcast News Speech Part 1	165	0.5
Total	416	12.8

Figure 1: Initial GALE Arabic total hours and speech of Saudi channels.

2.2. Program used

We used Gale Arabic recipe (s5b) in kaldi branch 5.0, which was developed by Ahmed Ali, et al. in Ref [1]. The run script consists of many steps the can be executed together. They started by preparing the training data and converting the flac files to wave files. Then they prepared the data by splitting it into train, dev, and test sets for the news and conversation parts. Table 1 shows the lists of acoustic models, which we investigated in developing an Arabic speech recognition system.

Table 1. Acoustic model lists

Acoustic model		Command used	Comments
GMM-HMM	Monophone	`steps/train_mono.sh`	
	Triphone	`steps/train_deltas.sh`	Number of gauss = 30000
	tri2a	`steps/train_deltas.sh`	Using deltas+delta+deltas #gauss = 40000
	tri2b	`steps/train_lda_mllt.sh`	Using LDA+MLLT #gauss = 50000
	tri3b	`steps/train_sat.sh`	Using LDA + MLLT + SAT #gauss = 100000
DNN	TDNN	`local/nnet3/run_tdnn.sh`	Using chain lattice-free recipe #epoch = 3 Activation function = relu ivector dim = 100, #hidden layers = 6 Total training time of this network is: 10:23:11

2.3. Feature extraction

MFCCs are extracted by the command (steps/make_mfcc.sh) for each set. Then, cepstral mean and variance statistics are computed per speakers using this script (steps/compute_cmvn_stats.sh). For the deep neural network, I-vectors are used as a feature. The next step is to create high resolution MFCC, then compute the diagonal UBM using 512 Gaussians, after that I Vectors are extracted. More information can be founded in this script (local/nnet3/run_ivector_common.sh)

2.4. Initial Results

Table 2 reports the word error rate (WER) presented in the original Kaldi recipe and the WER for our trained system using different acoustic models.

Table 2. WER of an Arabic ASR using Gale Arabic database

Acoustic model	Type of speech	WER % (our training)	WER % (amali recipe)
Tri1	Report	31.33	26.38
Tri2a	Report	30.58	25.66
Tri2b	Report	27.83	23.32
Tri3b	Report	25.75	21.64
TDNN	Report	12.85	10.72
Tri1	Conversational	49.25	46.86
Tri2a	Conversational	47.94	45.92
Tri2b	Conversational	44.28	42.23
Tri3b	Conversational	41.52	39.26
TDNN	Conversational	27.48	24.77
Tri1	Conversational + Reports	43.54	40.35
Tri2a	Conversational + Reports	42.42	39.42
Tri2b	Conversational + Reports	39.05	36.17
Tri3b	Conversational + Reports	36.27	33.61
TDNN	Conversational + Reports	22.78	20.26

2.5. Testing the initial ASR system on speech out of the corpus

We successfully tested the best model, which is TDNN, on speech out of the corpus to show the effectiveness and performance of this ASR. We selected a new recent speech segment, which was related to covid-19 from Al-Arabiya channel and a speech segment from the KSU database. The results of the system are presented in the Figure 2 and Figure 3 below and show that our system produced excellent output.

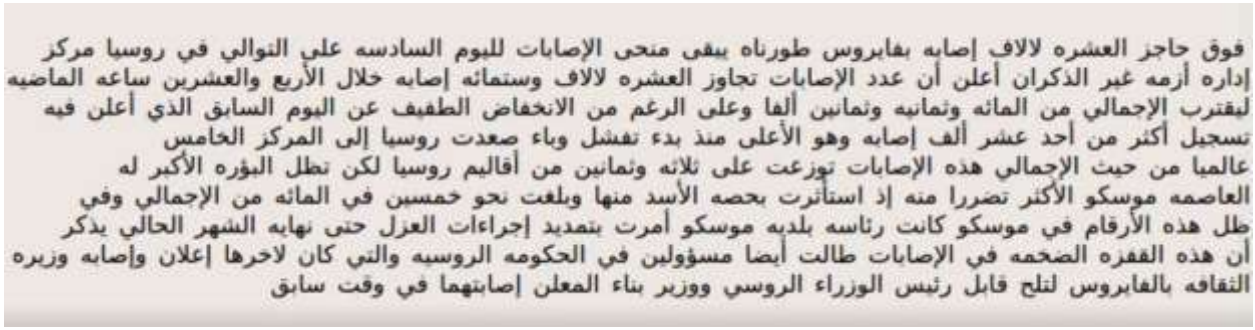


Figure 2:Test case 1: The output of the system for YouTube clip from Al-Arabiya channel

Figure 3:Test case 2: The output of the system Free speech from KSU Arabic database .

3 Arabic speech recognition system using large scale speech corpus

From the above experiments of the six-month report, we conclude the following points:

- The built GALE based ASR model needs more training data in order to recognize the speech efficiently.
- The GALE based ASR needs to be tuned using speech of Saudi dialect in order to achieve the objective of this project.
- For the integrated final system of this project, we need ASR model that can be integrated with the other parts (i.e. action recognition, object detection, OCR).

Hence, in this report, we investigated using of the whole GALE database to train the ASR system and evaluate the trained system by feeding speech from Arabic videos and Saudi dialect.

3.1. Corpus description

In this section, we provided a short description of Gale the speech database used for training and testing the ASR model. GALE is a huge Arabic database that contains recording of Arabic broadcast conversation and news. It consists of many phases with a total number of hours of speech equal to 1227 hours; out of it, 53.9 hours are from Saudi channels. Table 3 presents the details of the different phases of Arabic Gale.

Table 3: GALE Arabic total hours and speech of Saudi channels.

GALE Arabic Broadcast Conversation Speech		
Database name	Total hours	Total saudi hours
GALE Phase 2 Arabic Broadcast Conversation Speech Part 1	123	7
GALE Phase 2 Arabic Broadcast Conversation Speech Part 2	128	5.3
GALE Phase 3 Arabic Broadcast Conversation Speech Part 1	123	12.7
GALE Phase 3 Arabic Broadcast Conversation Speech Part 2	129	3.6
GALE Phase 4 Arabic Broadcast Conversation	75	2.8
Total	578	31.4

GALE Arabic Broadcast News Speech		
Database name	Total hours	Total saudi hours
Gale phase 1	17	0
GALE Phase 2 Arabic Broadcast News Speech Part 1	165	0.5
GALE Phase 2 Arabic Broadcast News Speech Part 2	170	8.1
GALE Phase 3 Arabic Broadcast News Speech Part 1	132	6.5
GALE Phase 3 Arabic Broadcast News Speech Part 2	128	7.4
GALE Phase 4 Arabic Broadcast News	37	0
Total	649	22.5

3.2. Machine configuration

Because of huge size of the GALE corpus, the model takes a lot of time to train, hence we used a powerful machine to train the model using this corpus. The machine that we used in the experiments has the following specifications: CPU: E5-2660, GPU: TITAN RTX-24GB, and RAM: 220 GB.

3.3. Training procedure

We follow the “run.sh” program in gale recipe (s5b) in the Kaldi environment of this research [1]. The run program consists of 10 stages containing the feature extraction, training, and decoding. In Table 4 we list all stages and their function.

Table 4: Summary of "run.sh" script.

Stage	Function
0	Preparing data, lexicon, and language model
1	Generating MFCC features for train and test files
2	Training the monophone system (mono)
3	Align data using the mono system and training triphone system (tri1) using delta features
4	Building graph of the tri1 system and decoding the test files.
5	Align data using tri1 system, and training the triphone system using the LDA+MLLT features (tri2b)
6	Building graph of the tri2b system and decoding the test files.
7	Align data using tri2b system, and doing speaker adapting training (SAT) using the fMLLR-adapted features (tri3b)
8	Building graph of the tri3b system and decoding the test files.
9	Training and decoding the chain model using end-to-end alignments

The training and validation loss of the chain model (last stage) are presented in Figure 4. The training time of each iteration is presented in Figure 5.

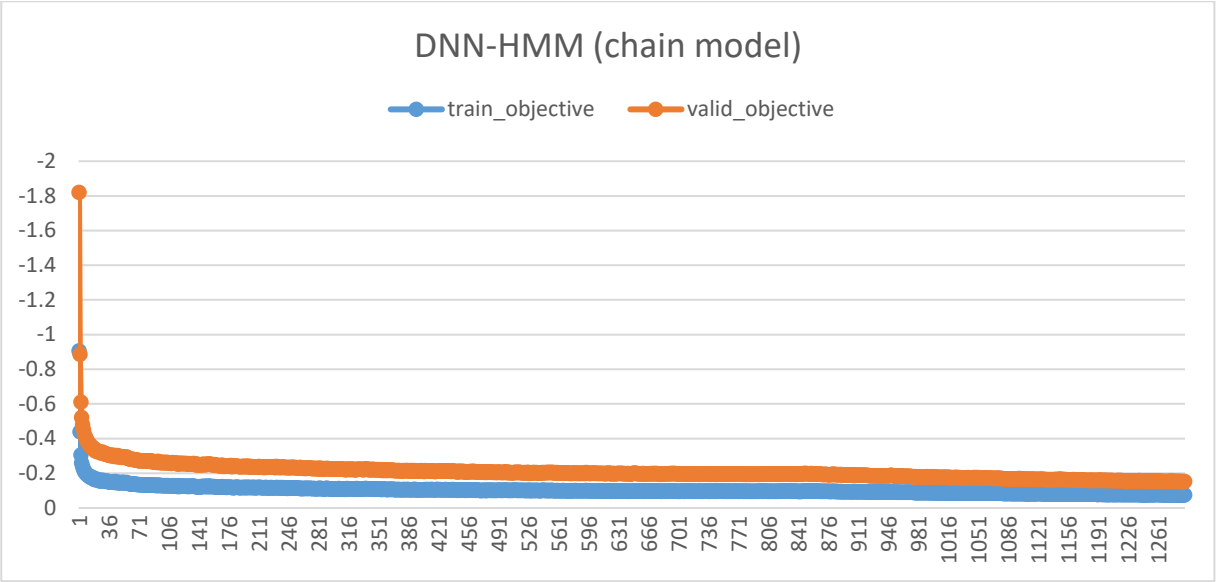


Figure 4: Training and validation loss of chain model using GALE corpus.

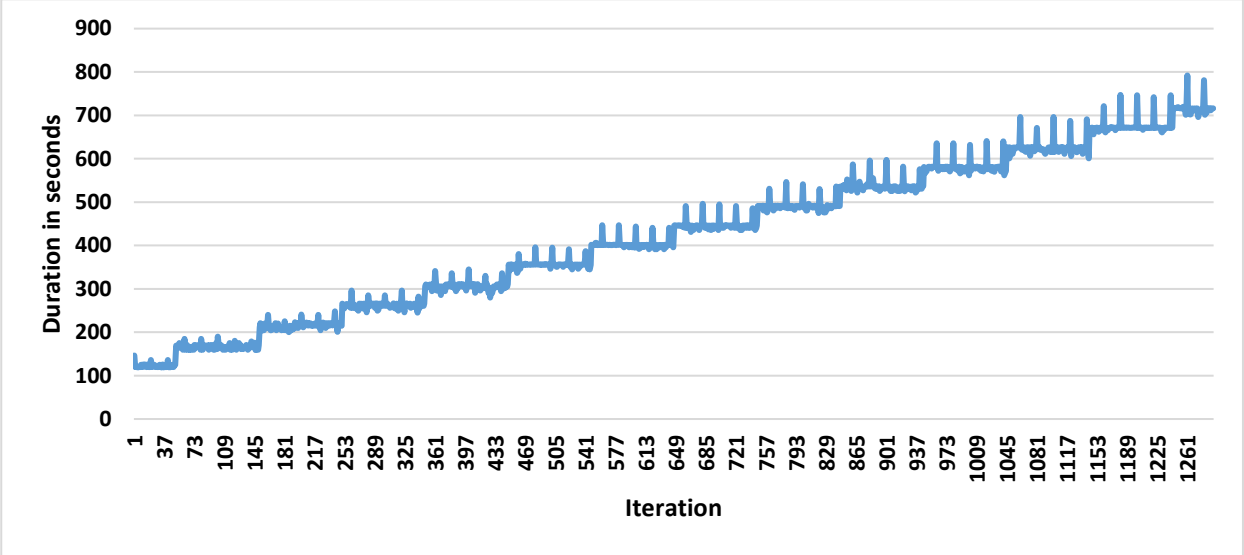


Figure 5: Duration in seconds for each of training iterations of the chain model.

As shown in Figure 5, the total time increased for each iteration resulting in a very long training time, approximately 6.3 days.

3.4. Results

To evaluate the trained model, we calculate the word error rate (WER) and character error rate (CER) for all testing files using each of the trained models as presented in the Figure 6. We can see that the acoustic model based on DNN (chain model) reached 14.65% WER, which is around 50% lower than the (Tri3b) model. Not that the WER presented in the original recipe [1] using the same model was 14.95%

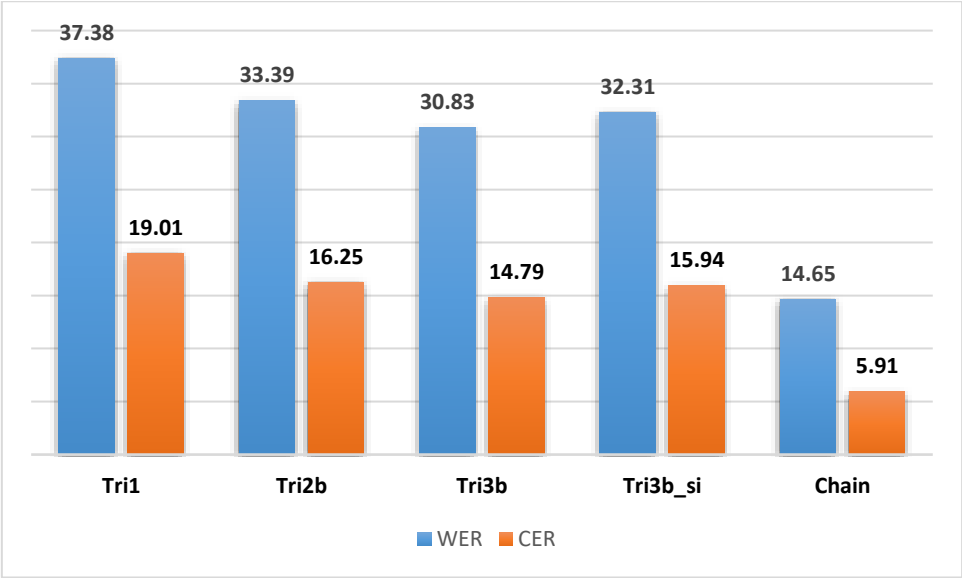


Figure 6: Performance of the trained model using the test files of the GALE corpus.

In Table 5, we present the details of the performance achieved by the chain model in terms of number of testing utterances, insertions errors, deletion errors, and substitution errors.

Table 5: Details performance of the chain model.

	# of testing samples	# of correction (%)	# of insertions (%)	# of deletions (%)	# of substitutions (%)
CER	316,317	297,623 (94.1%)	6102 (1.9 %)	6983 (2.2%)	5609 (1.8%)
WER	69,668	59,464 (85.4%)	1370 (2.0%)	1949 (2.8%)	6885 (9.9%)

3.5. Demos

In this section, we demonstrate the result of the best model (chain) using online speech from a YouTube video and a speech segment from a paragraph in KSU database to evaluate the systems using data that is not part of the GALE corpus.

Example1

Canonical text:

Speaker: "Prince Mohammed bin Salman announces the strategy of the Public Investment Fund"
Duration: [from 0.15 to 0.48] seconds
URL: https://www.youtube.com/watch?v=3ys1ivxQPPQ

استراتيجية صندوق الإستثمارات العامة عشرين واحد وعشرين إلى عشرين خمسة وعشرين التي تمثل مرتكزاً رئيسياً في تحقيق طموحات وطننا الغالي نحو النمو الاقتصادي ورفع جودة الحياة وتحقيق مفهوم التنمية الشاملة والمستدامة في مختلف القطاعات التقليدية والحديثة.
حقق صندوق الإستثمارات العامه خلال الأعوام السابقة إنجازات استثمارية و إقتصادية ضخمة تمكن من خلالها الوصول لمستهدفات استراتيجية مهمة حيث ضاعف صندوق الإستثمارات العامه إصوله إلى ما يقارب تريليون وخمسمائة مليار ريال سعودي بنهاية عام عشرين عشرين كما أطلق صندوق الإستثمارات العامة.

Recognized text:

استراتيجية صندوق الإستثمارات العامة عشرين واحد وعشرين إلى عشرين خمسة وعشرين
التي تمثل مرتكزا رئيسيا في تحقيق طموحات وطننا الغالي نحو النمو الاقتصادي
ورفع جودة الحياة وتحقيق مفهوم التنمية الشاملة والمستدامة في مختلف القطاعات التقليدية والحديثة
حقق صندوق استثمارات العام خلال الأعوام السابقة إنجازات استثمارية إقتصادية ضخمة تمكن من خلالها
الوصول مستهدفا استراتيجية مهمة حيث ضاعف صندوق استمرت العام وصوله إلى ما يقارب تريليون وخمسمائة مليار ريال سعودي
بنهاية عام عشرين عشرين كما أطلق صندوق استمارات العامة

Example2

Speaker: Native Arabic (Man) from KSU database

Canonical text:

مما يشرحُ الصَّدْرَ ، ويزيحُ سُحُب الهمِّ والغمِّ ، السَّفَرُ في الديارِ ، وقَطْعُ القفارِ ، والتقلبُ في الأرضِ الواسعةِ ، والنظرُ في كتابِ الكونِ المفتوحِ لتشاهد أقلام القدرةِ وهي تكتبُ على صفحاتِ الوجودِ آياتِ الجمالِ ، لترى حدائق ذات بهجةٍ ، ورياضاً أنيقةً وجناتٍ ألفاً ، اخرجْ من بيتكَ وتأملْ ما حولك

وما بين يديك وما خلفك ، اصْعَدِ الجبال ، اهبطِ الأودية ، تسلّقِ الأشجارَ ، عُبَّ من الماءِ النميرِ ، ضعْ أنفك على أغصانِ الياسمين ، حينها تجدُ روحك حرةً طليقةً ، كالطائرِ الغرّيدِ تسبحِّ في فضاءِ السعادةِ ، اخرجْ من بيتِك ، ألقِ الغطاء الأسودَ عن عينيك ، ثم سرْ في فجاجِ اللهِ الواسعةِ ذاكراً مسبحاً.

Recognized text:

مما يشرح الصدر ويزيح سحوب وللهم والغم السفر في الديار وقطع القفاري
والتقلب في الأرض الواسعة والنظر في كتاب الكون المفتوح
لتشاهد أقلام القدرة وهي تكتب على صفحات الوجود آيات الجمال
لترى حدائق ذات بهجة ورياضا أنيقة وجنات ألفا
أخرج من بيتك وتأمل ما حولك وما بين يديك وما خلفك
يصعد الجبال يهبط الأودية تسلق الأشجار عبى من الماء النمير ضاع أنفك على أغصان الياسمين
حينها تجد روحك حرة طليقة كالطائر الغريد تسبح في فضاء السعادة
أخرج من بيتك ألقي الغطاء الأسود عن عينيك ثم سر في فجاج الله الواسعة ذاكرا مسبحة

3.6. Demo using Saudi dialect

In this section, we demonstrate the result of the best model (chain) using speech from a YouTube video with Saudi dialect to evaluate the system performance on Saudi dialect.

Speaker: Saudi (women)

Canonical text:

راح يكون معك حق اذا قلت لهم راح يكون معك حق اذا قلت لهم
راح يكون له تأثير عليهم راح يكون له تأثير عليهم
راح تعرف منه كل شيء راح تعرف منه كل شيء
راح تصير مشكلة إذا كلمته عن خويه راح تصير مشكلة إذا كلمته عن خويه
أعتقد إنه راح الشغل أعتقد إنه راح الشغل
أظن ان هو قال كذه أظن ان هو قال كذه
راح يفكر بالموضوع راح يفكر بالموضوع
سمعت آخر خبر سمعت آخر خبر
تدري إيش صار أمس تدري ليش صار أمس
تعرف إنه خلاص ترك الشغل تعرف إنه خلاص ترك الشغل
لا تسوي إزعاج لا تسوي إزعاج
لا تقول إنك طفشان لا تقول إنك طفشان

لاتصدق كلامه لاتصدق كلامه
لا ترد عليه لا ترد عليه
لا تكثر كلام لا تكثر كلام
لا تتجاهلني لا تتجاهلني
لا تضايقهم لا تضايقهم

Recognized text:

راح يكون له تأثير عليهم راح يكون له تأثير
راح تعرف من كل شيء راح تعرف من كل شيء
راح تصير مش مشكلة إذا تكلمت عن خاوية راح تصير مشكلة إذا تكلمت عن خاوية
أعتقد إنه راح الشغل أعتقد إن راح الشغل
أظن هو قال إذا قال كده أظن هو قال كده
راح يفكر بالموضوع راح يفكر بالموضوع
سمعت آخر خبر سمعت آخر الخبر
تدري إيش صار أمس تدري ليش صار أمس
تعرف إنه خلاص ترك الشغل تعرف إنه خلاص ترك الشغل
لا تساوي إزعاج لا تسوي إزعاج
لا تقول إنك كوفشان لا بتقول إنك قشان
لا ترد عليه لا ترد عليه
لا تكفي كلام كثير كلام لا
لا تضايقهم لا تضايقهم

4 Data annotation (task 1.5)

As we mentioned in section 2.1, we used Arabic Gale database because it was the best for our video mining project because it has recording of broadcast news and conversation. Moreover, Gale is transcribed and annotated at the sentence level. Hence, we did not have to perform data annotation on this database. From the Gale description in section 3.1, we notice that the Saudi dialect speech of the Gale is not enough to build an accurate ASR for Saudi dialect. This was confirmed by the result in the previous section. Therefore, we started to collect more Saudi dialect speech from the social media and YouTube channels. After the collection of speech with Saudi dialect, we will transcribe and annotate the speech.

5 Data collection (task 1.6)

Outputs from speech engines can be enhanced by the use of a strong language model, which takes the outputs from the speech engine and converts them to a sequence of characters that conforms to the language. We are using the dictionary included in the Arabic Gale database, unfortunately, Saudi dialect words may not be included, and this leads in some instances to some nearby words that are out of context. In order to achieve better state of the art results, we participated in the data collection conducted by the semantic team (i.e. team of phase 3), and we collected nearby 1 billion words in diverse contexts.

In the next months, we will fine-tune the current language model by training it with Saudi dialect, we also will use the videos collected at the semantic part to extract the Saudi Arabic dialect audios from the videos. This will also complete the task 1.6 of this phase. Data collected by the semantic team is well described in their respective report.

6 Conclusion and future work

In this report, we demonstrated the results of the Arabic ASR system which we built using the KALDI toolbox. We used the whole version of Arabic GALE corpus to train the model. Then, we tested the trained model using continuous speech from streaming videos. Moreover, we demonstrated the performance of the trained model using Saudi dialect. We note that the performance of a Saudi dialect recognition requires enhancement. Hence, for future work, we plan to collect more Saudi dialect speech in order to fine-tune the trained model for better performance to achieve the objectives of the project. In addition, we aim to investigate an end-to-end speech recognition system such as a transformer for an Arabic ASR system. We anticipate that the end-to-end system will be more efficient and has better performance.

References

[1] A. Ali, Y. Zhang, P. Cardinal, N. Dahak, S. Vogel, and J. Glass, "A complete KALDI recipe for building Arabic speech recognition systems," in *2014 IEEE spoken language technology workshop (SLT)*, 2014, pp. 525–529.

[2] M. Alsulaiman, G. Muhammad, M. A. Bencherif, A. Mahmood, and Z. Ali, "KSU rich Arabic speech database," *Inf.*, vol. 16, no. 6 B, pp. 4231–4253, 2013.

[3] S. Khurana and A. Ali, "QCRI advanced transcription system (QATS) for the Arabic multi-dialect broadcast media recognition: MGB-2 challenge," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 292–298.

[4] N. Zerari, S. Abdelhamid, H. Bouzgou, and C. Raymond, "Bidirectional deep architecture for Arabic speech recognition," *Open Comput. Sci.*, vol. 9, no. 1, pp. 92–102, 2019.

[5] B. Dendani, H. Bahi, and T. Sari, "Speech Enhancement Based on Deep AutoEncoder for Remote Arabic Speech Recognition," in *International Conference on Image and Signal*

Processing, 2020, pp. 221–229.
[6] A. Hannun *et al.*, “Deep speech: Scaling up end-to-end speech recognition,” *arXiv Prepr. arXiv1412.5567*, 2014.
[7] Q. X. J. C. J. K. G. S. V. L. R. C. Vineel Pratap Awni Hannun, “wav2letter++: The Fastest Open-source Speech Recognition System,” *CoRR*, vol. abs/1812.0, 2018.
[8] M. Ravanelli, T. Parcollet, and Y. Bengio, “The PyTorch-Kaldi Speech Recognition Toolkit,” in *In Proc. of ICASSP*, 2019.
[9] O. Kuchaiev *et al.*, “Mixed-Precision Training for NLP and Speech Recognition with OpenSeq2Seq.” 2018.
[10] H. Inaguma *et al.*, “ESPnet-ST: All-in-One Speech Translation Toolkit,” *arXiv Prepr. arXiv2004.10234*, 2020.
[11] S. Yousfi, S.-A. Berrani, and C. Garcia, “ALIF: A dataset for Arabic embedded text recognition in TV broadcast,” in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 1221–1225.

Phase#2

Improving Arabic text extraction from videos

Rajab 1442 – Feb 2021

Table of Contents

1. Literature Review

The need for an Arabic optical character recognition (OCR) system has increased due to the large number of Arabic texts that are stored in images, not in textual forms. Many old books, and new books that are stored in PDF format are being transformed into text-based modern formats like EBUD, MOBI, or AMZ, but manually, which is time-consuming and bound to human-error. The importance of having an Arabic OCR system for the Arabic language has increased due to the large number of Arabic speaking population which is estimated at around 407-420 million people [1]. The Arabic language is also important for non-Arab Muslims, because all the religious texts are written, and cited in Arabic.

But Arabic language is known to be a difficult language, not just to learn, but also to deal with in digital systems. That is why researchers have considered developing an OCR system for Arabic language as an open problem until today. This difficulty arises from the nature of Arabic text, the way of writing Arabic texts specifically, which can be summarized by following:

1. Arabic Texts, unlike Latin languages, are written from right to left.
2. The characters within a single word can be connected or disconnected depending on the character itself. For example, the character "ي" is considered as a connected letter, and can be written in the letter shape: this shape "يـ", this shape "ـيـ", or this shape "ـي". Unlike the character "ر" which can be only written in the letter shape, or this shape "ـر", and no character can be connected to it from the left.
3. Each character has multiple forms, depending on its location within a word, and the previous character if it is a connected character or a disconnected character. For example, the character "ه" has four forms: this form "ه" which is used when it's located at the end of the word, and the previous character is a disconnected character, this form "هـ" which is used when it is located at the beginning of the word, this form "ـهـ" which is used when it is located at the middle of the word, and this form "ـه" which is used when the character is located at the end of the word, and the previous character is a connected character. Table 1 shows all the characters written in all forms. Arabic Texts have a cursive feature, which differs depending on the font style.
4. A diacritic (Harakat) can additionally be added to any Arabic character, which are: " ◌َ "," ◌ً "," ◌ُ "," ◌ٌ "," ◌ِ "," ◌ٍ ", and " ◌ّ ", and some characters can have more than one diacritic. Some Arabic words may consist of the same characters in the same order but have more than one meaning depending on the diacritics of the word.
5. The Arabic language has also ligatures, which is a special way to write two characters. It differs depending on the font style, for example: "محـ", and "نـي".
6. Some Arabic characters have similar shapes, but they can be distinguished by the existence of the dots, their quantity, and their position, which can be above the character, or below the character. For example, these three characters: "حـ" ,"خـ" ,"جـ".

Table 3. All Arabic characters in all forms

Name	Isolated	Initial	Medial	Final
alif	ا			ـا
baa	ب	بـ	ـبـ	ـب
taa	ت	تـ	ـتـ	ـت
thaa	ث	ثـ	ـثـ	ـث
jiim	ج	جـ	ـجـ	ـج
haa	ح	حـ	ـحـ	ـح
khaa	خ	خـ	ـخـ	ـخ
daal	د			ـد
dhaal	ذ			ـذ
raa	ر			ـر
zaay	ز			ـز
siin	س	سـ	ـسـ	ـس
shiin	ش	شـ	ـشـ	ـش
saad	ص	صـ	ـصـ	ـص
daad	ض	ضـ	ـضـ	ـض
taa	ط	طـ	ـطـ	ـط
dhaa	ظ	ظـ	ـظـ	ـظ
Ayn	ع	عـ	ـعـ	ـع
ghayn	غ	غـ	ـغـ	ـغ
faa	ف	فـ	ـفـ	ـف
qaaf	ق	قـ	ـقـ	ـق
kaaf	ك	كـ	ـكـ	ـك
laam	ل	لـ	ـلـ	ـل
miim	م	مـ	ـمـ	ـم
nuun	ن	نـ	ـنـ	ـن
haa	ه	هـ	ـهـ	ـه
waaw	و			ـو
yaa	ي	يـ	ـيـ	ـي

These features have made the development of an Arabic OCR system a real challenge for researchers, especially for the segmentation process, since unlike the Latin languages, most of the Arabic words consist mainly of connected characters. Due to this many works were dedicated to the segmentation process exclusively, but there are others who developed a complete OCR system, like the work of [2] which consists of skew detection and line segmentation, word segmentation, character segmentation, and finally character classification. The accomplished accuracy for line segmentation is good, which varies between 97% and 100%, depending on the font style and if the standard deviation verification of the line length is applied or not. As for the character segmentation, it uses the Zidouri segmentation algorithm [3], which uses the vertical projection, as well as the number of horizontal pixels to find guide bands, and then several rules are applied. The accuracy varies between 46% and 84%, depending on the font style, this variation makes the algorithm font-dependent. The classification stage has an accuracy of 82%, it uses the decision tree classifier, generated by 24

features. The complete system accuracy has not been computed, but the author has used the segmentation accuracy multiplied by the classification to have an approximated accuracy of the overall system which is 61%.

In general, most of the researchers focus on the segmentation problem as it is considered as the main source of errors in any Arabic OCR system [4]. The developed algorithms use many approaches in order to segment the characters correctly, but in general, they can be classified into two main classes: Implicit segmentation, and explicit segmentation. In implicit segmentation, the characters are being segmented during the recognition phase, there is no real segmentation of the characters and it usually uses a variable-size window that passes the whole word, to provide the tentative segmentation points which are confirmed or not during the classification phase. In [5] the authors have used the restricted Boltzmann machine for classification, and compared their results of using RBM with the use of Hidden Markov Model (HMM). The character rate accuracy of using the RBM is 95.2%, while the character rate accuracy of HMM is 87%. In [6] the authors have designed an Arabic OCR system that uses three sliding windows on sub-word image, which are fed into a multichannel neural network to predict the likelihood of the input window that is a candidate cut place. The authors compared their model with a similar model but using only one window. Using the size of 18pt for the text from the APTI dataset [7] for testing, the three-windows' model has accomplished an accuracy of 98.9%, while the single-window model has accomplished an accuracy of 90.2%. both are trained using only one font. There is another version of the three-window model that is trained on 4 fonts, which has an accuracy of 95.5%.

In explicit segmentation, the characters are segmented before they are being classified or predicted. This approach has been used by many researchers and is classified into many subclasses. The first attempt to segment Arabic characters is traced back to 1975 by Nazif [8] [4], yet it is not known what type of approach have been used, nor the achieved accuracy. The most common method for Arabic characters' segmentation is done by extracting the vertical projection profile, then several operations are used, and rules can be applied to segment the connected characters. This method is computationally simple and accomplished good results on simple fonts, however without applying any rule it failed in the case of overlapped characters and cursive fonts. This method was used in [9], where it firstly approximates the limits of characters in each word using a vertical projection, modulated by the vertical width of writing, which sets a certain threshold for the width of the characters. Then the maximum number of black segments on a line of pixels is calculated, each character must have only one black segment except for some characters located at the end of the word. This method was tested on handwritten and typewritten text, but it was designed for typewritten specifically. The accuracy accomplished for handwritten text is 96%, but the testing data seems very small, containing only 4 words written by 5 persons, and some rules were followed by the writers to

make the testing data. As for the typewritten text the algorithm was tested on 500-1000 words typewritten using four different fonts that do not have overlapped characters, and it accomplished an accuracy between 99% and 100%. In [10] the authors presented a segmentation algorithm that is based on the vertical projection profile, and some rules were applied. The rules are related to the structural characteristics between the background region and the character component, as well as the isolated characters. The testing accuracy of the algorithm is 94.7%. It was tested on 500 samples, written on different fonts, but the rules were not reported. In [11] the authors designed an Arabic OCR system that does the word segmentation using the vertical projection profile, and the character segmentation using the same method with a set of statistical features. The system was tested on APTI dataset, using 24816 words written by 10 different fonts of different sizes, and the achieved accuracy is 97.7%.

Contour tracing is another method for character segmentation. In [12] the authors proposed two methods for Arabic/Latin handwritten texts. The first one uses the pen thickness to detect the junction (baseline) that connects the characters, while the second detects the upper contour of each word, then two filters are applied to get two types of strokes, which the authors denote as “valleys” and “summits”. Those strokes will be analyzed to get the primary segmentation points. Both methods have been tested on handwritten Latin and Arabic texts, the first one achieved an accuracy of 93.5%, while the second one achieved an accuracy of 99.3%. The second one is also faster according to the authors.

Template matching can be used as well to segment Arabic characters. In [13] Bushofa and Spann have proposed an OCR system for Arabic texts, using a sliding window of size 7×7 that searches along the baseline to find the segmentation points. The reported accuracy of the overall system is 97.23%, it was tested on a private dataset that consists of four different fonts and different sizes as well.

In [14] the authors have proposed a technique where a wavelet transform is used to detect the underlying horizontal edges of the word, then those edges that overlap with the baseline are considered as candidate segmentation points, and using NN classifiers, those candidates are filtered into definite segmentation points. The reported accuracy of 97.83%, was compared to other works that used spatial-based techniques, and has shown not just a higher accuracy, but also a higher recognition rate.

2. Building an end-to-end Character recognition system

2.1. Introduction

In our initial investigation while preparing the proposal, we were intending to use a 3 step system, the text detection within the images, the detected text extraction as small images, and the text segmentation and binarization, but in order to be more efficient, we proposed a more efficient method that uses a deep neural technique where the three steps are fused into a single system. The proposed method gave very high accuracy and has shown state of the art efficiency on public available datasets. So by this proposed method we are accomplishing tasks 2.1, and 2.2, and 2.3 in one single task.

Hence, we investigated the application of recent deep learning techniques for Arabic optical character recognition (OCR). We proposed a system that uses deep learning models to recognize the Arabic text in video frames at real time. To evaluate the proposed system, we used two datasets, a publicly available dataset called ALIF which was proposed by S. Yousif et el., in 2015 [1]., this dataset contains 2308 annotated images for training and 3221 images for testing, the test bed is split into three sets, with various difficulty in the texts within the images.

The ALIF dataset is segmented at the character level, but lacks various fonts and sizes of the texts and most of the text images are already cropped within small images with limited height and width. Any extension to huge video frames, would make the trained model unable to generalize. In order to avoid such problem, we opted to develop our own dataset, by two methods, by labelling selected video frames containing various texts at the character level, and by generating synthetic texts with various fonts and sizes over different backgrounds, and if time permits with diverse orientations of the texts. The new KSU Arabic-OCR database will be developed to fulfil the requirements of the whole project of semantic representation of Arabic video mining, in which OCR is playing an important role of the project.

2.2. Details of the system

Our proposed Arabic deep OCR consists of two phases as shown in

Figure 1. At the first phase, we start the training process by using the public Arabic dataset ALIF, by using the text images and their respective annotations as bounding boxes. At the second phase, frames without corresponding annotations are fed to the detection system, in order to test the model for detecting and automatically segmenting Arabic characters within videos frames. Output results of the test images will then be compared to the annotated characters of the test annotations as sequences of successive letters.

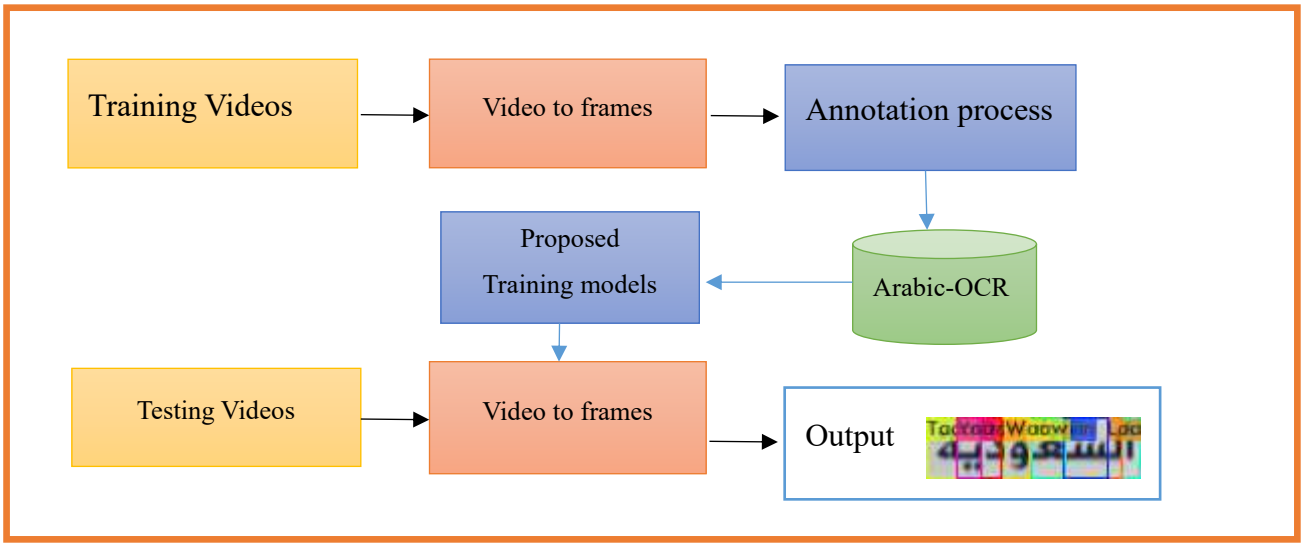


Figure 1: Research Methodology.

2.3. Character level Arabic Datasets:

This phase of the project aims essentially to detect the Arabic text, including the Saudi dialect, from the whole frame of video, rather than from the segmented regions. This is why the training models require to be trained on Arabic character level. In our proposed system, we started by ALIF as a baseline, and at a later stage, we will use our developed Arabic OCR dataset for a second more general trained model, able to be tested on video frames with various sizes and texts.

2.3.1. ALIF Dataset:

The ALIF dataset is originally split into four sets, a train set and three tests sets. The sets contain different numbers of images, depending on the complexity of the texts, including blurring, diverse sizes, etc…, that make the test sets with variable difficulties of text detection.

Dataset Split	Number of text images (Character Level Annotated)	Number of characters Including space
Train	2308	-
ALIF test 1	900	14110
ALIF Test 2	1299	19306
ALIF Test 3	1022	13362

A graphical distribution of the letters of the datasets is presented in Figure 2 to Figure 5

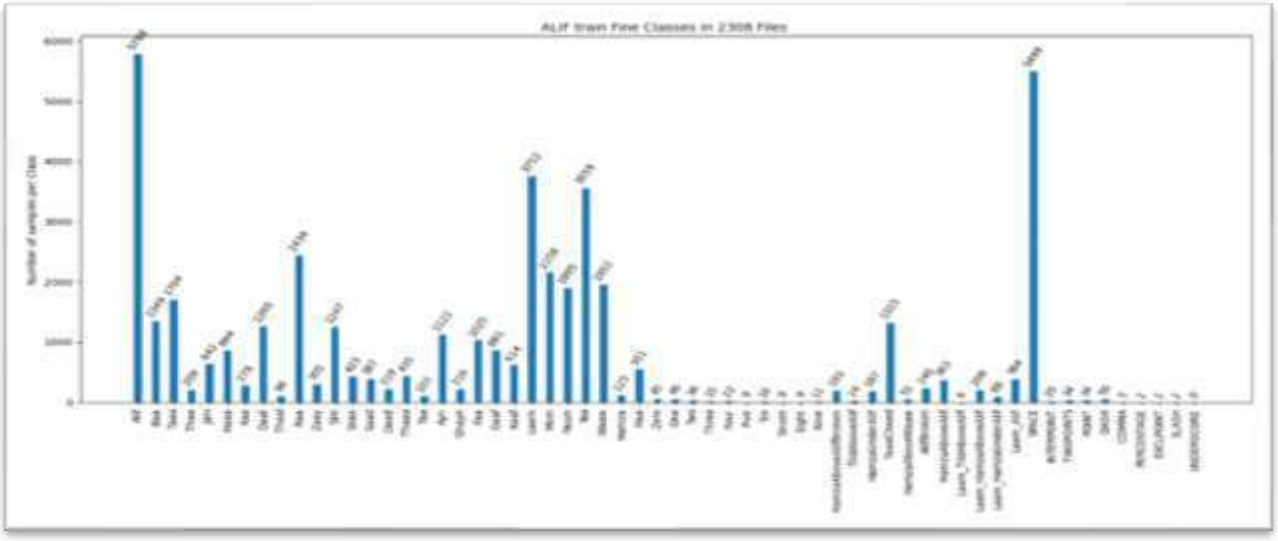


Figure 2:ALIF Train dataset

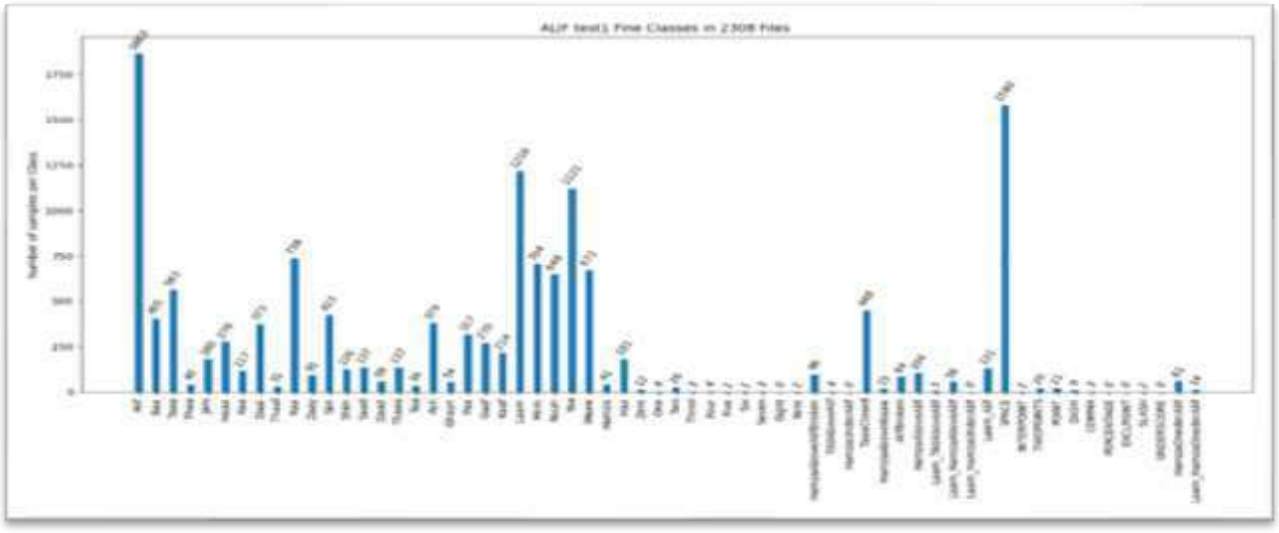


Figure 3, ALIF Test 1 dataset

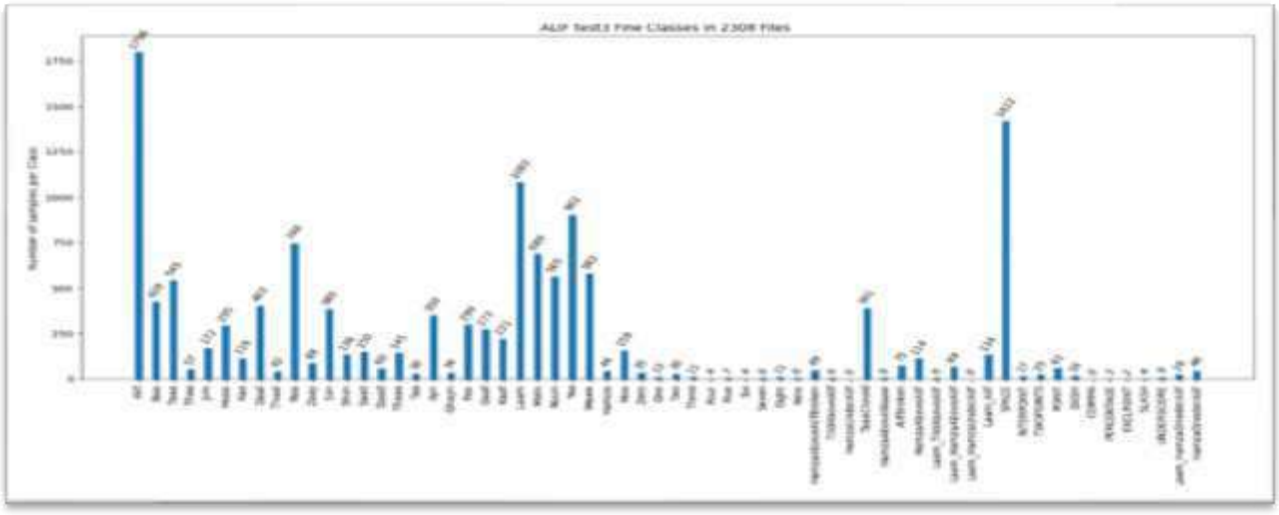


Figure 4, ALIF Test 2 dataset

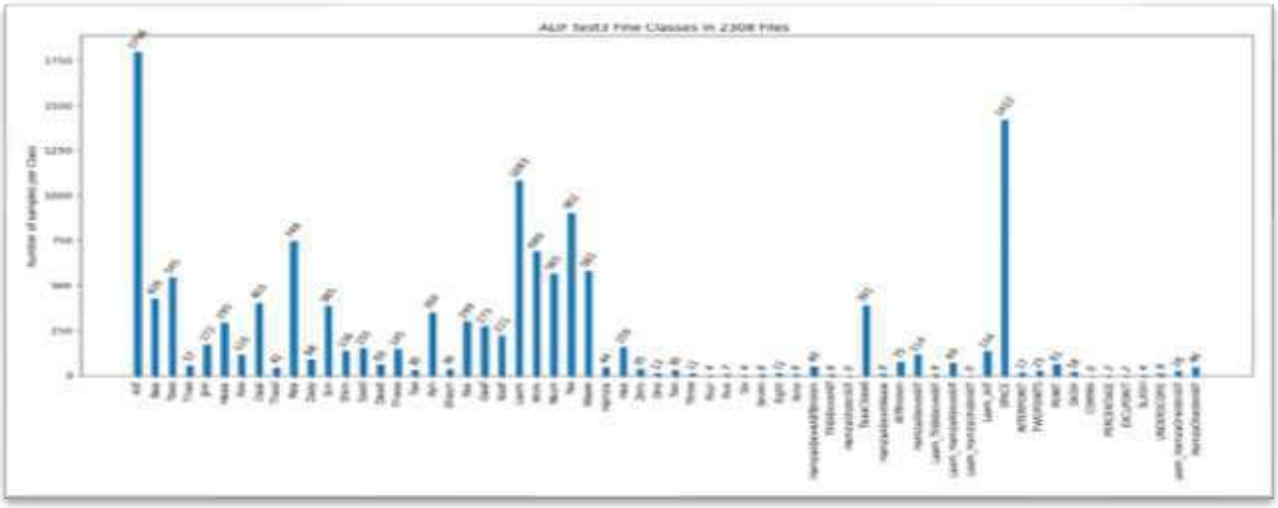


Figure 5, ALIF Test 3 dataset

2.3.2. KSU-OCR dataset

Given that ALIF does not cover a varied set of fonts and sizes of texts, and with the need to extend our work to local Saudi Colloquial Arabic, as well as to Arabic international video content, we started by selecting a set of TV shows and news videos, with some variety in size of text and the backgrounds. The videos were given to two annotators to make the bounding boxes at two levels, a sentence level and a character level. We have preferred to use the same ALIF set of characters, to have a uniform strategy, if we want to compare models. A list of the characters is presented in Table 2, where we also present the statistics per character. Most characters can occur at start of a word, can be isolated, in the middle or at the end. In our annotation style, we omitted such complications. In a future version, such additions can be done by an automatic manner with 10% manual sample checking.

Alif:2518	Miim:1211	Six:7
Baa:735	Nuun:991	Seven:2
Taaa:775	Yaa:1537	Eight:14
Thaa:80	Waaw:951	Nine:12
Jiim:287	Hamza:62	HamzaAboveAlifBroken:64
Haaa:447	Haa:401	TildAboveAlif:10
Xaa:114	Zero:44	HamzaUnderAlif:94
Daal:667	One:39	TaaaClosed:526
Thaal:61	Two:32	HamzaAboveWaaw:20
Raa:919	Three:14	AlifBroken:109

Zaay:171	Faa:505	HamzaAboveAlif:252
Siin:434	Gaaf:372	Laam_TildAboveAlif:5
Shiin:239	Kaaf:454	Laam_HamzaAboveAlif:79
Saad:249	TWOPOINTS:10	Laam_HamzaUnderAlif:26
Daad:115	POINT:15	Laam_Alif:174
Thaaa:29	DASH:69	SPACE:2728
Taa:186	COMMA:21	*INTERPOINT:0
Ayn:645	PERCENTAGE:8	*EXCLPOINT:0
Ghayn:66	Four:10	*UNDERSCORE:0
Laam:1758	Five:39	*SLASH:0

*Table 2: Set of ALIF characters (without position constraints), *: listed to comply with ALIF Dataset character encoding*

The difficulty of the manual selection of the position of each bounding box depended essentially on the size of the text and its font. In this initial work, we investigated only the horizontal texts. In a future version, logos and rotated texts will be taken into account. The initial texts that were chosen to be labelled varied in terms of height and width, in both Figure 6**Error! Reference source not found.** and Figure 7, where the distribution of the collected text areas (size of the bounding box of each text sentence within the image), are shown.

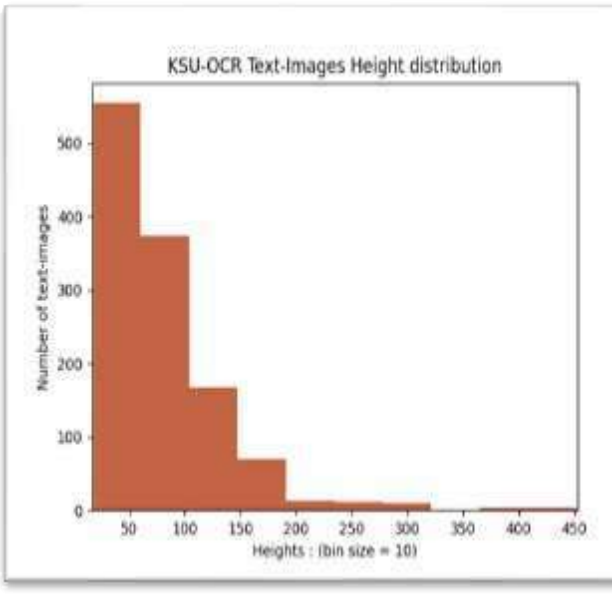


Figure 6, Height Histogram, Min: 17, Max: 452

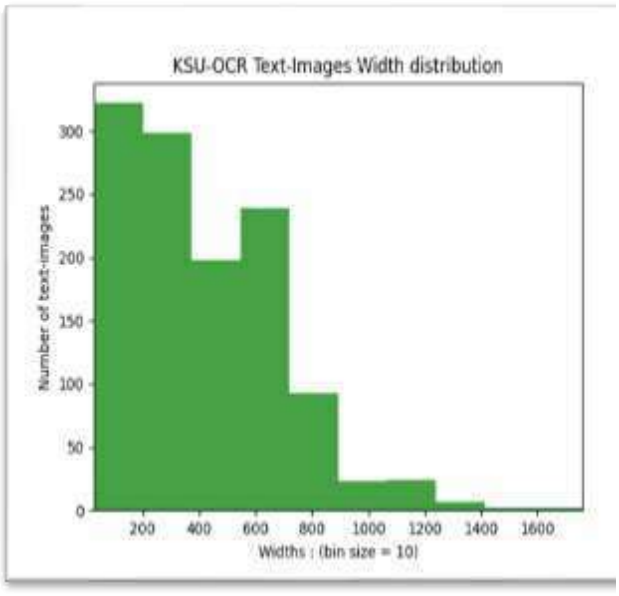


Figure 7, Width histogram, Min: 27, Max: 1758

A better representation of the character distribution of our collected Arabic KSU OCR dataset is presented in Figure 8.

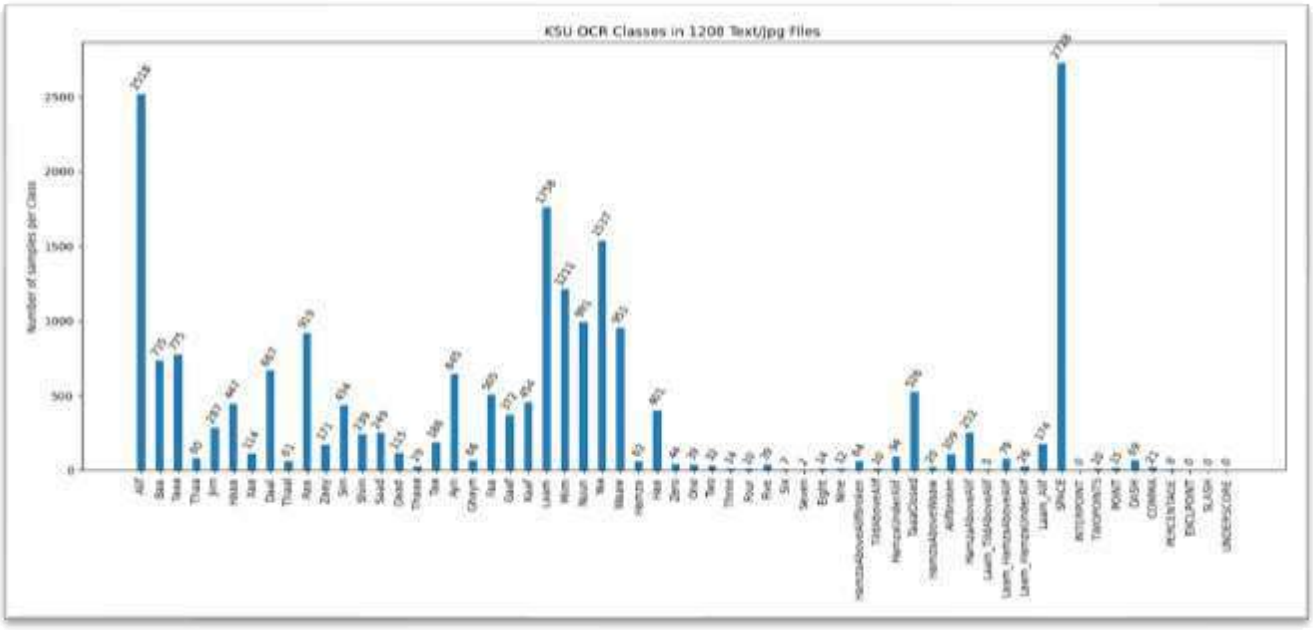


Figure 8, Histogram of the character distribution of the Arabic KSU-OCR

We need also to investigate in a more uniform set of text images, in order to have a better model trained on an evenly distributed set of Arabic characters .For example the character Alif and the Space are the most dominant characters in our dataset (this is the same in ALIF dataset). Adding also numbers might be a very interesting approach as many plates and road signs have numbers and can contribute to better understanding of the video content.

2.3.3. KSU-OCR dataset Labeling Methodology

In order to improve the data collection procedure on our KSU–OCR dataset, we investigated further in splitting the video texts of the dataset into different sentence images, this approach might be useful when a text detector is introduced at the frame level to detect probable text areas, and send them as separate bounding boxes to be recognized into successive letters words.

Each video of the KSU-OCR dataset has been split into frames, each frame split into areas of texts, then each area of text split into boundary boxes for letters. The annotations were made by the labelMe website, and the results of each frame saved as a separate xml file. The content of a sample XML is presented in Figure 9.

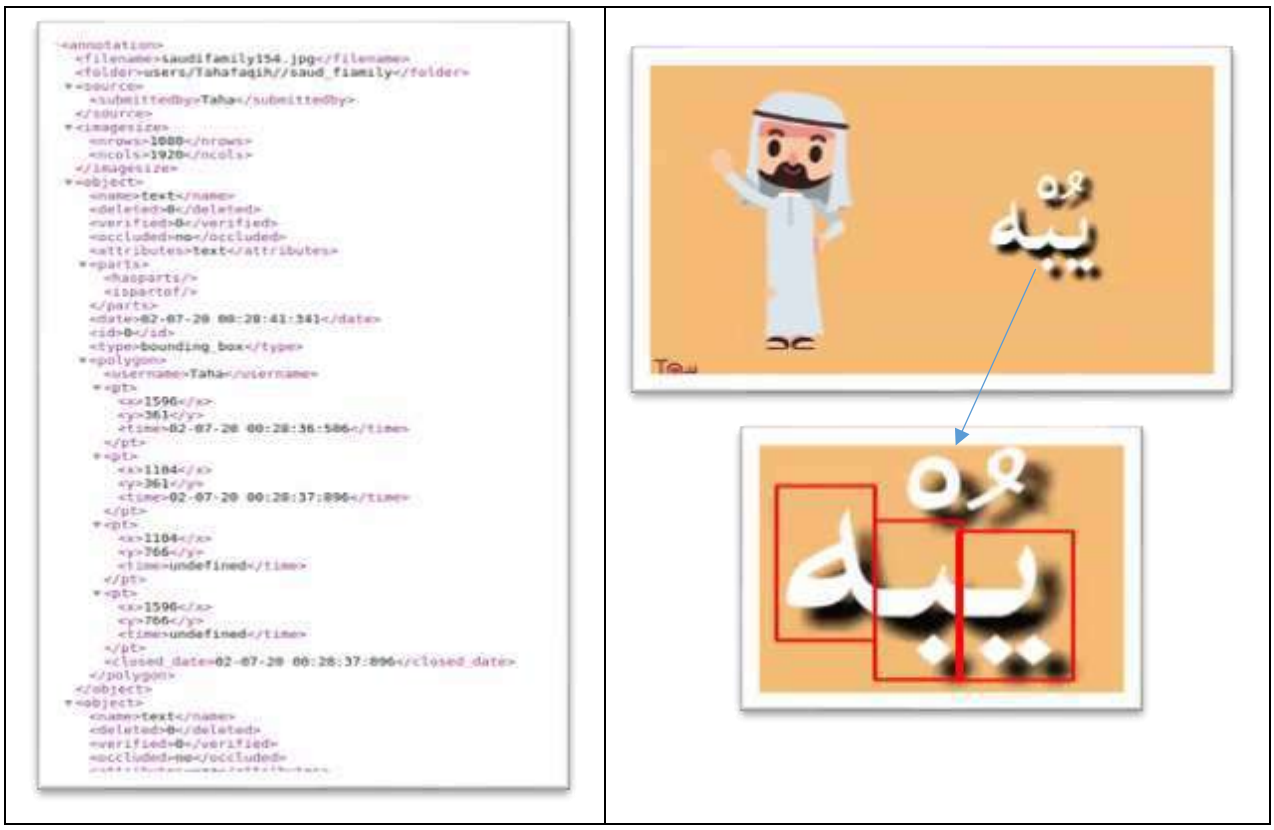


Figure 9, Sample of the Arabic KSU-OCR Dataset annotation

2.3.4. Future Work on KSU-OCR

We are exploring generating Arabic text data synthetically, by superposing Arabic texts over diverse backgrounds, and generating images with their respective bounding boxes. This approach can be a complementary method to balance the dataset and enhance the control over the text fonts, text sizes and orientation.

3. Experimental Results

3.1 Training the proposed system using the ALIF dataset

As mentioned earlier, ALIF text images have limited height, which can be a benefit while training but a handicap for testing on large texts. This point will be taken into account when we will make our own dataset.

In the initial training of our model, we used a fixed image size of (64×512) for the deep neural network. Due to the small size of the database, we used a pertained weight to initialize the weights of the backbone network detector. Then, we trained the model for 10k iterations. Figure 10 shows the training loss using ALIF database. The mean average precision is presented for the testing set.

The performance of the proposed system after 10K iterations is presented in Table 3. We can notice that the system still needs enhancement.

Table 3: System performance after 10K iterations

Precision	75 %
Recall	90 %
F1-score	82 %
mAP@0.50	7.44 %

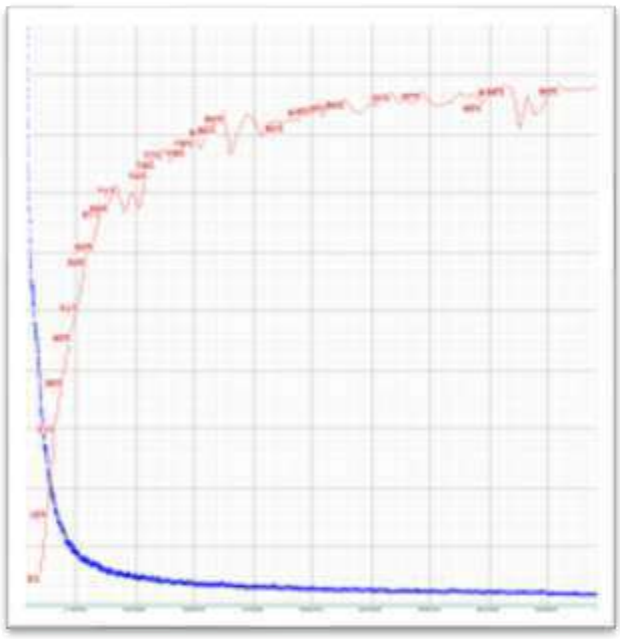


Figure 10. Loss and mAP of ALIF training model

3.2 Character detection and recognition Results on ALIF tests

In order to compare our work to state of the art works on the ALIF dataset, we improved the models and added varied augmentation schemes to have a variety of images, including saturation hue, blurring, etc…We also used an additional approach of multi-labelling, where a character is detected by its variant word positions, at **S**tart, Isolated **M**iddle or **E**nd position, as this information is also provided in the ALIF dataset.

3.2.1 Single label

In this set of experiments, there were 57 labels selected from 60 original labels of the ALIF dataset, the decrease in labels was based on the number of occurrences in the training dataset since some labels had an appearance of less than 10 occurrences in the text images histogram, and hence were discarded. After repeated sessions of training and parameter tuning, the initial results are presented in Table 4. The results are very interesting, as accuracies surpassed the original accuracies of the ALIF paper.

Table 4: ALIF dataset initial results using single label : 57 classes

Test Results	
ALIF_test1	SENT: %Correct=45.05 [H=405, S=494, N=899] WORD: %Corr=94.38, Acc=93.80 [H=13371, D=692, S=104, I=82, N=14167]
ALIF_test2	SENT: %Correct=37.67 [H=486, S=804, N=1290] WORD: %Corr=91.65, Acc=90.93 [H=17746, D=1378, S=239, I=140, N=19363]
ALIF_test3	SENT: %Correct=2.85 [H=29, S=987, N=1016] WORD: %Corr=71.72, Acc=69.07 [H=9624, D=2676, S=1119, I=355, N=13419]

3.2.2 Multi Label

In this set of experiments, there were 62 labels, including position information of each character as a second label for our network. Results of the multi label network are shown in Table 5.

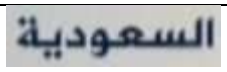
Table 5: ALIF dataset initial results using Multi- label : 62 classes

Test Results	
ALIF_test1	SENT: %Correct=43.51 [H=392, S=509, N=901] WORD: %Corr=94.02, Acc=92.24 [H=26603, D=1450, S=243, I=504, N=28296]
ALIF_test2	SENT: %Correct=36.85 [H=479, S=821, N=1300] WORD: %Corr=91.89, Acc=89.78 [H=35600, D=2615, S=529, I=817, N=38744]
ALIF_test3	SENT: %Correct=3.42 [H=35, S=987, N=1022] WORD: %Corr=76.16, Acc=70.65 [H=20444, D=4326, S=2074, I=1478, N=26844]

The Multi Label results are also very interesting, as our proposed system can detect the character and its position in the word. This is very encouraging, as the system self-detect boundary boxes, and the order of appearance of the characters.

In Table 6, we show some testing samples of the proposed model. We can clearly see that this model recognized and detected the Arabic text with very high accuracy.

Table 6: ALIF result samples

Input image	Output image	Prediction	
		Character	Score
		Alif	100%
		Laam	100%
		Siin	100%
		Ayn	100%
		Waaw	100%
		Daal	100%
		Yaa	100%
		TaaaClosed	100%
		Character	Score
		Alif	100%
		Laam	100%
		Gaaf	96%
		Alif	100%
		Haa	93%
		Raa	100%
		TaaaClosed	100%

3.3 Training the proposed system using KSU database

In order to have a training dataset for the Arabic OCR, as realistic as possible, i.e. similar to daily videos from the YouTube or any other video provider, we selected the size of (640×640) as input image to our network, where text can be at any position, contrary to ALIF where the text is surely within the small text images. With this constraint, we are forcing our network to automatically detect the texts within the big images, then convert the localized text image as a sequence of characters.

With this new size constraint, the training took longer time compared to the ALIF database. Therefore, we trained the system only for 2K iterations to investigate the performance of the proposed model in large images. The performance after 2K iterations is presented in Table 7. We note that the system is in its early stages, and that the system needs further enhancement.

Table 7: System performance after 2K iterations.

Precision	**57 %**
Recall	66 %
F1-score	61 %
mAP@0.50	55.71 %

3.4 Initial Results on the KSU database (Arabic-OCR)

Herein, we show the example of testing the proposed model using our database. We can clearly see that the model has recognized and detected the Arabic text with good accuracy. A very interesting point is that the system can differentiate between English and Arabic text as shown in Table 8.

Table 8 : Arabic-OCR result sample.

Input image	Output Image
	

Predictions

Character	Score	Character	Score	Character	Score
jiim:	95%	laam:	71%	hamzaunderalif	27%
alif:	97%	laam:	96%	baa:	85%
baa:	47%	miim:	37%	ghayn:	43%
laam:	83%	haaa:	66%	xaa:	54%
baa:	99%	miim:	96%	daad:	100%
space:	80%	laam:	75%	alif:	94%
space:	71%	space:	99%	ayn:	99%
				taaaclosed:	94%

Figure 11 shows other examples of the output of the detection and extraction process of our proposed model (using KSU OCR database). From the results, we can clearly notice that the proposed system is encouraging but still needs more enhancements. In addition, post-processing step may be used to remove duplicate detections. We anticipate that these improvements will increase the system performance.

Figure 11: Another example of the detected results.

4. References

[1] "Demographics of the Arab League - Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Demographics_of_the_Arab_League#cite_note-UN-1.

[2] I. Supriana and A. Nasution, "Arabic Character Recognition System Development," *Procedia Technology,* vol. 11, pp. 334-341, 2013.

[3] A. Zidouri, "On Multiple Typeface Arabic Script Recognition," *Research Journal of Applied Sciences,* 2010.

[4] Y. M. Alginahi, *A survey on Arabic character segmentation,* vol. 16, 2013, pp. 105-126.

[5] A. M. Rashwan, M. S. Kamel and F. Karray, "An arabic optical character recognition system using restricted boltzmann machines," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013.

[6] M. A. Radwan, M. I. Khalil and H. M. Abbas, "Predictive segmentation using multichannel neural networks in Arabic OCR system," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.

[7] F. Slimane, R. Ingold, S. Kanoun, A. M. Alimi and J. Hennebert, "Database and Evaluation Protocols for Arabic Printed Text Recognition," 2012.

[8] A. Nazif, "A System for the Recognition of the Printed Arabic Characters," 1975.

[9] B. A. Najoua and E. Noureddine, "A robust approach for Arabic printed character segmentation," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 1995.

[10] L. Zheng, A. H. Hassin and X. Tang, "A new algorithm for machine printed Arabic character segmentation," *Pattern Recognition Letters,* vol. 25, no. 15, pp. 1723-1729, 1 11 2004.

[11] A. Qaroush, B. Jaber, K. Mohammad, M. Washaha, E. Maali and N. Nayef, "An efficient, font independent word and character segmentation algorithm for printed Arabic text," *Journal of King Saud University - Computer and Information Sciences,* 31 8 2019.

[12] K. Romeo-Pakker, H. Miled and Y. Lecourtier, "A new approach for Latin/Arabic character segmentation," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 1995.

[13] B. M. Bushofa and M. Spann, "Segmentation and recognition of Arabic characters by structural classification," *Image and Vision Computing,* vol. 15, no. 3, pp. 167-179, 1 3 1997.

[14] A. Broumandnia, J. Shanbehzadeh and M. Nourani, "Segmentation of printed Farsi/Arabic words," in *2007 IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2007*, 2007.

[15] S. Yousfi, S.-A. Berrani, and C. Garcia, “ALIF: A dataset for Arabic embedded text recognition in TV broadcast,” in 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 1221–1225.

Phase#3

Learning a Semantic Representation for Arabic

Rajab 1442 – Feb 2021

Table of Contents

List of Figures

List of Tables

1. Introduction

This document reports the progress of Phase#3 (learning a semantic representation for Arabic). In the first year plan, phase#3 included two main objectives: "Arabic text data collection and preprocessing" (phase#3.1) and starting with "Arab word vectors estimation" (phase#3.2).

In order to accomplish phase#3.1, we **first** conducted a comprehensive survey of 33 Arabic corpora to help researchers understand the progress of Arabic corpora and the current limitations of Saudi Dialect corpora. **Second**, we intensively collected text from different sources and diverse domains that contains over 1.2B words. During the collection process, pre-exiting corpora, Facebook, YouTube, and Twitter as well as other websites were used to collect text discussing recent events. **Thirds**, we built a new incremental preprocessing system to create Saudi Dialect lexicons and use them to clean and normalize the text. **Fourth,** we accomplished the targeted KPI by building a contemporary linguistic corpus for the Saudi language, which is named the KSUSC corpus. To our knowledge, this corpus is the newest and largest Saudi Dialect corpora to date, with +1B words, +161M sentences, and +14M unique words, covering 26 different domains. Moreover, the collection process for building the KSUSC is discussed in details, and the challenges in collecting SD text with respect to each platform are highlighted in secion2.

When it comes to phase#3.2, it was found that new models and representation have emerged recently and, particularly, sentence embeddings (BERT-Like models) outperformed the word vector estimation. BERT-Like models are generally self-supervised machine learning techniques that make use of the huge amounts of unlabeled text data available on the internet. Thus, we **first** surveyed BERT-Like models to help researchers understand these models and how they can improve the semantic representation of Dialect text. **Second**, we have empirically tested the most powerful BERT-Like models with different datasets over semantic sentence similarity search task. **Third**, we conducted another empirical test to evaluate different BERT-Like models on a dataset of Arabic pairs of sentences. From the preliminary results, it was possible to conclude that SBERT-paraphrase model had the best performance and, thus, further test will be conducted in the future to generate a sentence embeddings model for Saudi dialect.

The remaining of this report will report the details of phase#3.1 contribution. Then, phase#3.2 details and experiment results will be presented. Finally, the report will be concluded and future plan and research direction will be presented at the end.

2. Phase 3.1: Arabic Test Data Collection and Pre-Processing

Language corpus is a term used to describe a collection of texts written in one or more languages [1]. It is considered one of the most important sources of data in different areas, including information retrieval (IR), natural language processing (NLP), and computational linguistics (CL). This is because it can represent the written language and, hence, be used to process opinions and implement related applications. Compared to English corpora, Arabic corpora are poorly resourced and lack sufficient research and data, which negatively affects Arabic-based NLP practitioners [2].

The Arabic language is the mother language of Arab countries and one of the six official languages of the United Nations (UN) [3]. There are three main versions of the Arabic language: Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialect Arabic (DA). CA is the official language used in the Quran and during the medieval period, MSA is the official language used in the modern period and by news outlets, and DA is the spoken language that is used in daily life and differs from one country/city to another. Moreover, there are many subcategories under these Arabic languages, and thus, Arabic is considered to be highly inflectional and to have a very complex morphology [4]. Therefore, the development of large Arabic corpora has been the focus of many researchers in recent years. However, most of these developments have focused on CA and MSA rather than DA; thus, Arabic NLP solutions perform poorly with DA data [5].

In DA, also known as colloquial language, one lemma can have hundreds of surface forms, and thus, it is a morphologically rich language [6]. DA differs from MSA syntactically, morphologically, and phonologically and does not have standard orthographies [7]. As emphasized in the literature [8], it is nearly impossible to have one NLP solution that can process all variations of Arabic. It is also important to have a corpus that reflects the current use of language [9], and thus, new terms that reflect recent events around the world (such as the COVID-19 pandemic and the Saudi Vision 2030 framework) must be incorporated into corpora.

Although MSA used to be considered the dominant written language, due to the social media platforms emerging around the world, DA has become more frequently written than MSA [10]. Moreover, Saudis are considered some of the most active users of social media. It was reported by Alruily [4] that there are more than 11 million Arabic accounts in Twitter, with 27.4 million tweets being published a day; Saudi Arabia has the most active users, with 30% of these tweets. In [11], 61% of the 175M Arabic tweets considered were found to be from Saudi Arabia.

Thus, this section focuses on Saudi Dialect (SD) and highlight the current challenges of SD corpora. In particular, it presents a new system designed to build a Saudi Dialect Corpus named KSU Saudi Corpus (KSUSC). The system proposed is designed to incrementally create two types of lexicons to identify common ASCII and unwanted symbols. These lexicons will be used to clean MSA and SD text from different resources and platforms and then validate the data to eliminate irrelevant characters or incomplete text. The incremental process introduced in this system allows it to be scaled to other languages in DA.

This section is organized as follows. First, the literature review of relevant tools and corpora is presented. Second, the collection process is described and the details of the collected resources are also summarized. Third, the architecture of the proposed system is presented and simple examples are illustrated. Finally, the statistics of the built corpus are presented and discussed to finally conclude the report at the end.

2.1 Literature Review

In this section, Arabic corpora and preprocessing tools available in the literature are discussed and compared to motivate the contribution of this phase.

2.1.1 Existing Arabic Corpus

- **PATB** (MSA) [12] The Penn Arabic Treebank (PATB) corpus began in the fall of 2001 and has had three full releases of morphologically and syntactically annotated data: (1) the Arabic Treebank: Part 1, which consists of 166K words of written MSA newswire from the Agence France Presse corpus; (2) the Arabic Treebank: Part 2, which consists of 144K words from Al-Hayat distributed by Ummah Arabic News Text; and (3) the Arabic Treebank: Part 3, which consists of 350K words of morphologically annotated newswire text from An-Nahar.

- **Gumar (**Gulf Arabic Dialect) [13] Gumar Corpus is a large-scale corpus of Gulf Arabic (GA) that includes a number of sub-dialects in the six countries of the Gulf Cooperation Council: Bahrain (BH), Kuwait (KW), Oman (OM), UAE, Qatar (QA), and Saudi Arabia (SA). It's present preliminary results on GA morphological annotation. Building a morphologically annotated GA corpus is a first step towards developing NLP applications, for searching, retrieving, machine-translating, and spell- checking GA text among other applications. the Gumar Corpus, a large-scale corpus of Gulf Arabic consisting of 110 million words from 1,200 forum novels, the corpus annotated for sub-dialect information at the document level, it presents results of a preliminary study in the morphological annotation of Gulf Arabic which includes developing guidelines for a conventional orthography. The text of the corpus is publicly available through a web interface. The Gumar Corpus is a morphologically annotated Gulf Arabic (GA) corpus. In its current state, the corpus contains more than 112 million words that spans over 1,200 documents. Its content is collected from online novels, where it has 60.52% SA, 13.35% EA, 5.91% KW, 1.13% OM, 0.65% QA, and 0.49% BH. Moreover, there is around 10% that is identified as GA (other) which are the cases of a novel containing a combination of several GA dialects that is due to multiple writers with different dialects or due to the existence of different characters in the novel. It was sometimes hard to differentiate through the text between the three dialects of OM, QA and AE even with a native speaker annotating, hence these cases we marked as GA (other) also. The rest of the corpus (7.93%) is mostly MSA and other DA such as Egyptian, Iraqi, Levantine, ... etc.

- **Gumar Emirati** (Emirati dialect) [14] The first large-scale morphologically manually annotated corpus of Emirati Arabic. This corpus includes about 200,000 words selected from eight Gumar corpus novels in the Emirati Arabic variety. The selected texts are being annotated for tokenization, part-of-speech, lemmatization, English glosses and dialect identification. It defines to be the native spoken variety in the Gulf Cooperation Council, is still lagging behind other Arabic dialects with respect to resource and tool creation, given the considerable amount of Dialect content online.

- **Curras** (Palestinian Dialect) [15] The first morphologically annotated corpus of the Palestinian Arabic dialect. Palestinian Arabic is one of the many primarily spoken dialects of the Arabic language. Arabic dialects are generally under-resourced compared to Modern Standard Arabic,

the primarily written and official form of Arabic, it consists around 43,000 words. The corpus data was collected from variety of resources (Facebook, Twitter, Blogs, Forums, Palestinian stories, Palestinian terms, TV Shows).

- **YADAC** (Egyptian Arabic) [16] It's another Dialect Arabic Corpus is a multi-genre Dialect Arabic corpus – that is compiled using Web data from microblogs (i.e. Twitter), blogs/forums and online knowledge market services in which both questions and answers are user-generated. 15M search queries are randomly selected and used to crawl the Web over a period of 7 months – May 2011 to November 2011. After applying the threshold model of dialect identification, the total size of YADAC is 6M wordform tokens and 457K wordform types. It is distributed as 41% from online knowledge market services, 32% from microblogs and 27% from blogs and forums.

- **COLABA** (Egyptian, Iraqi, Levantine, and Moroccan) [17] Cross Lingual Arabic Blog Alerts (COLABA) is a large effort to create resources and processing tools for Dialect Arabic Blogs. It is an initiative to process Arabic social media data such as blogs, discussion forums, chats, etc. Given that the language of such social media is typically DA, one of the main objective of COLABA is to illustrate the significant impact of the use of dedicated resources for the processing of DA on NLP applications. Accordingly, Information Retrieval (IR) was chosen as the main testbed application to process DA.

- **Arabic corpus for Egyptian tweets** (Twitter-based Egyptian Arabic) [18] This is an Arabic corpus for Egyptian tweets and it covered a blend of different general topics discussed on Twitter. The corpus contains 22,834 tweets complied over the period from May 2011 to December 2011. It is a subset of the microblog portion of YADAC. This corpus uses a function-based annotation scheme in which words are labeled based on their grammatical functions rather than their morpho-syntactic structures given that these two do not necessarily map. While a standard morpho-syntactic scheme makes comparisons with other work easier.

- **Multi-Dialectal Corpus of Arabic** (Arabic Dialects) [11] a multi-dialectal corpus of Arabic is based on the geographical information of tweets for classification. The information was collected based on the user locations around Arab countries, and extracted tweets that have Dialect word(s). A total of 175 million tweets were collected in March 2014. By excluding

tweets with non-deterministic user locations, a 62M tweets that have deterministic mappings between user locations and Arab countries were retained. Furthermore, these tweets were filtered based on dialectal words to extract a 6.5M tweets (i.e. 3.7% of the original tweets); in which 3.99M (61%) were from SA, 880K (13%) from EG, 707K (11%) from KW, 302K (5%) from AE, 65k (2%) from QA, and the remaining (8%) from other countries such as Morocco and Sudan.

- **AraSenTi-Tweet** (Saudi Dialect) [6] A Corpus for Arabic Sentiment Analysis of Saudi Tweets. The corpus consists mainly of tweets written in MSA and the Saudi dialect. It contained around 2.2 million tweets and was used to generate an Arabic corpus of tweets. A corpus of Saudi tweets was extracted from the datasets and has reached 17,573 tweets The corpus was manually annotated for sentiment and labelled with four labels for sentiment: positive, negative, neutral and mixed. Baseline experiments were conducted to provide benchmark results for future work.

- **SANA** (Arabic Dialects) [19] is a large-scale, multi-genre, multi-dialect multilingual lexicon for the **subjectivity and sentiment analysis** of the Arabic language and dialects. Language use varies across genres and SANA caters for that fact by encompassing lexica derived from four main genres: Online newswire, chat turns, Twitter tweets, and YouTube comments. In addition to Modern Standard Arabic (MSA), where most NLP efforts have been focused for the past few years, SANA also covers both Egyptian Dialectal Arabic (EDA) and Levantine Dialectal Arabic (LDA), along with providing English glosses.

- **TunDiaWN** (Tunisian Dialect) [20] A corpus-based approach to create WordNet resource for Tunisian dialect, which deviates from the strategies commonly adopted. Tunisian dialect (TD) textual data collect consists in producing MultiTD corpus (Multi-source Tunisian dialect corpus) which gathers TD texts from many sources: social networks (Twitter, Facebook, etc.), written pieces of theater, dictionaries, transcriptions of spontaneous speech, etc.

- **AWATIF** (MSA) [21] is a multi-genre corpus of MSA labeled for subjectivity and sentiment analysis (SSA) at the sentence level. This corpus is labeled using both regular and crowd sourcing methods under three different conditions with two types of annotation guidelines. The data was collected from three different sources: (1) 2855 sentences were collected and labeled

from Part 1 V 3.0 (ATB1V3) of the Penn Arabic TreeBank (PATB); (2) 1019 sentences were harvested and labeled from 30 Wikipedia Talk Pages (WTP); (3) 1508 sentences were crawled and labeled from web forum (WF) that comprises 2532 threaded conversations from 7 WFs. AWATIF is expected to help bridge a gap in research that it can be exploited for building genre-nuanced SSA systems for Arabic. It is also expected to help uncover how a MRL like Arabic can be handled in the context of social meaning extraction tasks like that of SSA.

- **Saudi Twitter Corpus** (Saudi Dialect) [22] Saudi Twitter Corpus for Sentiment Analysis is partially bridges this gap due to its focus on one of the Arabic dialects which is the Saudi dialect. This corpus presents annotated data set of 4700 for Saudi dialect sentiment analysis. The purpose of this work is to present the first Saudi annotated corpus. This will be achieved by reporting a procedure of manual corpus annotation. This corpus includes data from Twitter and covers several domains such as sport, economy, and politics. The intention of this corpus is to create the first reliably annotated Twitter data for the Saudi dialect which will be subsequently released to the LREC community as part of this submission.

- **MD-ArSenTD** (Egypt and the United Arab Emirates) [23] The Multi-Dialect Arabic Sentiment Twitter Dataset is composed of tweets collected from 12 Arab countries (KW, SA, QA, UAE, Jordan, Lebanon, Palestine, Syria, Algeria, Morocco, Tunisia, Egypt) and annotated for sentiment and dialect. The Twitter4J API [24] was used to collect 470K tweets, starting from 3/1/2017 until4/30/2017. Then, the target size of the MD-ArSenTD was set to 14,400 tweets and, thus, 1200 tweets were selected and annotated for each country. This dataset will help the Arabic NLP research community understand the specificities of Arabic tweets by providing insights into Twitter's topics, dialects and writing styles in the different Arab countries. The implementation was only performed on tweets from Egypt and the United Arab Emirates (UAE), with the aim of discovering distinctive features that may facilitate sentiment analysis. These models are based on feature engineering and deep learning, and have already achieved state-of-the-art accuracies in English sentiment analysis. Results indicate the superior performance of deep learning models, the importance of morphological features in Arabic NLP, and that handling dialectal Arabic leads to different outcomes depending on the country from which the tweets are collected.

- **SANAD** (Arabic Corpus) [25] A large Single-labeled Arabic News Articles Dataset (SANAD) of textual data collected from three news portals. The dataset is a large one consisting of almost 200k articles distributed into seven categories that are offered to the research community on Arabic computational linguistics. SANAD is composed of three main datasets scraped from three news portals, which are AlKhaleej, AlArabiya, and Akhbarona. It is made public and freely available online.

- **KACST** (Arabic Corpus) [2] It is the King Abdulaziz City for Science and Technology (KACST) Arabic corpus, which was designed and created to overcome the limitations of existing Arabic corpora. The design of the distribution of the corpus materials across all time periods, mediums, domains, and topics is influenced by the information and knowledge production available across those time periods as well as the available text mediums. The main source of the KACST Arabic Corpus texts is the Internet. Several Arabic web sites provide free-to-download, machine readable format texts covering numerous mediums. The corpus texts have been collected from numerous sources and contains more than 731 million words from 869,800 texts. The main purpose of the KACST Arabic Corpus project is to develop a free access, KACST Arabic corpus design and construction large-sized, and sufficiently diverse Arabic corpus to represent the many varieties of Arabic language across three main dimensions: time, region, and genre. Such a corpus could be used for different research interests, beginning with linguistic studies at various levels and extending to the development of NLP applications.

- **1.5 Billion Arabic Corpus** (Arabic Corpus) [26] A contemporary linguistic corpus for Arabic language that includes more than five million newspaper articles. It has over five million articles from ten news sources. The total number of words exceeds 1.5 billion words, and the total number of unique words exceeds 3.3 million words. The data were collected from newspaper articles in ten major news sources from eight Arabic countries, over a period of fourteen years. It should be noted that the news websites crawling was done between December 2013 and June 2014. The main purpose for creating this corpus, is to have a free tool for Arabic language available for researcher. It is made specifically for work in the field of information retrieval, or natural language processing. The corpus is not limited to one subject. It is multi-topic news corpus covering Politics, literature, arts, technology, sports, economy, culture, and many other

subject matters. It is also, a good representation of Arabic language. It covers a period of fourteen years and eight countries. These countries have a very large portion of Arabic native speakers. Finally, all ten sources used in creating the corpus are well represented.

- **Arabic Text Corpus** (Arabic Corpus) [27] An Arabic text corpus for Arabic variety detection which includes more than 233.000 words built from three varied sources of Arabic language: Quranic text, Classical Arabic text, and Modern Arabic text. The Arabic Quranic text has been downloaded from the Tanzil project web site. It is one of the most important Quranic text verification projects. The modern Arabic fragments have been taken from Corpus of Contemporary Arabic. The corpus consists of 1,500,000 words split into 14 classes: autobiography, children stories, computer, economics, education, health, interview, politics, religion, sciences, short stories, sociology, sports, and tourism. It is an open source corpus developed at Leeds University by Latifa el Sulaiti. The classical Arabic text has been extracted from InAra Arabic corpus. Even if it is designated for intrinsic plagiarism detection but the majority of its content is ancient Arabic files. Ancient Arabic files have been selected manually. Files taken from the original corpus are classified into 8 classes: religion, geography, and travel, history, sociology, literature, philosophy, and science. This corpus will be freely available online for researchers soon and any comments from researchers are welcome for future improvements to meet their needs for deeper research on Arabic texts and its different varieties.

- **JCCA Corpus** (Arabic Corpus) [9] The Jordan Comprehensive Contemporary Arabic Corpus is a 100-million-word corpus that is balanced, annotated, comprehensive, and representative of contemporary Arabic as written and spoken in Arab countries today. This corpus contains slightly more than 100-million words of the same text types, domains, and genres. The corpus contains 87% of texts from written sources and 13% of transcribed spoken language. The written part includes texts from Applied Sciences, Arts, Belief and Thought, Commerce and Finance, Imaginative works, Leisure, Natural and Pure Sciences, Social Sciences, and World Affairs. The spoken sub-corpus includes transcripts of Spontaneous Conversations (4.2%) and Context-Governed Spoken Language (6.2%) from the categories of Educational/Informative, Business, Public/Institutional, and Leisure. The corpus was automatically annotated both morphologically and syntactically. A sample of one million words was manually and semi-manually verified; it

was additionally annotated for sentiment and glossed in English. DIWAN was used to annotate a one-million-word snapshot of the corpus. DIWAN is a dialectal word annotation tool, but we upgraded it by adding a new tag-set that is based on traditional Arabic grammar and by adding the roots and morphological patterns of nouns and verbs.

- **KSUCCA** (Arabic Corpus) [28] The King Saud University Corpus of Classical Arabic (KSUCCA) is a classical Arabic corpus that consists of around 50 million words. The corpus includes texts of six genres, namely religion, linguistics, literature, science, sociology, and biography. The main purpose of KSUCCA was to be used for studying the distributional lexical semantics of words in the holy Quran. It is part of ongoing research that attempts to study the meanings of words used in the holy Quran, through analysis of their distributional semantics in contemporaneous texts.

- **BRAD** (MSA and Dialects Corpus) [29] Book Reviews in Arabic Dataset (BRAD) is a corpus used for sentiment analysis and machine language applications. It comprises of almost 510,600 book records collected from GoodReads application. BRAD's 510,598 Arabic-reviews are made for 4993 books (authored by 2043 writers) contributed by 76530 users. It should be noted that the dataset has reviews expressed in Modern Arabic Standard (MSA) as well as dialects. Each record corresponds for a single review and has the review in Arabic language with a total of around 2,781,805 sentences. The balanced clean subset contains 156,506 reviews. Each review is annotated with a scale from 1 to 5.

- **OSAC** (MSA) [30] Open Source Arabic Corpora (OSAC) is a free accessible Arabic corpus that contains about 18M words and about 0.5M district keywords after stopwords removal. It was collected from multiple websites to includes 22,429 text documents. Each text document belongs to 1 of 10 categories (Economics, History, Entertainments, Education & Family, Religious and Fatwas, Sports, Heath, Astronomy, Low, Stories, Cooking Recipes). This corpus has been used to address the impact of text preprocessing on the Arabic text classification.

- **Tashkeela** (Arabic Corpus) [31] A corpus collected from freely published texts in ancients books, these books had been rewritten and vocalized by volunteers manually, to ensure that words are vocalized. This corpus can be used as a linguistic resource tool for natural language processing

such as automatic diacritics systems, dis-ambiguity mechanism, features and data extraction. It is a collection of Arabic vocalized texts, that is freely available, and covers modern and classical Arabic language. The Data contains over 75 million of fully vocalized words obtained from 97 books, structured in text files. The corpus is collected mostly from Islamic classical books, and using semi-automatic web crawling process. The Modern Standard Arabic texts crawled from the Internet represent 1.15% of the corpus, about 867,913 words, while the most part is collected from Shamela Library, which represent 98.85%, with 74,762,008 words contained in 97 books.

- **ANT Corpus** (MSA) [32] An online Arabic corpus of news articles (ANT Corpus) is collected from RSS Feeds. Each document represents an article structured in the standard XML TREC format. This Corpus can be used for Text Classification to assign to each article its accurate predefined category. Documents were collected from a Tunisian news website, "Jawhara-FM", which is updated daily. It contains about 10,000 articles from 9 categories, including +865,500 word in which around 144 word is per article. Each document is described by meta-data including its identifier, the publishing date, the document's title, the access link, the source, the author, the abstract, the main text and the category which is considered as the document's class.

- **OPUS** (Multilingual Corpus) [33] A growing language resource of parallel corpora and related tools. The focus in OPUS is to provide freely available data sets in various formats together with basic annotation to be useful for applications in computational linguistics, translation studies and cross-linguistic corpus studies. This corpus was compiled as additional data sets on a large scale in order to provide data for many other, often under-resourced languages and domains. The overall goal of the OPUS project is to make parallel resources freely available, especially emphasizing the support of low density languages. The largest domains covered by OPUS are legislative and administrative texts (mostly from the European Union and associated institutions), translated movie subtitles and localization data from open-source software projects. There is also a substantial amount of newspaper texts and some other smaller collections from various on-line sources. OPUS has been extended by several large collections; such as TED [33] which is a parallel corpus of TED talk subtitles provided by CASMACAT[1]. The files are

[1] http://www.casmacat.eu/corpus/ted2013.html

originally provided by the Web Inventory of Transcribed and Translated Talks[2] and contains 15 languages and a total number of 3.81M sentence fragments. Another example is MultiUN [34], which is a collection of translated documents from the United Nations with 6 languages and a total number of 81.41M sentence fragments. One last, and mostly known dataset is OpenSubtitles [35], which is a collection of translated movie subtitles from http://www.opensubtitles.org/. This is a slightly cleaner version of the subtitle collection using improved sentence alignment and better language checking with a 62 languages and a total of 3.35G number of sentence fragments.

- **HARD** (Arabic and Dialectal Corpus) [36] Hotel Arabic-Reviews Dataset (HARD) is one of the largest Book Reviews in Arabic Dataset for subjective sentiment analysis and machine language applications. It adds to the large dataset BRAD, and comprises of 490,587 hotel reviews. The dataset is a collection of Arabic reviews compiled from Booking.com; a website that specializes in online accommodation booking. Collected reviews are structured as follows: A rating out of 10 of the accommodation, a title of the review, a positive aspect(s) of the accommodation, a negative aspect(s) of the accommodation, the reviewer's username, and country of residence. Reviews are available in many languages, including MSA, and can be filtered to the user's preference. However, the majority of the dialectal reviews are in Gulf dialects. The data was collected and originally organized in the following columns: Hotel name, rate (reviewer's rating out of 10), user type (family, single, couple), room type, nights (number of nights stayed), review title, positive review, negative review. The main objective was to make HARD datasets publicly available and perceived as benchmark in the field of Arabic computing.

- **Habibi** (Arabic and Dialectal Corpus) [37] An open-source Arabic song lyrics dataset from 18 different Arab countries. Habibi corpus comprises of more than 30,000 Arabic songs' lyrics in 6 Arabic dialects for singers from 18 different Arabic countries (Egypt, SA, Lebanon, Iraq, Sudan, KW, Syria, UAE, Morocco, Tunisia, Yamen, Jordan, QA, Bahrain, Algeria, Oman, Palestine, Libya). These songs are segmented into more than 500,000 sentences (song verses) with more than 3.5 million words that are free from spam words, ads, hashtags or emojis resulting in a clean and noise-free dataset. The resource provides a rich and diverse venue for researchers working on

[2] https://wit3.fbk.eu

Dialects Identifications and Authorship Attribution, and available under License Creative Commons Attribution-NonCommercial

- **Arabic Sentiment Analysis Dataset** (Saudi dialect) [38] A dataset built from tweets discussing several social issues in Saudi Arabia. These issues include changes that were implemented by the country as part of a newly established vision, known as Saudi Arabia Vision 2030. The constructed dataset was manually annotated according to the sentiment conveyed in the text. To achieve the best sentiment classification accuracy, several procedures were implemented within the proposed framework including light stemming, feature extraction, parameter optimization and feature-set reduction. This corpus was available for download[3] and contained 15,149 words.

- **SUAR** (Saudi dialect) [10] The corpus SUAR (SaUdi corpus for NLP Applications and Resources) consists of 104,079 words collected from different online resources (Blogs, Forum, Instagram, Twitter, WhatsApp, YouTube). The corpus was automatically annotated using the MADAMIRA tool, after which it was manually inspected to validate the resulting analysis. And this corpus conducts a pilot study to explore possible directions to facilitate the morphological annotation of the Saudi corpus.

- **MADAR** (Arabic Dialects) [39] The Multi Arabic Dialect Applications and Resources (MADAR) is the first is a large parallel corpus of 25 Arabic city dialects in the travel domain. The goal of MADAR is to create a large number of dialects, a unified framework with common annotation guidelines and decisions, and targeting applications of Dialect Identification (DID) and Machine Translation (MT). The MADAR Corpus is a collection of parallel sentences covering the dialects of different cities from the Arab Regions (including Aleppo, Alexandria, Algiers, Amman, Aswan, Baghdad, Basra, Beirut, Benghazi, Cairo, Damascus, Doha, Fes, Jeddah, Jerusalem, Khartoum, Mosul, Muscat, Rabat, Riyadh, Salt, Sanaa, Sfax, Tripoli, and Tunis.), in addition to English, French, and MSA. This corpus is created by translating selected sentences from the Basic Traveling Expression Corpus (BTEC) [17] in French and English to the different dialects. The MADAR Corpus will be made available soon to the research community under a non-

[3] https://www.kaggle.com/snalyami3/arabic-sentiment-analysis-dataset-ss2030-dataset

commercial license. While the Arabic portions of the corpus will be provided but the English corpus will not be available due to copy right restrictions. The latest release contains two datasets: (1) Corpus-26: a set of 2,000 sentences and translated to all 25 city dialects (each of these sentences has 25 corresponding parallel translations), in addition to MSA; (2) Corpus-6: a set of 12,000 sentences translated to the dialects of five selected cities: Doha, Beirut, Cairo, Tunis, and Rabat, in addition to MSA.

- **Dialectal Saudi Twitter Corpus** (Saudi Dialect) [4] This corpus presents a dialectal Saudi corpus that contains 207,452 tweets generated by Saudi Twitter users. In addition, a comparison between the Saudi tweets dataset, Egyptian Twitter corpus and Arabic top news raw corpus, representing MSA in various aspects, such as the differences between formal and colloquial texts was carried out. This corpus used Twitter API to compile Twitter users' tweets for building the current corpus. All the tweets were collected in 2017, generated by 101 Saudi users. The corpus contains 101 UTF-8 text files and the total number of the tweets is 207,452.

From Table *1*, it can be seen that the SD corpus is a domain that needs further contributions and that the available Arabic corpora are not enough to cover the gaps. First, MSA language dominates most of the large datasets, while SD text is not introduced in many of them. This creates an issue because SD differs from MSA syntactically, morphologically, and phonologically [7]; especially since people are increasingly using Dialectal Arabic, while MSA is limited to formal resources. Regardless of the similarity between the two languages, more SD text needs to be collected to allow Arabic NLP models to process such text.

The second issue in the literature that can be highlighted is that even though there have been attempts to collect SD text, these attempts have been limited in size. As summarized in Table 1, SD corpora only range from thousands of words to a couple million words. Although 2~3M words might sound a large number, it is not enough for NLP tasks that target dialectal language. This is because, for example, DS does not have standard orthographies, which makes processing it more challenging and thus requires a very large amount of data.

Table 1. Existing Arabic Corpus

Corpus	Language	Source	Description	Size	Year	Accessibility	Reuse
PATB	MSA	Newswire	A large-scale annotated Arabic corpus based on newswire from different countries.	+ 1.3 M words	2004 - 2011	Commercial	Private
Gumar	GA MSA	Forum novels	A morphologically semi-automatically annotated Gulf Arabic (GA) corpus.	112 M words	2016	Restricted	Private
Gumar Emirati	EA	Forum novels	Morphologically manually annotated corpus of Emirati Arabic.	200K words	2019	Restricted	Private
Curras	Palestinian	Facebook Twitter Blogs Forums Documents TV Shows	The first morphologically annotated corpus of the Palestinian Arabic dialect	43K Words	2016	Public	Personal
YADAC	Egyptian	Twitter Blogs Forums	A multi-genre Dialectal Arabic corpus that is compiled using Web data from microblogs, blogs/forums and online knowledge market services in which both questions and answers are user-generated	6M wordform tokens	2012	Not Available	Not Available
COLABA	Egyptian Iraqi Levantine Moroccan	Blog Alerts	An Arabic corpus that was built for NLP resources covering four Arabic dialects	--	2010	Private	Private
Arabic corpus for Egyptian tweets	Egyptian	Twitter	An Arabic corpus for Egyptian tweets of different general topics discussed on Twitter	22,834 tweets	2012	Private	Private
Multi-Dialectal Corpus of Arabic	MSA Arabic Dialectal	Twitter	A multi-dialectal corpus collected based on Twitter geographical information to collect a dialectal corpus for different Arab countries	62 M tweets	2014	Private	Private
AraSenTi-Tweet	MSA Saudi	Twitter	A Corpus that is manually annotated for Arabic Sentiment Analysis of Saudi Tweets	2.2 M	2017	Private	Private
SANA	English	Newswire	A large-scale, multi-genre,	224,564	2014	Private	Private

Corpus	Language	Source	Description	Size	Year	Accessibility	Reuse
	Egyptian Levantine MSA	chat turns Twitter YouTube	multi-dialect multilingual lexicon for the subjectivity and sentiment analysis of the Arabic language and dialects	entries (sentence)			
TunDiaWN	Tunisian	Twitter Facebook TripAdviser Theater Dictionaries Transcripts	A corpus to create WordNet lexical resource for Tunisian dialect language.	32,848 words	2014	Private	Private
AWATIF	MSA	PATB Part1 Wikipedia Web forum	An MSA multi-genre corpus collected from different resources and labeled for subjectivity and sentiment analysis	5,382 sentence	2012	Private	Private
Saudi Twitter Corpus	Saudi	Twitter	A Saudi tweets corpus that is annotated for sentiment analysis.	4,700 tweets	2016	Private	Private
MD-ArSenTD	KW, SA, QA, UAE, Jordan, Lebanon, Palestine, Syria, Algeria, Morocco, Tunisia, Egypt, MSA	Twitter	The first Multi-Dialect Arabic Dataset that is annotated for both sentiment and dialects.	14,400 tweets	2017	Private	Private
SANAD	MSA	News	A large collection of Arabic news articles that can be used in different Arabic NLP tasks such as Text Classification and Word Embedding	200K articles	2019	Public	Public
KACST	Classical Arabic MSA	Website crawling	King Abdulaziz City for Science and Technology (KACST) Arabic corpus	+731 million words	2014	Restricted	Private
1.5 Billion Arabic Corpus	Classical Arabic	Newspapers	A contemporary linguistic corpus for Arabic language	+1.5 billion	2014	Private	Private

Corpus	Language	Source	Description	Size	Year	Accessibility	Reuse
				words			
Arabic text corpus	Classical Arabic MSA	Quran Contemporary Arabic corpus InAra Arabic corpus	An Arabic text corpus for Arabic variety detection	1,500,000 words	2018	Private	Private
JCCA	MSA	Newspapers Books Online sources	Jordan Comprehensive Contemporary Arabic Corpus that is balanced, annotated, comprehensive, and representative of contemporary Arabic as written and spoken in Arab countries today	100 million word	2019	Private	Private
KSUCCA	Classical Arabic MSA	Almaktabah Alshamilah website	The King Saud University Corpus of Classical Arabic	+50M words	2014	Public	Public
BRAD	MSA Dialectal	Book reviews	Book Reviews in Arabic Dataset used for sentiment analysis and machine language applications	2,781,805 sentences	2016	Public	Public
OSAC	MSA	Websites	A free accessible Arabic corpus	+18M words	2010	Public	Public
Tashkeela	Classical Arabic MSA	Ancients books Online Libraries	A corpus collected from freely published texts in ancients books	75M words	2017	Public	Public
ANT Corpus	MSA	Tunisian news website	An online Arabic corpus of news articles that is collected from RSS Feeds	+86500 words	2017	Public	Public
OPUS (TED)	15 language	TED talk subtitles	A parallel corpus of TED talk subtitles	3.81M sentence fragments	2013	Public	Public
OPUS (MultiUN)	6 languages	Documents	A collection of translated documents from the United Nations	81.41M sentence fragments	2010	Public	Public
OPUS (OpenSubtitles)	62 language	Movie subtitles	A collection of translated movie subtitles from opensubtitles.org	3.35G sentence fragments	2018 2020	Public	Public
HARD	MSA Dialectal	Booking.com website	Hotel Arabic reviews in Dataset for subjective sentiment	490,587 reviews	2016	Public	Public

Corpus	Language	Source	Description	Size	Year	Accessibility	Reuse
			analysis and machine language applications				
Habibi	6 languages	song lyrics	Arabic Song Lyrics corpus from 18 different Arabic countries	+3.5M words	2020	Public	Public
Arabic Sentiment Analysis Dataset	Saudi	Twitter	Collection of tweets discussing several social issues in Saudi Arabia	15,149 words	2020	Public	Public
SUAR	Saudi	Blogs Forum Instagram Twitter WhatsApp YouTube	SD corpus that is collected from multiple online resources and automatically annotated with manual validation.	104,079 words	2018	Private	Restricted
MADAR	AD English French MSA	Traveling Expression Corpus	A large parallel corpus that was translated from English and French travel corpus to a 25 Arabic city dialects	+12,000 sentence	2018	Restricted	Restricted
Dialectal Saudi Twitter Corpus	Saudi	Twitter	A dialectal Saudi corpus that was generated by 101 Saudi users.	207,452 tweets	2017	Public	Public

The third limitation of the current SD corpora is the lack of diversity. As shown in Table *1*, corpora that include SD language are usually collected from one source (mostly Twitter); only SUAR introduced corpora that include SD language are usually collected from one source (mostly Twitter); only SUAR introduced different sources, yet it only includes 104,079 words. Although Twitter can be a rich source of data, it can show only one social aspect of the community. YouTube and websites are as frequently used as Twitter, and they can show different aspects, especially when targeting resources with diverse titles and topics.

In conclusion, regardless of the rich literature on Arabic corpora, SD corpora are still lacking and need further contributions. Thus, the following section will propose a new model to collect and build a new Saudi corpus that is large, up to date, and diverse.

2.1.2 Existing Preprocessing Tools

When collecting text data from different resources, irrelevant text and dirty data might also be collected during the process. Thus, collected data must be pre-processed and cleaned before analysis. From the literature of NLP, different preprocessing tools were developed to clean and normalize text data, as follows.

- **MADAMIRA** [40] It is a system for morphological analysis and disambiguation of Arabic text that combines some of the best aspects of two previously commonly used systems for Arabic processing, MADA [5], [41], [42] and AMIRA [43].

- **Farasa** [44] Is an Arabic segmenter tool which means insight and is defined for the precise and include separate tools that contains a set of elements and most important Dependency Parser, POS tagger, and tokenization or segmentation module, It's a more flexible usage of components compared to MADAMIRA.

- **CAMeL** [45] It is a collection of open-source tools for Arabic natural language processing in Python. CAMeL Tools currently provides utilities for pre-processing, morphological modeling, dialect identification, named entity recognition and sentiment analysis.

- **AraNLP** [46] It is a free Java-based library named "AraNLP" that covers various Arabic text preprocessing tools. Although a good number of tools for processing Arabic text already exist, integration and compatibility problems continually occur.

- **RDI MORPHO3** [47] This system uses rules in conjunction with statistics in order to build a list of possible prefix-suffix template combinations [48]. The main disadvantage of this system is that the rules are built manually which is time consuming and demanding a deep knowledge of the Arabic language.

- **Sebawai root extractor (SR)** [47] It is very similar to RDI MORPHO3 root extractor. However, it uses automatic rules rather than manual rules [49]. Rules have been obtained through training the system with a list of word-root pairs.

2.2 Data Collection

To collect as much data as possible while using diverse sources, data were collected from the available open corpora in addition to new sources on the web. Due to the difficulty of acquiring valid SD text and how expensive it is to clean such text, it was also decided to introduce MSA text into the collected data. This is because there is some similarity between MSA and dialectical language [50] that can be useful to semantic-based tasks.

The overall statistics of the collected resources are summarized in Table 2. The total size of the collected text is 184,146,256 sentences, 1,238,539,863 words, and 26,674,484 unique words; in which 126,090,964 words were in SD language. The texts were collected from five different resources, including existing open source corpora, websites, Facebook, YouTube, and Twitter. When comparing the number of unique words from each source, as illustrated in Table 2, YouTube is the richest source of vocabularies for SD language, while pre-existing corpora are the richest for MSA language.

Table 2. Total statistics of collected data

Source	Language	No. of Unique Words	No. of Sentences	No. of Words
Pre-Existing Datasets	MSA	14,761,449	164,615,349	1,026,315,738
	SD	383,690	515,782	4,144,771
Web Crawling	MSA	1,938,577	722,340	32,529,087
	SD	112,025	547,433	2,524,225
Facebook	MSA	988,479	1,356,175	21,640,486
	Mixed	1,368,222	2,302,685	27,122,297
YouTube	SD	5,453,422	12,565,229	106,599,767
	Mixed	251,193	247,937	4,367,554
Twitter	SD	1,320,775	1,231,816	12,822,201
	Mixed	96,652	41,510	473,737
Total	MSA	17,688,505	166,693,864	1,080,485,311
	SD	7,269,912	14,860,260	126,090,964
	Mixed	1,716,067	2,592,132	31,963,588
	SUM	**26,674,484**	**184,146,256**	**1,238,539,863**

From Table 2, it can be noticed, at some cases, the collected data included both MSA and SD in its text (i.e., mixed). This usually happens when the text includes official announcement that is written in MSA language in addition to comments written by people in SD language. The only exception is in the pre-existing corpora; where it only included either MSA or SD text. In YouTube

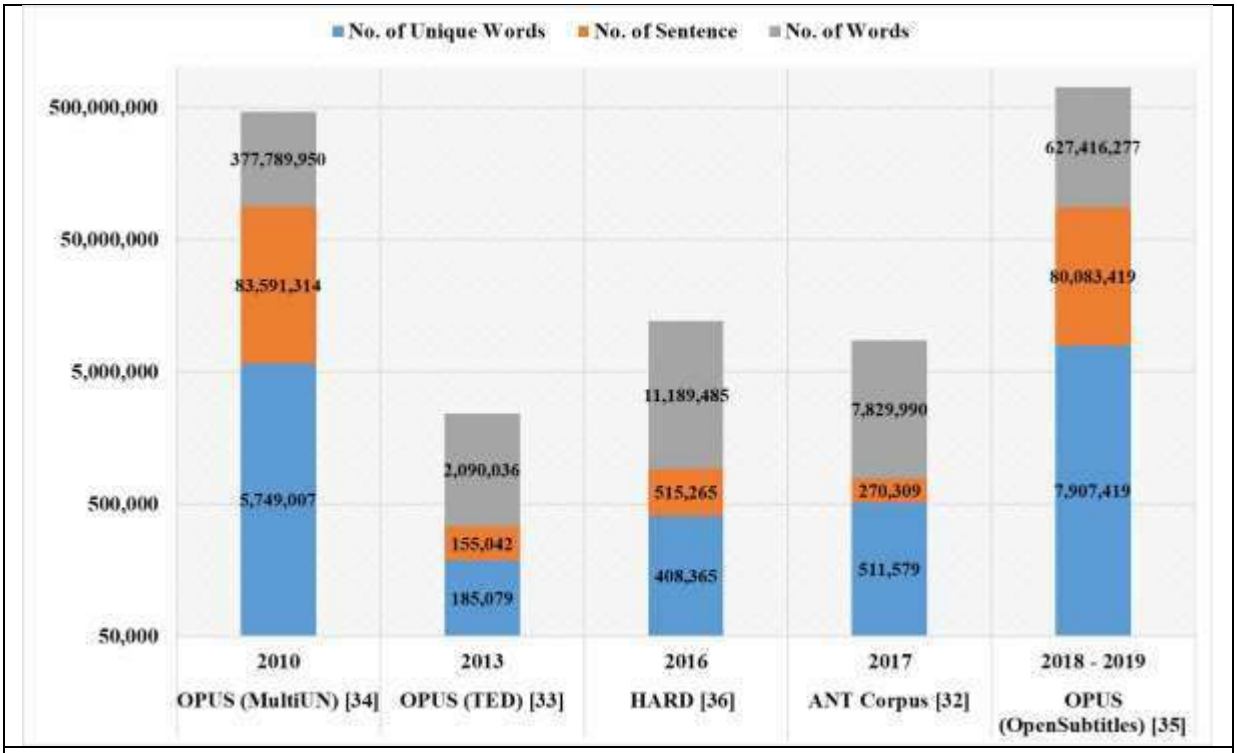
and Twitter, it was difficult to find pure MSA text. The collected text was either identified as SD or mixed of SD and MSA. This is because both YouTube and Twitter are not formal platforms as the case with news websites.

One last interesting observation, from Table 2, is that Facebook was not a rich source of SD text. This is because Saudis rarely use this platform for communication and, instead, use YouTube and Twitter. For that reason, most of the resources collected from Facebook were text written with MSA followed by the small number of comments in SD language (mixed with MSA). Further information on the details of each source is described in the following sections.

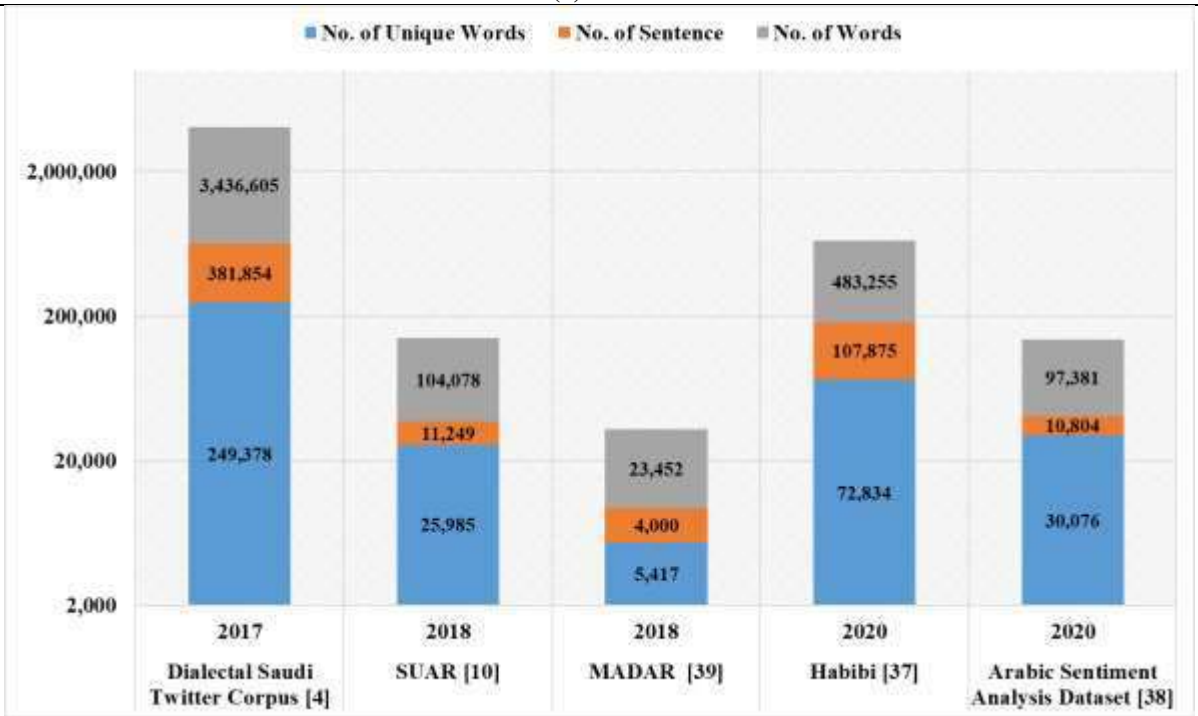
2.2.1 Pre-Existing Corpus Collection

The bibliography details of the existing corpora that was collected and its statistics is summarized in Figure 1.(a) and Figure 1.(b). Note that the information shown in this figure represent the portion collected for our corpus and not the whole corpus information. This is because some corpora include different languages that are not related of Saudi language, so it was eliminated in the collection phase.

Starting with MSA corpora, illustrated in Figure 1.(a), it is clear that the literature is rich with MSA text that can be reused. OPUS corpora is the richest source of MSA text and included data from 2010 to 2019. Thus, it MSA corpora can include vocabularies related to past events. On the other hand, when considering the SD corpora illustrated in Figure 1.(b), the text is more recent where it is dated between 2017 to 2020. Yet, collected SD corpora are limited in size comparing to the MSA, in which they ranged between 23,000 words to 3M words (5000~249000 unique words) only. Thus, more text need to be collected from to cover recent events and enrich the literature of SD text.



(a) MSA



(b) SD

Figure 1. Collected existing corpora; (a) statistic of MSA corpora, (b) statistic SD corpora

In addition to the size limitation, collected corpora are still limited in domains. Figure 2, illustrated the categories covered by the collected corpora, which is only 10 domain. However, to build a diverse and scalable corpus more categories need to be considered.

Figure 2. Categories of collected pre-Existing corpora

2.2.2 Website Crawling

In addition to existing corpora, it was also important to collect relatively new text from the web. Thus, web crawling technique was used to collect text from different websites and news outlet. In particular, ten different websites were crawled. These websites are rich in content and cover news, medicine topics, or sell products. Only Arabic pages were extracted and then manually scanned to identify if it has MSA or SD language or both. After crawling all pages, XML tags were removed from the text and only the body of the pages were preserved.

A special crawler tool was used to deal with every websites' selectors to get the required word.

- Since each website has its own selectors. Also, in the same website's pages some of them has different selectors, most of existing crawlers get all sentences on the page without considering the repetition of the sentences (e.g., related links the same page)
- Our crawler goes through every link in the website
- It extracts the content according to given selectors
- We have saved the body according to the language type, or if there is no body (in case the page

is main page or the page just contain links to the pages contain required sentences)

- We check the extracted content and then fix the issues for every website's results
- The last step we merge the bodies from the result to a one file.

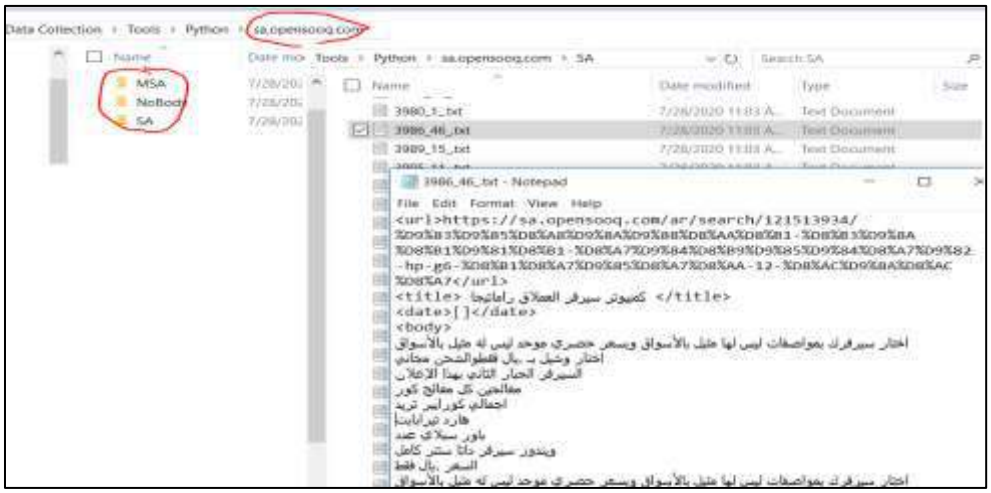


Figure 3. Sample of Website Crawling

The bibliography details of the crawled websites and its statistics is summarized in Table 3. Note that the information shown in this table represent the data before cleaning. From Table 3, it can be noticed that medicine and news outlets included only MSA, which is expected since these platforms are formal. However, in commercial platform that sells product, the text extracted was SD because it included the seller description about their products in addition to the consumers' comments on the products, in which they wrote in the spoken language.

Table 3. Websites used for crawling

Website	Language	Category	No. of Sentences	No. of Word	No. of Unique Words
https://www.webteb.com/	MSA	Medicine	476,298	8,929,536	131,933
https://ajel.sa/	MSA	News	1,163	34,691	10,614
http://aletq.com/	MSA	News	92,609	961,372	163,414
http://al-madina.com/	MSA	News	7,752	273,945	40,349
https://www.bbc.com/arabic	MSA	News	33,255	968,393	101,730
https://twasul.info/	MSA	News	11,150	316,322	56,733
https://ar.wikipedia.org/	MSA	General	100,113	21,044,828	1,433,804
https://sa.opensooq.com/	SD	Economy	363,095	1,698,155	53,483
https://haraj.com.sa/	SD	Economy	79,618	298,346	30,804
https://www.aswaqcity.com/	SD	Economy	104,720	527,724	27,738
Total			**1,269,773**	**35,053,312**	**2,050,602**

Combining all categories together, as illustrated in Figure 4, it can be seen that the total number of words collected from SD websites (+2M words) is much less than MSA sources. It was found websites can be a rich source of MSA text but not SD. This can be attributed to the fact that the young generation of Saudi's are not using website platforms as before, and social media are becoming more used, as will be emphasized in the following section.

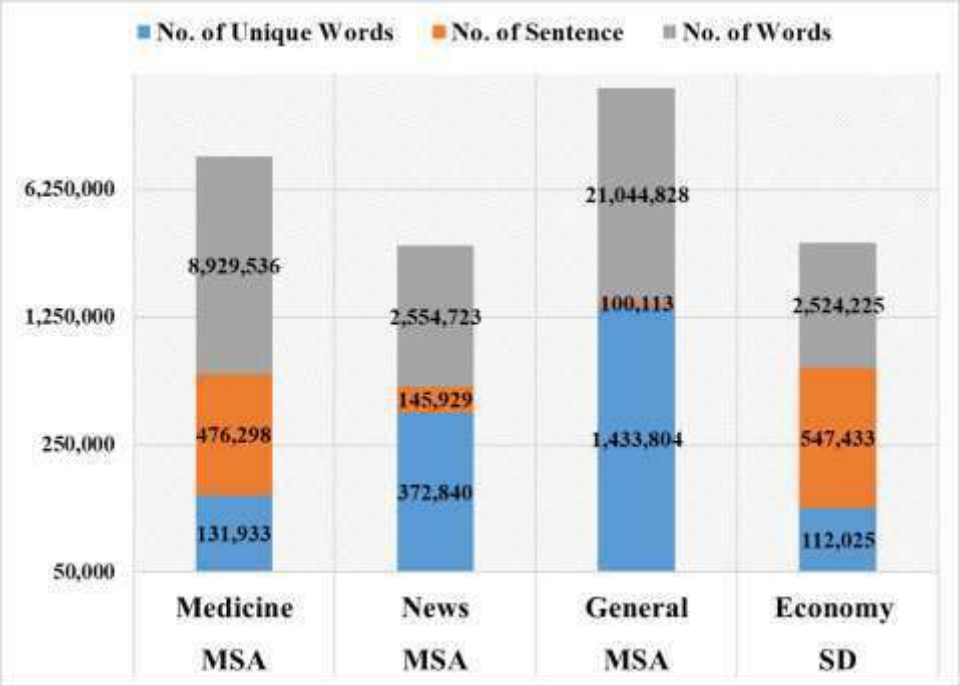


Figure 4. Website crawling statistics

2.2.3 Social Media Collection

Social media are becoming a major source of research and a rich source of information especially for dialectal text [38]. In particular, the Saudi community witnessed a massive increase in use of social media in the last 10 years [6]. For research purposes, social media platforms provided API's for developers allowing them to collect users comments and text blogs given certain criteria. In this report, it was decided to collect text from three social media platforms: YouTube, Twitter, and Facebook. The bibliography details of the collected social media accounts and its statistics is summarized in Table 4. Note that the information shown in this table represent the statistics before cleaning.

Table 4. Social media data statistics

Source	Language	No. of Sentence	No. of Words	No. of Unique Words
Facebook	Mixed	2,302,685	27,122,297	1,368,222
	MSA	1,356,175	21,640,486	988,479
		3,658,860	**48,762,783**	**2,356,701**
Twitter	Mixed	41,510	473,737	96,652
	SD	1,231,816	12,822,201	1,320,775
		1,273,326	**13,295,938**	**1,417,427**
YouTube	Mixed	247,937	4,367,554	251,193
	SD	12,565,229	106,599,767	5,453,422
		12,813,166	**110,967,321**	**5,704,615**

From Table 4, it can be noticed that SD text was differentiated better with Twitter and YouTube; while in Facebook almost 50% of the collected text were mixed and differentiating SD text was a challenge. In general, the total number of words collected from such resources exceeded +17M sentence, +173M words, and +9M unique words. Most of the collected text is SD rather than MSA. The details of each sources, and how the API was used to collect the data is shown in the following section. Mixed text, however, occurred in all platforms of social media. When comparing the sources of social media, from Table 4, it becomes clear that YouTube was the richest source of data for SD text and twitter comes afterwards.

In addition to the size of the data, the diversity is also important. From Figure 5, it is clear that the collected data is divers and cover different categories and topics. Data collected from social media have been categorized into 25 different categories. Each category included one set of data or up to five different resources. The details of the number of words in each category will be explained next, and is different depending on the source it was collected from.

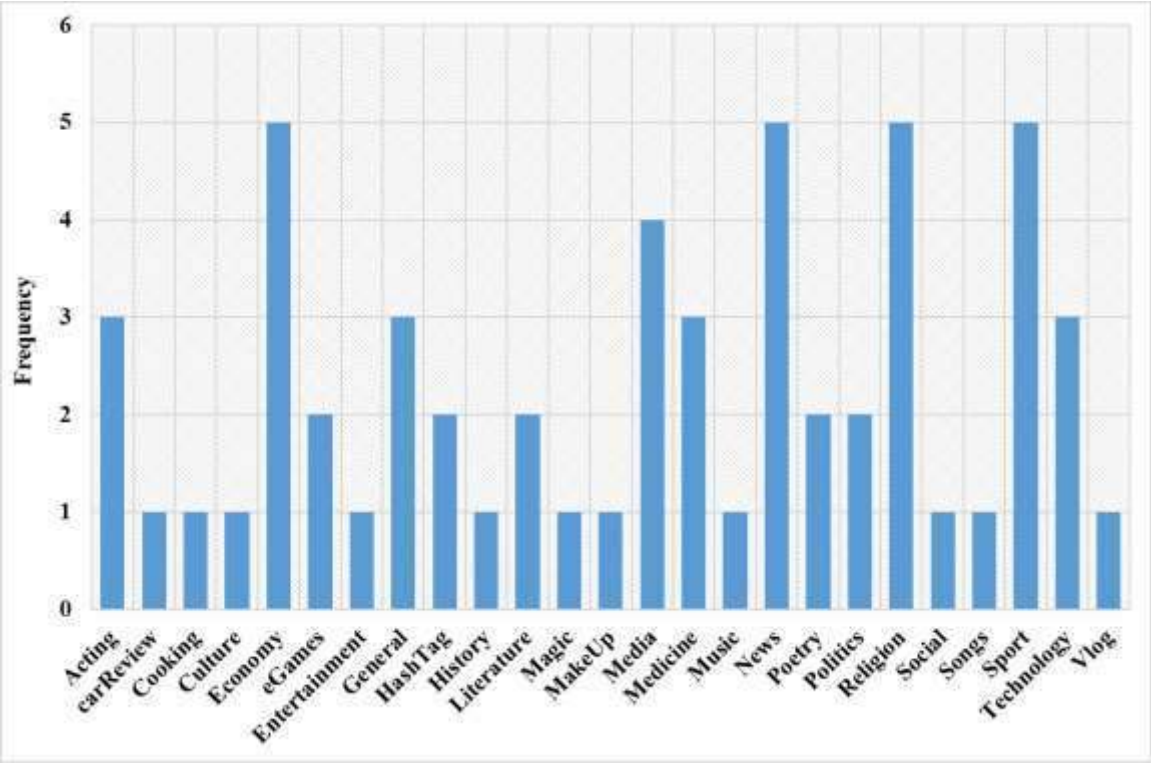


Figure 5. Categories distribution across collected social media dataset

A. YouTube

In order to collect data from YouTube, commentThread [51] API was used. This tool allows the developers to fetch the users replies and comments from their channels and videos. In order to fetch the comments from a channel or a video, an ID must be explicitly given to the commentThread in the link URL. However, a challenge that was faced with users whose name are identified as user name instead of ID. With this type of accounts, the ID does not explicitly appear in the URL so it is difficult to fetch its information.

For example, Figure 6 shows how the account URL can include the user ID; while if the URL included the name instead, then the URL would be (https://www.youtube.com/user/khalejiatv). To overcome this challenge, a script was used to convert the user URL to a channel then fetch its data. For example, the script below shows how to collects data from Khalijia Channel on YouTube:

{

"kind": "youtube#commentThread",

" allThreadsRelatedToChannelId": string,

"snippet": {

"channelId": UCsVQJ4sjopcYEQHxWnS2Spg,

"videoId": 9dgGvZiBIj4,

"topLevelComment": comments Resource,

"canReply": boolean,

"totalReplyCount": unsigned integer,

"isPublic": boolean

},

"replies": { "comments": [
comments Resource

]
}
}

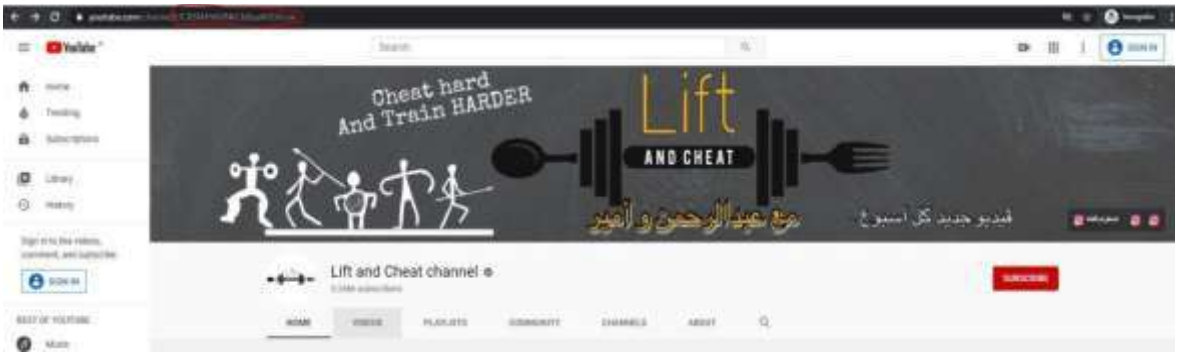


Figure 6. Sample of YouTube account

After overcoming the challenges of fetching data from YouTube, it was important to decide what channels and videos to collect from. First, Arabic channels were examined, and if these channels included videos or playlist introduced by Saudi announcer or broadcaster then it will be flagged. Then, the comment sections of the flagged videos will be examined, and if they have large amount of comments and they are written in Saudi language then the video URL will be sent to commentThread to collect the data. As a result of this process it was possible to collect a total of 110,967,321 words, 12,813,166 sentences, and 5,704,615 unique words from YouTube. As illustrated from Figure 7, the only category that could not differentiate SD from MSA is when the comments about religious content. This is because people write in their spoken language and included MSA text from Quran or Prophet saying.

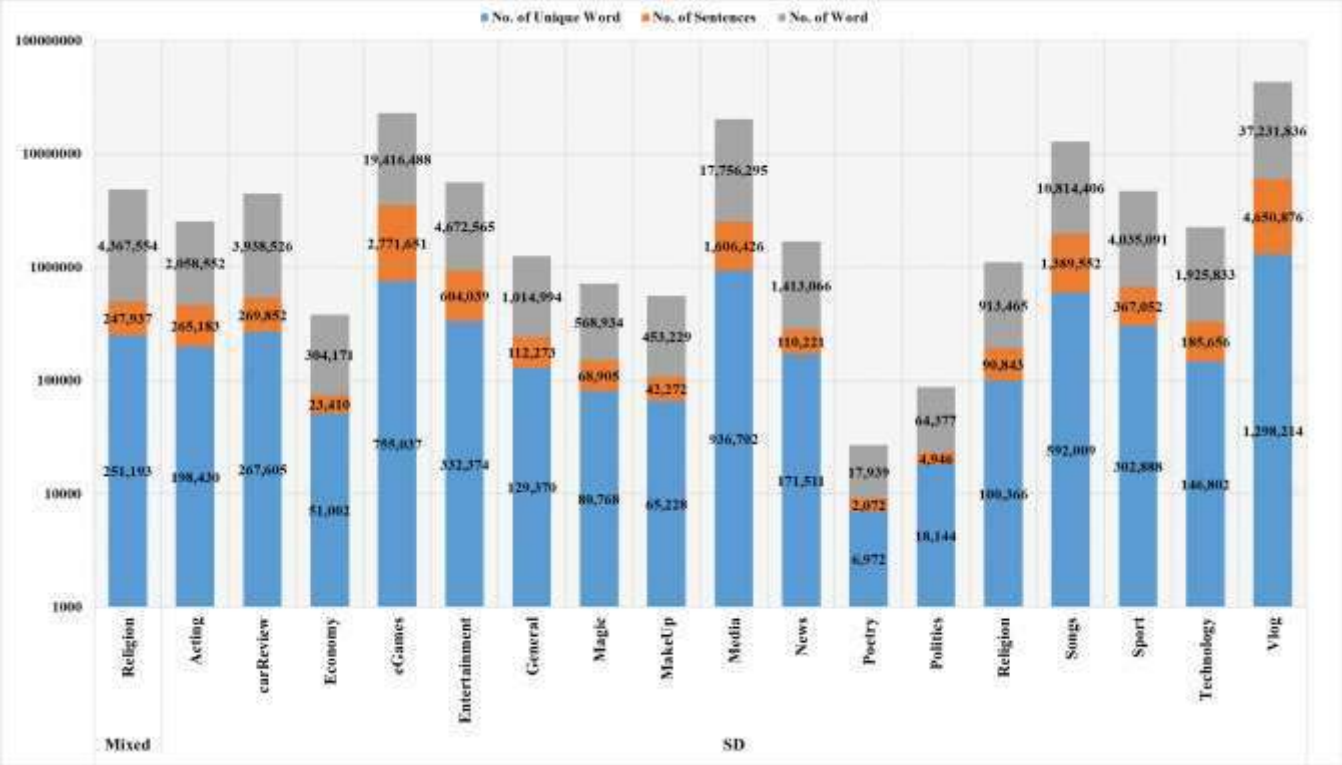


Figure 7. Collection statistics for YouTube data

Moreover, from Figure 7, one of the richest content for SD text in YouTube is the vlog style videos. In which users were very active commenting about these videos and engaging with the channels owner. On the other hand, commenting on poetry content was relatively scarce while the rest of the categories depends on how much the challenge is active.

B. Twitter Resources

Twitter provides different options to collect their data via their API. In order to collect relevant and important text for KSUSC, different criteria were used during the collection. The first criterion used is streaming using the geographical information, where SA location was used to identify Saudi tweets. Of course during this process, none SD texts were found and, thus, identified as mixed. The second criteria used is archived tweets based on important keywords (hashtags) that is related to Saudi events that happened recently; such as #المرور_السعودي or #شكرا_لكل_معلم_سعودي. A third criterion used to collect archived data is using relevant users account; which are the well-known Saudi influencers that have millions of Saudi followers and tweets. Note that, regardless of the criterion used, all tweets, re-tweet, and comments were collected during the process.

During the collection process, it was noticed that there were a lot of redundant and unnecessary data gathered in the text. This was due to the Twitter protection policy, which ensure that no one has full access to a user comments and only a sample of the comments are retrieved. Nevertheless, twitter comments had a significant importance in our task, because tweets are usually in MSA while comments are in SD. Thus, to compensate for this shortcoming, it was decided to increase the number of collected users, and minimize the collection time periods. For each user, replies and timeline are collected only once. This way, it was possible to reduce the number of request quota, while collecting more relevant SD text.

As a result of Twitter data collection, illustrated in Figure 8, the challenge of differentiating SD text from MSA is not relevant to the category but rather to the sources. If the source who published and retweeted or commented is an entity account, the collected text will be a mix between MSA and SD. Moreover, it also a challenge to categorize more that 5M words and, thus, it was labeled as general. This is because tweets are many times a reply or discussion about different topics, so many outcomes did not belong to a certain category. News category comes in second place with respect to the number of SD words (with +2M words), while it comes in first with mixed text (with +300,000 words).

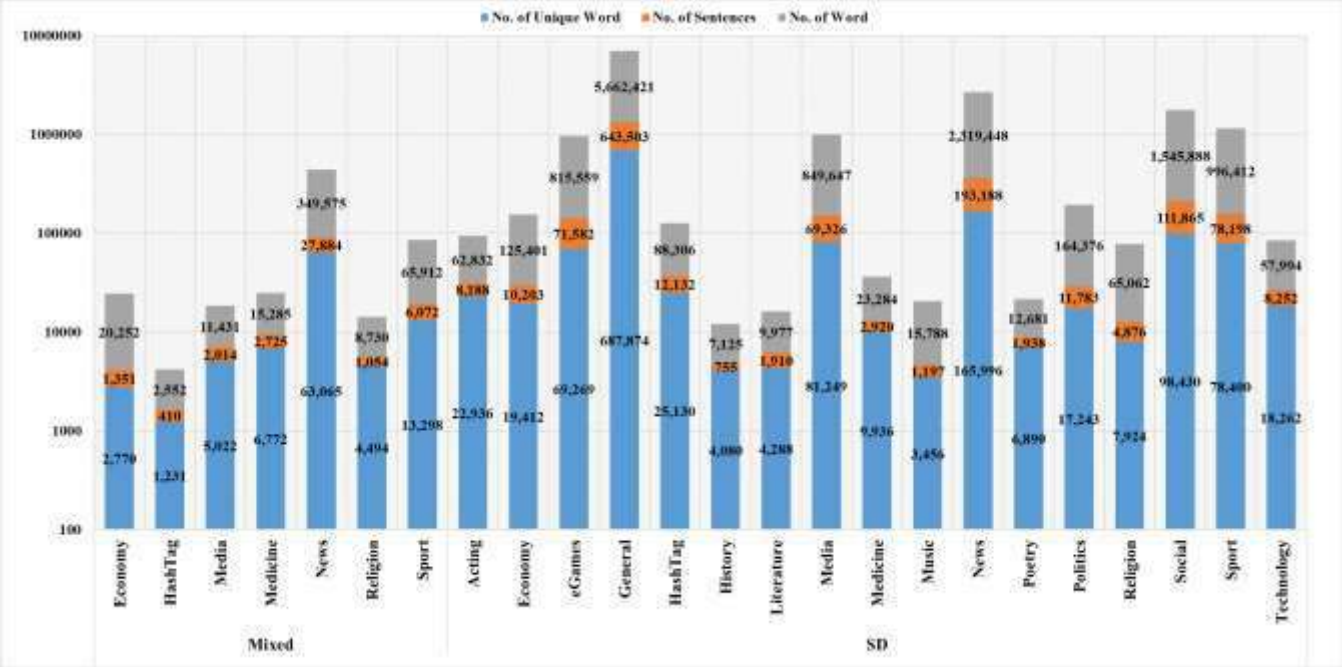


Figure 8. Collection statistics for Twitter data

C. Facebook Resources

In the collection process, it was decided to collect text also from Facebook because it is usually considered as one of the best source for collecting social data. It provides posts and comment related a given page through GraphAPI [52]. The first step was to collect pages in different domains including news, politics, culture, sport, cooking, and advertisement. FindMyFBID tool [53] was also used to convert the page URL to ID so that GraphAPI can retrieve its data. The post and its corresponding comments cannot be retrieved at once, so it was decided to first retrieve all the posts from the last three years; then, for each post a new request is sent to fetch its comments and replies.

From Figure 9, it is clear that Facebook was not popular in SA and, thus, the data collected from this platform was either MSA or mixed. As in Twitter, MSA text was relevant to users rather than the category; in which text that is published by official entities were in MSA while post written by users were a mixed of MSA with SD. Categories such as acting, technology, and culture are the poorest in SD content, while the rest of the categories are interchangeable. There was no common behavior in Facebook text, because it was not a rich source of SD language.

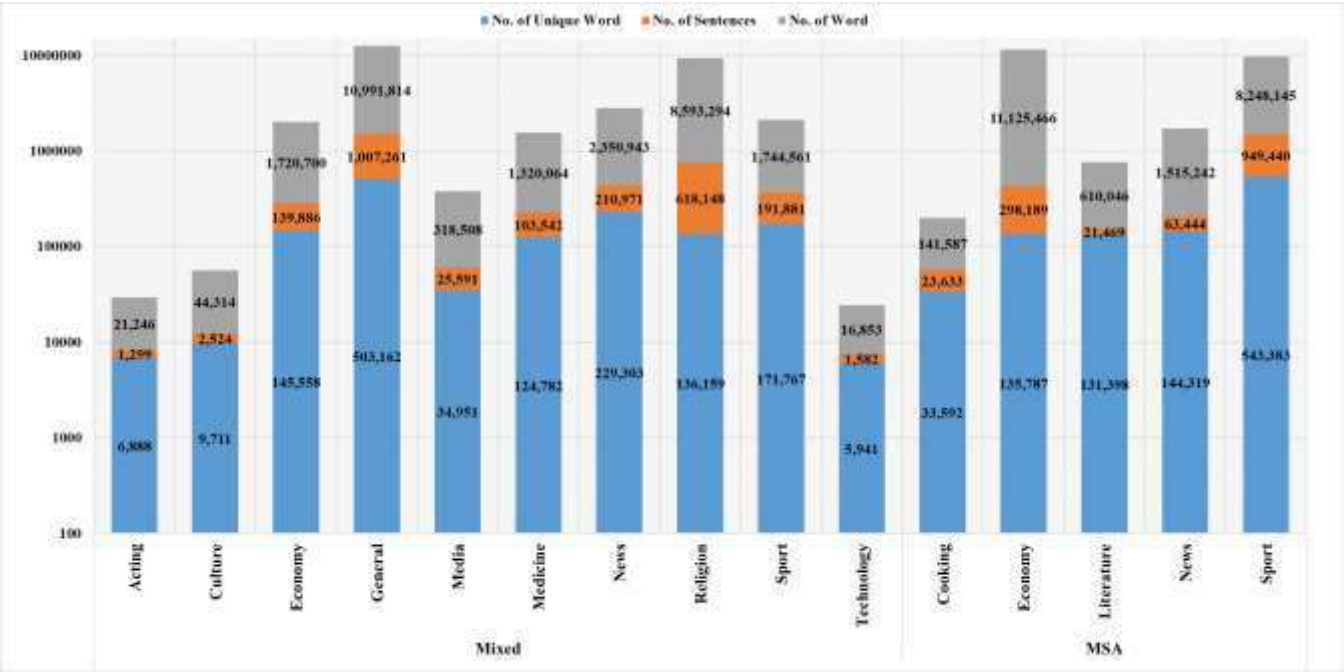


Figure 9. Collection statistics for Facebook data

2.3 KSUSC System Architecture

To overcome the current gaps discussed in the literature, this section proposes a new system that will be used to preprocess and build the King Saud University Saudi Corpus (KSUSC). The system proposed is divided into five main processes, as illustrated in Figure 10. First, data are processed to build lexicons for all accented characters and unwanted symbols. Second, the data are normalized to unify identical characters that are found with different representations. Third, irrelevant information and repeated sentences and words are removed. Fourth, numbers and English words are masked. Finally, in the fifth step, the corpus is encoded and validated to ensure that all characters are recognized and normalized. This is of particular importance because semantic corpora such as the KSUSC can be used with non-Arabic models such as BERT, which need to be strictly normalized and encoded.

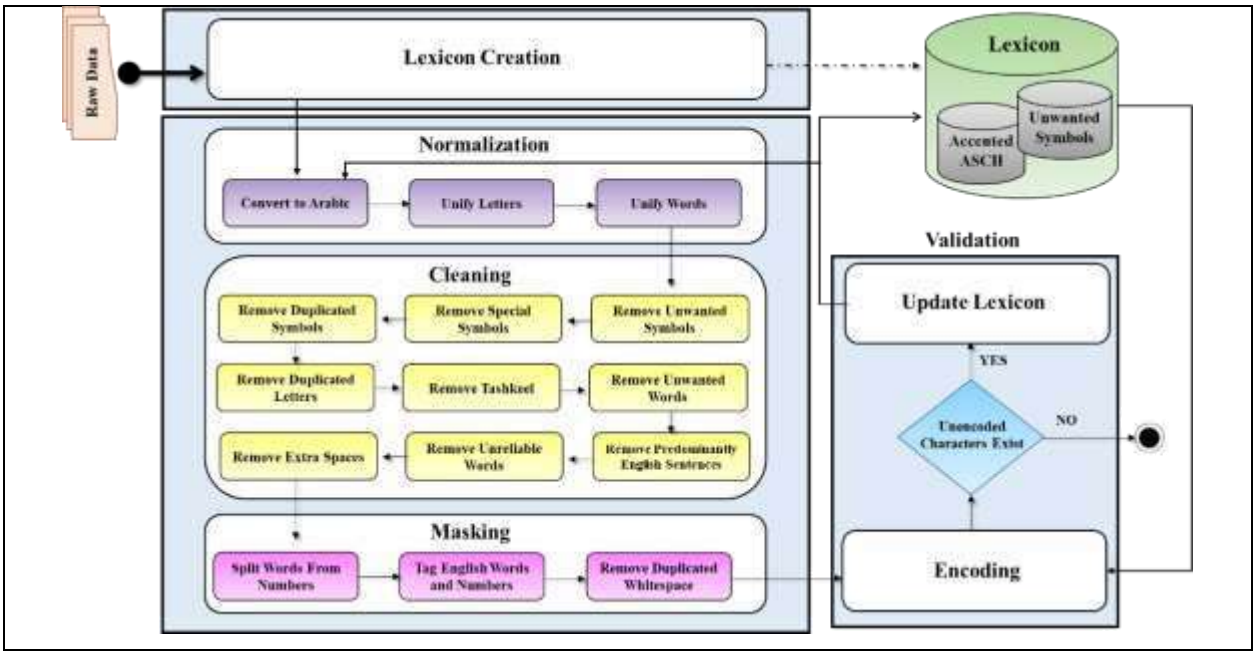


Figure 10. KSUSC system architecture

Due to the diversity in resources collection, using one round of normalization and cleaning would not work on all kinds of data; even the pre-existing corpora needs a certain level of pre-processing to be compatible with the proposed KSUSC corpus. Therefore, an incremental process was adopted into the proposed system to continuously update the lexicons. Thus, KSUSC system can be scalable to new SD documents or even other Arabic languages. The following sections will explain the details of the proposed system along with illustrated examples.

2.3.1 Lexicon Creation

Before starting to normalize and clean the text, it is important to build a lexicon for all variations of letters and symbols and for special unwanted symbols. This is of particular importance because the data are collected from various places, and thus, it is possible that the letters were written in different formats or using different keyboard settings. Moreover, irrelevant symbols and tags could have been included during the collection process. Thus, the collected text is scanned, and all unrecognized symbols and letters are extracted incrementally in the validation phase. Then, the list is investigated to create two lexicons: Accented ASCII and Unwanted Symbols. The 'Accented ASCII' lexicon includes all accented characters and their equivalent ASCII codes; an example is shown in Table 5. T he 'Unwanted Symbols' lexicon includes all characters that are not recognized. After creating these lexicons, it is possible to use them with any subsequent new data.

Table 5. Sample of ASCII lexicons for 'ا' and 'ي'-accented characters

Letter	ASCII	Variants	Variants ASCII Code	Letter	ASCII	Variants	Variants ASCII Code
ا	1575	آ	1649	ي	1610	ﻳـ	65268
		ﺎ	65166			ﻴـ	65267
		ﺍ	65165			ﻲ	65266
		ﺂ	64337			ﻱ	65265
		ٱ	1493			ى	1744
		أ	64336			ﻳـ	64510
		ٲ	1503			ی	1745
						ﻴـ	64511
						ﯼ	64484
						ﻴ	64486

2.3.2 Normalization

While collecting the data, it was noticed that a character's Unicode code can be different from one source to another. For example, some people write numbers in Hindi ASCII, while others write them in Arabic ASCII; sometimes spaces or punctuation are written with Arabic keyboards, while other times, they are written with English keyboards. All these characters have the same meaning but might confuse the semantic model if they are represented differently. Thus, it is important to first normalize the collected data regardless of the source from which it was collected. To do that, three processes are implemented as follows.

A. **Convert to Arabic:** All characters, space, and punctuations written in English are converted to Arabic, and the numbers written in Hindi ASCII are converted to English ASCII.

B. **Unify Letters:** In Arabic text, letters can be written differently depending on their position in the word or the region of the writer. For example, the letter 'ك' can be written as ک, ڪ, ك, ک, etc. However, it is impossible for Python regular expressions to recognize the similarity since each variation has a different ASCII code. Thus, accented characters are converted to ASCII characters using the previously built Accented ASCII lexicon.

C. **Unify Words:** In Arabic text, some words can be abbreviated using one character and, thus, should be unified as either abbreviated or full words. For example, Table 6 shows some abbreviation and their corresponding words.

Table 6. Abbreviation vs. real words

Abbreviation	Words
ﷺ	صلى الله عليه وسلم
ﷻ	جل جلاله
هـ	هجري

After applying the three processes of normalization, all letters should be recognizable and have the same meaning, making them ready to enter the cleaning phase. Figure 11 shows a sample of the text before the normalization phase applied and after it. Note the changes colored in blue.

مقالة جميلة توضح مقارنة رائعة بين Ionic و React_Native وأيهم أفضل، معلومة React Native عمل به Instagram Airbnb. 26//1428هـ، المعدل بالمراسيم الملكية رقم (م/70) وتاريخ 1437/11/6هـ، ورقم (م/73) وتاريخ 1439/7/18هـ، ورقم (م/115) وتاريخ 1439/12/5هـ، وبعد الصادرة بالقرار الوزاري رقم (7019) وتاريخ 1429/7/3هـ، وبناءً على ما تقتضيه المصلحة العامة.. يُقرر ما يلي:	**(a)**
مقالة جميلة توضح مقارنة رائعة بين Ionic وReact_Nativeوأيهم أفضل، معلومة React Native عمل به Airbnb،Instagram 26//1428هـ، المعدل بالمراسيم الملكية رقم (م/70) وتاريخ 1437/11/6 هجري، ورقم (م/73) وتاريخ 1439/7/18 هجري، ورقم (م/115) وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم (7019) وتاريخ 1429/7/3 هجري، وبناءً على ما تقتضيه المصلحة العامة.. يُقرر ما يلي:	**(b)**

Figure 11. Sample of data in normalization phase; (a) before normalization, (b) after normalization

2.3.3 Cleaning

The text cleaning phase is used to eliminate unwanted words and characters and improve the performance of a semantic model. This step removes all HTML tags, extra whitespaces, special characters, punctuation, and duplicate letters as well as any sentences that are not related to the meaning of the text or have been corrupted during collection. In particular, the following nine processes are implemented to clean the data and remove any noise.

A. Remove Unwanted Symbols: Unwanted symbols such as (;$*()"#!@^&-+={}[],|'<>- ﴾﴿ »«؛؟؟ É ﷺ□👍~) were replaced by whitespace in this task; while full stop and comma were left in the text because they give indication about the length of the sentences. Note that some symbols where not recognized at the beginning and, thus, the Unwanted Symbols lexicon is updated accordingly. For example, Table 7 shows some symbols that were not recognize at the beginning and how it were removed from the text. Moreover, Figure 12 shows a sample of the text before removing unwanted symbols and after removing them, colored in red.

Table 7. Removing of unknown symbols

Symbol	In the word	Text Before Cleaning	Text after cleaning
ɵ	ɵ ˬ ɵ	ضحكت ضحك هه يآبطيɵˬ ɵ بطناهة ؟	ضحكت ضحك هه يآبطي بطناهة
ˬ	ˬɵ	ضحكت ضحك هه يآبطيɵˬ ɵ بطناهة ؟	
؟	؟	ضحكت ضحك هه يآبطيɵˬ ɵ بطناهة ؟	
°	° °	انا حاسه انك حاط 18 عشان تجذب المتابع ° ° ؽ	انا حاسه انك حاط 18 عشان تجذب المتابع
ؽ	ؽ	انا حاسه انك حاط 18 عشان تجذب المتابع ° ° ؽ	

وتفيد مجلة «Heliyo» ، بأن علماء جامعة البلطيق الفيدرالية ، بالتعاون مع علماء جامعة مقاطعة كيميروفو ، درسوا تأثير الأشعة فوق. بالقرار الوزاري رقم(7019) "أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا ، أأدخل…	**(a)**
وتفيد مجلةHeliyo ، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق. بالقرار الوزاري رقم 7019 أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل…	**(b)**

Figure 12. Sample of data when removing unwanted symbols; (a) before removal, (b) after removal

B. Remove Special Symbols: Symbols such as (:\/%-$) needed to be removed from the text, however they had a meaning when they come with a number. For example, a colon can come

in numbers when it represents a date (12:30 pm) and backslash or dash can come in numbers when it represents dates (1/1/2020) or (1-1-2020). Thus, it was important to reserve these special symbols when they occur in numbers while removing them otherwise. Figure 13 shows a sample of the text before removing unwanted symbols and after removing them. Note how is the symbols in numbers (highlighted green) are kept, while the other symbols (highlighted red) were removed.

26//1428هجري ، المعدل بالمراسيم الملكية رقم (م/70) وتاريخ 6/11/1437هجري ، ورقم (م/73) وتاريخ 1439/7/18هجري ، ورقم (م/115) وتاريخ 5/12/1439هجري ، وبعد الصادرة بالقرار الوزاري رقم (7019) وتاريخ 3/7/1429هجري ، وبناءً على ما تقتضيه المصلحة العامة.. يُقرر ما يلي: خطورة الغيبة -محمد مختار الشنقيطير	**(a)**
26//1428هجري، المعدل بالمراسيم الملكية رقم (م/70) وتاريخ 6/11/1437هجري، ورقم (م/73) وتاريخ 1439/7/18هجري، ورقم (م/115) وتاريخ 5/12/1439هجري، وبعد الصادرة بالقرار الوزاري رقم (7019) وتاريخ 3/7/1429هجري، وبناءً على ما تقتضيه المصلحة العامة.. يُقرر ما يلي خطورة الغيبة محمد مختار الشنقيطير	**(b)**

Figure 13. Sample of data when removing special symbols; (a) before removal, (b) after removal

C. **Remove Duplicated Symbols:** After removing all unrelated symbols, it was noticed that full stop and commas are sometimes unnecessary repeated. Thus, it was decided to remove the duplication, as shown in Figure 14 (colored in red).

أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا ، أأدخل...	**(a)**
أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل.	**(b)**

Figure 14. Sample of data when removing duplicate symbols; (a) before removal, (b) after removal

D. **Remove Duplicated Letters:** Speakers sometimes repeat a letter in a word to emphasize their emotion. For example, in English, people might write ‘Good’ as ‘Goooooood’ to emphasize that they like something. Similarly, this type of redundancy occurs in Arabic and should be removed to standardize the data. For instance, the redundant letters ‘و’ in the word ‘كفووووو’ will be removed to obtain the correct word ‘كفو’. One exception is the word ‘هههه’, which is an

expression of laughing out loud. In this expression, the letter ‘ﻫـ’ is repeated several times depending on the user; however, if it is removed, the system might confuse it with the letter ‘ﻫـ’, which means Hijri, and mistakenly transform it to the word ‘هجري’. Thus, the expression for laughing out loud was transformed into two letters instead of one. For example, Figure 13 shows how the word ‘ههههههههه’ is transformed to a word with two letters, ‘هه’, which also means laughing.

ضحكت ضحك هههههههههه يآبطني θ ͺ θبطناهة⍰	(a)
ضحكت ضحك هه يآبطني θ ͺ θبطناهة⍰	(b)

Figure 15. Sample of data when removing duplicate letters; (a) before removal, (b) after removal

E. Remove Tashkeel (Arabic Punctuation Marks): In Arabic, there are special punctuation marks called “Tashkeel” that are used to emphasize certain vowels sound; such as “◌̃, ◌ً, ◌ٌ, ◌ٍ, ◌َ, ◌ُ, ◌ِ, ◌ّ, ◌ْ, ◌ۖ, ◌ٓ, ◌ـ, ◌ۗ, ◌ٔ, ◌ٕ, ,, ◌۠ ,◌ۢ, ◌ۚ, ◌ٰ, ◌ٕ, ◌ۘ, ◌ٗ, ,, ◌ٗ, ◌ٖ, ◌ٝ, ◌ۜ, ◌ۜ, ◌۫, ◌ٞ, ◌ٛ, ^, ◌ۦ, ◌۟, ◌ٚ, ◌ٛ, ◌ٙ, T, ◌ٜ, ◌ٚ, ◌ٍ, ◌ٌ, ◌ٜ, ◌ٜ, ◌ۣ, o, ◌ۨ ◌ٓ, ◌ۨ, ◌ۨ, ◌ٕ, ◌ٕ, ◌ٔ”. These marks were removed to standardize the text, as shown in Figure 16.

بارِكَ اللّهُ فيكِ يادكتوره خطورةُ الغَيْبَةِ كلُ الشرقيةِ تنتظرِكَ بِشوقٍ سددَ الباري خطاكِ	(a)
بارك الله فيك يادكتوره خطورة الغيبة كل الشرقية تنتظرك بشوق سدد الباري خطاك	(b)

Figure 16. Sample of data when removing Arabic punctuation marks; (a) before removal, (b) after removal

F. Remove Unwanted Words: I n the text, there are words and sentences with special uses that show emotions or are used as metadata, which has no relevance to the semantic meaning of the text itself and includes HTML tags, URLs, emails, hashtags, emojis, and other special metadata. These words and sentences are removed in this step, as shown in Figure 17. There is one exception during the removal process, which is related to hashtags. Since hashtags have meaning when they appear inside a sentence, it was decided that hashtags should be removed only when they appear separately as one word in a line. However, when a hashtag appears in a

sentence, it is processed so that the hash and the underscore are replaced with whitespace and the word itself is preserved.

قم بزيارة الموقع https://website.com/main وارسل بياناتك إلى الإيميل mymai@site.com RT حيو حارتي 2020 أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل … بارك الله فيك يادكتوره خطورة الغيبة -محمد مختار الشنقيطي تصاميم دعوية، كل الشرقية تنتظرك بشوق، سدد الباري خطاك وياليت قومي يعلمون	**(a)**
قم بزيارة الموقع وارسل بياناتك إلى الإيميل حيو حارتي 2020 أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا ، أأدخل … بارك الله فيك يادكتوره خطورة الغيبة -محمد مختار الشنقيطي تصاميم دعوية ، كل الشرقية تنتظرك بشوق ، سدد الباري خطاك وياليت قومي يعلمون	**(b)**

Figure 17. Sample of data when removing unwanted words; (a) before removal, (b) after removal

G. Remove Dominant English Sentences: During the cleaning process, sentences that include more that 25% of its words English were removed for the text. Figure 18 shows and example of two sentences that have English words, and which one will be removed (as colored in red).

وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية ، بالتعاون مع علماء جامعة مقاطعة كيميروفو ، درسوا تأثير الأشعة فوق اكتبي باليوتويب how to make packground to my iphone اتوقع كذا 1428//26هجري ، المعدل بالمراسيم الملكية رقم (م/70) وتاريخ 1437/11/6هجري ، ورقم (م/73) وتاريخ 1439/7/18هجري ، ورقم (م/115) وتاريخ 1439/12/5هجري ، وبعد الصادرة بالقرار الوزاري رقم (7019) وتاريخ 1429/7/3هجري ، وبناءً على ما تقتضيه المصلحة العامة.. يُقرر ما يلي:	**(a)**
وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي	**(b)**

Figure 18. Sample of data when removing sentences dominated by English words; (a) before removal, (b) after removal

H. Removing Unreliable Words: During the collection process, it was noticed that whitespaces got removed at some part of the sentences and more than word glued together. To reduce the manual revision of these words, it was decided to remove words that is longer than 15 letter, as illustrated in Figure 19. This is because, in Arabic, no word can be longer than that.

(a)	وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعةكيميروفودرسواتاثيرالأشعةفوق اعمل لايكات 256655895566655222544325665665566566 مرة
(b)	وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة اعمل لايكات مرة

Figure 19. Sample of data when removing unreliable words; (a) before removal, (b) after removal

I. Remove Extra Spaces: Extra whitespaces occurred in between punctuations and text or numbers. Thus, these spaces were removed, as shown in Figure 20.

(a)	وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري ، ورقم م /73 وتاريخ 1439/7/18 هجري ، ورقم م /115 وتاريخ 1439/12/5 هجري ، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 3 / 1429/7 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي
(b)	وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي

Figure 20. Sample of data when removing duplicate spaces; (a) before removal, (b) after removal

2.3.4 Masking

After cleaning the text, it is important to mask English words and numbers for future use with semantic models. These words can also be manually translated into Arabic in the future. To automate this process, tags are used for masking, and the following processes are applied.

A. Split Words From Numbers: Before masking the numbers, it was important to separate them from text since it was glued to other words at some documents, as shown in Figure 21.

(a)	وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي
(b)	وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي

Figure 21. Sample of data when splitting words from numbers; (a) before splitting, (b) after splitting

B. Tag English Words and Numbers: All English words will be masked with the tag [unkown] while numbers will be tagged as [number], as shown in Figure 22.

(a)	حيو حارتي 2020 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي شوف نسبة البطارية بين الساعة 10:03 و الساعة 10:54، الف مبروك الله يوفقكم ويسعدكم وبما انك في سان فرانسيكو جرب المطاعم خاصة كفطور. mamas nopa zazie كل التوفيق والنجاح
(b)	حيو حارتي number]] انا حاسه انك حاط] [numberعشان تجذب المتابع [[numberهجري، المعدل بالمراسيم الملكية رقم م] [numberوتاريخ [[numberهجري، ورقم م] [numberوتاريخ] [numberهجري، وبعد الصادرة بالقرار الوزاري رقم] [numberوتاريخ [[numberهجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي شوف نسبة البطارية بين الساعة] [numberو الساعة [[number]، الف مبروك الله يوفقكم ويسعدكم وبما انك في سان فرانسيكو جرب المطاعم خاصة كفطور. [unknown] [unknown] [unknown] كل التوفيق والنجاح

Figure 22. Sample of data when tagging English words and numbers; (a) before tagging, (b) after tagging

C. Remove Duplicated Whitespace: After applying all previous steps, it was noticed that whitespaces were duplicated due to replacing unwanted symbols or letters with spaces. Thus, these duplicates were removed as shown in Figure 23.

وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي	**(a)**
وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي	**(b)**

Figure 23. Sample of data when removing duplicate whitespace; (a) before removal, (b) after removal

2.3.5 Data Validation

After cleaning the data, it is important to validate the data and correct any mistakes found. For example, it was noticed that some symbols were not recognized and letters had different ASCII although they mean the same. Thus, it was important to validate the text and correct any error. In particular, this phase introduces two main processes to validate the data and correct unrecognized letters or symbols.

A. Encoding: All Arabic letters, in addition to full stops and commas, were matched to a list of English letters. Then, the collected data were encoded to English, and any letter or symbol that was not recognized and was not found in the lexicon was added to a list called a non-encoded list.

B. Updating Lexicons: After encoding the documents, the non-encoded list was investigated. If the list was not empty, Arabic letters were added to the matching entry in the Accented ASCII lexicon, and other symbols were added to the Unwanted Symbols lexicon.

The normalization and cleaning phases were repeated after updating the lexicons, and all files were encoded incrementally until the non-encoded list was empty. Figure 24 shows the encoding results after validation. All letters were translated to a matching English letter, and thus, all words were recognized.

حيو حارتي [number] ضحكت ضحك هه يآبطني بطناهة انا حاسه انك حاط [number] عشان تجذب المتابع الفديوهات لاتثبت انه المدرس ولا واضح الشكل ويمكن احد مثلها اساسا مو اثبات صريح وتفيد مجلة [unknown]، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق [number] هجري، المعدل بالمراسيم الملكية رقم [number] وتاريخ [number] هجري، ورقم [number] وتاريخ [number] هجري، ورقم [number] وتاريخ [number] هجري، وبعد الصادرة بالقرار الوزاري رقم [number] وتاريخ [number] هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي شوف نسبة البطارية بين الساعة [number] والساعة [number]، الف مبروك الله يوفقكم ويسعدكم وبما انك في سان فرانسيكو جرب المطاعم خاصة كفطور. [unknown] [unknown] [unknown] كل التوفيق والنجاح الرحمة المهداة صلى الله عليه وسلم أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل بارك الله فيك يادكتوره خطورة الغيبة محمد مختار الشنقيطي، تصاميم دعوية، كل الشرقية تنتظرك بشوق، سدد الباري خطاك ويااليت قومي يعلمون	**(a)**
Mjh MGQJj number VMcJ VMc gg HWfGgI GfG MGSg Gfc MGW number YTGf JLPH GdeJGHY GdaOjhgGJ dGJKHJ Gfg GdeOQS hdG hGVM GdTcd hjecf GMO eKdgG GSGSG eh GKHGJ UQjM hJajO eLdI unknown , HCf YdeGA LGeYI GdHdWjb GdajOQGdjI, HGdJYGhf eY YdeGA LGeYI ebGWYI cjejQhah, OQShG JCKjQ GdCTYI ahb number gLQj , GdeYOd HGdeQGSje GdedcjI Qbe e number hJGQjN number gLQj , hQbe e number hJGQjN number gLQj , hQbe e number hJGQjN number gLQj , hHYO GdUGOQI HGdbQGQ GdhRGQj Qbe number hJGQjN number gLQj , hHfGA Ydi eG JbJVjg GdeUdMI GdYGeI. jbQQ eG jdj Tha fSHI GdHWGQjI Hjf GdSGYI number h GdSGYI number , Gda eHQhc Gddg jhabce hjSYOce hHeG Gfc aj SGf aQGfSjch LQH GdeWGYe NGUI caWhQ. unknown unknown unknown cd GdJhajb hGdfLGM GdQMeI GdegOGI Udi Gddg Ydjg hSde CQCjJ EPG UdjJ GdUdhGJ GdecJhHGJ hUeJ QeVGf hCMddJ GdMdGd hMQeJ GdMQGe hde CRO Ydi Pdc TjFG, CCONd HGQc Gddg ajc jGOcJhQg NWhQI GdZjHI eMeO eNJGQ GdTfbjWj, JUGeje OYhjI, cd GdTQbjI JfJXQc HThb, SOO GdHGQj NWGc hjGdjJ bhej jYdehf	**(b)**

Figure 24. Sample of data during the validation phase; (a) before encoding, (b) after encoding

2.3.6 Use Case Example

In order to understand the processes proposed, this section present a full use case the shows the result of each step in the method. Starting from the original sample, presented in Figure 25, the text will be normalized as shown in Figure 26, cleaned as shown in Figure 27, masked as shown in Figure 28, and lastly validated based on the encoding shown in Figure 29.

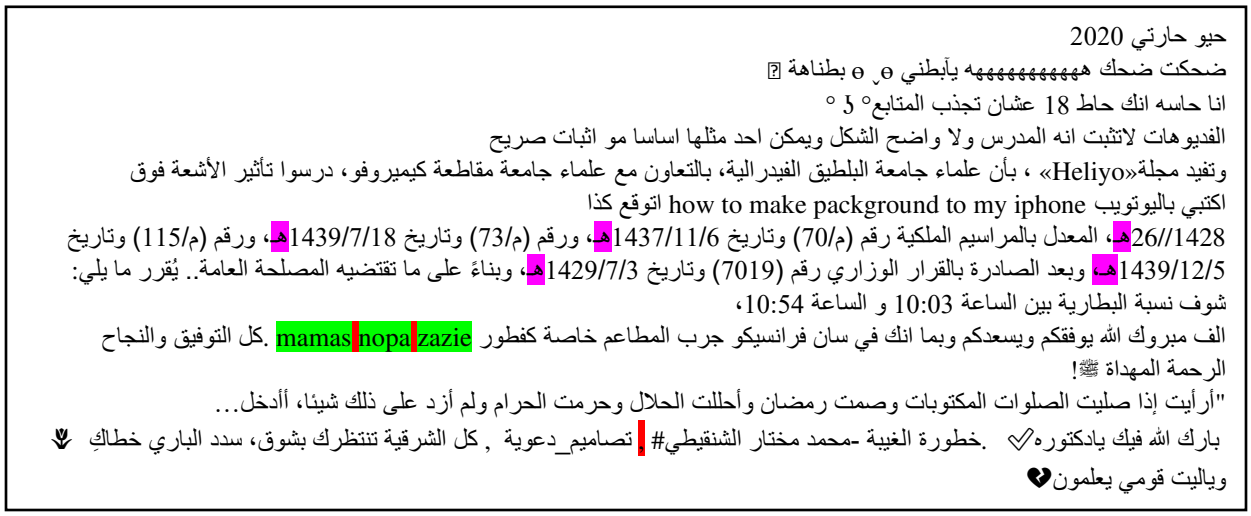

حيو حارتي 2020
ضحكت ضحك هههههههههههه يآبطني ɵ ِ ɵ بطناهة ⍰
انا حاسه انك حاط 18 عشان تجذب المتابع° Ɜ °
الفديوهات لاتثبت انه المدرس ولا واضح الشكل ويمكن احد مثلها اساسا مو اثبات صريح
وتفيد مجلة«Heliyo» ، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق
اكتبي باليوتويب how to make packground to my iphone اتوقع كذا
26//1428هـ، المعدل بالمراسيم الملكية رقم (م/70) وتاريخ 1437/11/6هـ، ورقم (م/73) وتاريخ 1439/7/18هـ، ورقم (م/115) وتاريخ
1439/12/5هـ، وبعد الصادرة بالقرار الوزاري رقم (7019) وتاريخ 1429/7/3هـ، وبناءً على ما تقتضيه المصلحة العامة.. يُقرر ما يلي:
شوف نسبة البطارية بين الساعة 10:03 و الساعة 10:54،
الف مبروك الله يوفقكم ويسعدكم وبما انك في سان فرانسيكو جرب المطاعم خاصة كفطور mamas nopa zazie .كل التوفيق والنجاح
الرحمة المهداة ﷺ!
"أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل...
بارك الله فيك يادكتوره✓ .خطورة الغيبة -محمد مختار الشنقيطي# . تصاميم_دعوية , كل الشرقية تنتظرك بشوق، سدد الباري خطاكِ
وياليت قومي يعلمون💔

Figure 25. Original sample of the use case

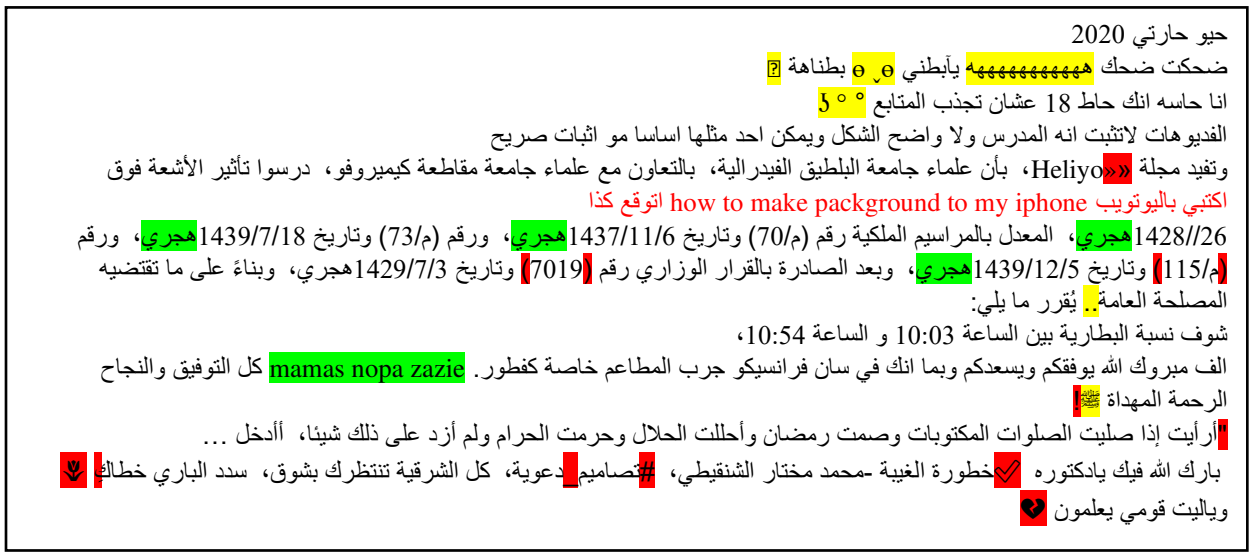

حيو حارتي 2020
ضحكت ضحك هههههههههههه يآبطني ɵ ِ ɵ بطناهة ⍰
انا حاسه انك حاط 18 عشان تجذب المتابع ° Ɜ °
الفديوهات لاتثبت انه المدرس ولا واضح الشكل ويمكن احد مثلها اساسا مو اثبات صريح
وتفيد مجلة «Heliyo»، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق
اكتبي باليوتويب how to make packground to my iphone اتوقع كذا
1428//26هجري، المعدل بالمراسيم الملكية رقم (م/70) وتاريخ 1437/11/6هجري، ورقم (م/73) وتاريخ 1439/7/18هجري، ورقم
(م/115) وتاريخ 1439/12/5هجري، وبعد الصادرة بالقرار الوزاري رقم (7019) وتاريخ 1429/7/3هجري، وبناءً على ما تقتضيه
المصلحة العامة.. يُقرر ما يلي:
شوف نسبة البطارية بين الساعة 10:03 و الساعة 10:54،
الف مبروك الله يوفقكم ويسعدكم وبما انك في سان فرانسيكو جرب المطاعم خاصة كفطور. mamas nopa zazie كل التوفيق والنجاح
الرحمة المهداة ﷺ!
"أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل ...
بارك الله فيك يادكتوره ✓خطورة الغيبة -محمد مختار الشنقيطي، #تصاميم_دعوية، كل الشرقية تنتظرك بشوق، سدد الباري خطاك
وياليت قومي يعلمون 💔

Figure 26. Use case sample after normalization

حيو حارتي 2020
ضحكت ضحك هه يآبطني بطناهة
انا حاسه انك حاط 18 عشان تجذب المتابع
الفديوهات لاتثبت انه المدرس ولا واضح الشكل ويمكن احد مثلها اساسا مو اثبات صريح
وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي
شوف نسبة البطارية بين الساعة 10:03 و الساعة 10:54،
الف مبروك الله يوفقكم ويسعدكم وبما انك في سان فرانسيكو جرب المطاعم خاصة كفطور. mamas nopa zazie كل التوفيق والنجاح
الرحمة المهداة صلى الله عليه وسلم
أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل
بارك الله فيك يادكتوره خطورة الغيبة محمد مختار الشنقيطي، تصاميم دعوية، كل الشرقية تنتظرك بشوق، سدد الباري خطاك
ويااليت قومي يعلمون

Figure 27. Use case sample after cleaning phase

حيو حارتي [number]
ضحكت ضحك هه يآبطني بطناهة
انا حاسه انك حاط [number] عشان تجذب المتابع
الفديوهات لاتثبت انه المدرس ولا واضح الشكل ويمكن احد مثلها اساسا مو اثبات صريح
وتفيد مجلة [unknown]، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق
[number] هجري، المعدل بالمراسيم الملكية رقم [number] وتاريخ [number] هجري، ورقم [number] وتاريخ [number] هجري، ورقم [number] وتاريخ [number] هجري، وبعد الصادرة بالقرار الوزاري رقم [number] وتاريخ [number] هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي
شوف نسبة البطارية بين الساعة [number] والساعة [number]،
الف مبروك الله يوفقكم ويسعدكم وبما انك في سان فرانسيكو جرب المطاعم خاصة كفطور. [unknown] [unknown] [unknown] كل التوفيق والنجاح
الرحمة المهداة صلى الله عليه وسلم
أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل
بارك الله فيك يادكتوره خطورة الغيبة محمد مختار الشنقيطي، تصاميم دعوية، كل الشرقية تنتظرك بشوق، سدد الباري خطاك
ويااليت قومي يعلمون

Figure 28. Use case sample after masking

Mjh MGQJj number
VMcJ VMc gg HWfGgI
GfG MGSg Gfc MGW number YTGf JLPH GdeJGHY
GdaOjhgGJ dGJKHJ Gfg GdeOQS hdG hGVM GdTcd hjecf GMO eKdgG GSGSG eh GKHGJ UQjM
hJajO eLdI unknown , HCf YdeGA LGeYI GdHdWjb GdajOQGdjI, HGdJYGhf eY YdeGA LGeYI
ebGWYI cjejQhah, OQShG JCKjQ GdCTYI ahb
number gLQj , GdeYOd HGdeQGSje GdedcjI Qbe e number hJGQjN number gLQj , hQbe e number
hJGQjN number gLQj , hQbe e number hJGQjN number gLQj , hHYO GdUGOQI HGdbQGQ
GdhRGQj Qbe number hJGQjN number gLQj , hHfGA Ydi eG JbJVjg GdeUdMI GdYGeI. jbQQ eG
jdj
Tha fSHI GdHWGQjI Hjf GdSGYI number h GdSGYI number ,
Gda eHQhc Gddg jhabce hjSYOce hHeG Gfc aj SGf aQGfSjch LQH GdeWGYe NGUI caWhQ.
unknown unknown unknown cd GdJhajb hGdfLGM
GdQMeI GdegOGI Udi Gddg Ydjg hSde
CQCjJ EPG UdjJ GdUdhGJ GdecJhHGJ hUeJ QeVGf hCMddJ GdMdGd hMQeJ GdMQGe hde CRO
Ydi Pdc TjFG, CCONd
HGQc Gddg ajc jGOcJhQg
NWhQI GdZjHI eMeO eNJGQ GdTfbjWj,
JUGeje OYhjI,
cd GdTQbjI JfJXQc HThb, SOO GdHGQj NWGc
hjGdjJ bhej jYdehf

Figure 29. Use case sample after encoding

2.4 KSUSC Corpus Design

After collecting the data from various resources and cleaning them using the KSUSC system, it was decided to design the KSUSC corpus to be the largest Saudi corpus to date. To design this new corpus, the following criteria were validated:

- Corpus Size: The designed corpus is a large corpus with more than 1B words.
- Corpus Languages: The languages of the final corpus are MSA and SD.
- Material Mode: The material of the KSUSC is written text because it can be easily collected and validated. However, in the future, spoken materials will also be considered for inclusion.
- Corpus Dates: The KSUSC corpus covers past materials taken from preexisting corpuses (up to 2010) in addition to recent new content written by the end of 2020.
- Corpus Source: The text was collected from five resources, including preexisting corpora, websites, and different social media platforms.
- Corpus Domains: The KSUSC is a diverse corpus that covers more than 26 domains.

After validating the proposed categories, it was possible to design and build KSUSC corpus to include a total of 161,795,667 sentences, 1,183,156,600 words, and 14,240,747 unique words. For more details about the detailed statistics of the corpus, see "Appendix A – KSUSC Detailed Statistics". Metadata was introduced to archive the text and the final outcome was copyrighted as will be explained next.

2.4.1 Text Distribution

To uncover the design criteria of KSUSC corpus, the general statistics are outlined in Table 8. From this table, it is clear how KSUSC is a large size corpus that include both MSA and SD languages. The size of MSA text is +8M unique words, +146M sentences, and ~1B words, while the size of the SD with mixed is +6M unique words, +14M sentences, +150M number of words. Although the total number of words in SD text might be relatively small comparing to the MSA in KSUSC, but it is still considerably bigger than all available corpora in SD. Moreover, when comparing the unique number of words between MSA and SD, it can be seen that MSA takes 56% of the total number of unique words while SD and mixed acquire the rest. Thus, KSUSC is still rich with SD vocabularies and morphologies.

Table 8. KSUSC distribution across languages

Language	No. of Sentences	No. of Words	No. of Unique Words
MSA	146,969,746	1,032,814,633	8,029,018
SD	12,441,955	119,877,091	5,014,593
Mixed	2,383,966	30,464,876	1,197,136
Total	**161,795,667**	**1,183,156,600**	**14,240,747**

With respect to the date design criteria, as illustrated in Figure 30, it is clear that KSUSC data is saturated at the end (between 2018-2020). This of particular importance to ensure that new vocabularies, reflecting recent events, are covered by KSUSC. Note that this figure reflects the number of unique words to avoid noisy or redundant data. There are some missing years from the source data, but this would not affect the quality of the corpus since we are more concern with data from the last three years that can help with new tasks and challenge, such as COVID-19 or Saudi Vision 2030.

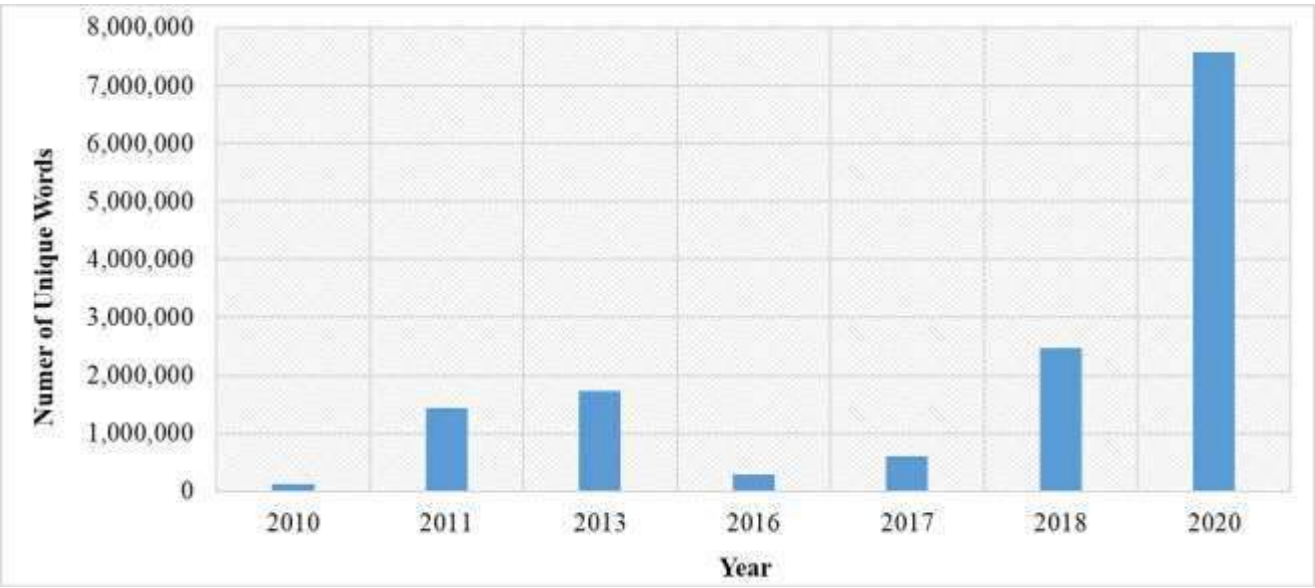


Figure 30. KSUSC designed distribution across timeline

When considering the source of the text, as illustrated in Table 9, more than 102M number of words in SD language resulted from the YouTube, +10M from Twitter, and +2M from web crawling. However, when considering the MSA text, it can be noticed that +900M resulted after pre-processing the pre-exiting datasets indicating that, indeed, they were further cleaned and

standardized. Moreover, +32M number of words in MSA resulted from web crawling and +20M words from Facebook.

Table 9. KSUSC designed distribution across sources

	Language	No. of Sentences	No. of Words	No. of Unique Words
Pre-Existing Dataset	MSA	145,193,442	980,222,925	5,642,982
	SD	502,562	4,142,961	335,718
		145,696,004	**984,365,886**	**5,978,700**
Facebook	Mixed	2,112,808	26,001,220	922,787
	MSA	1,127,112	20,409,001	622,403
		3,239,920	**46,410,221**	**1,545,190**
Twitter	Mixed	39,027	447,612	75,539
	SD	1,009,425	10,894,402	769,388
		1,048,452	**11,342,014**	**844,927**
Website	MSA	649,192	32,182,707	1,763,633
	SD	255,565	2,476,153	97,791
		904,757	**34,658,860**	**1,861,424**
YouTube	Mixed	232,131	4,016,044	198,810
	SD	10,674,403	102,363,575	3,811,696
		10,906,534	**106,379,619**	**4,010,506**

Finally, the categories distribution is another design criterion that must be highlighted with respect to the number of unique words. From Figure 31, it can be noticed that the category that include the highest number of unique words is the general (with 34%). This is because it included text that do not target a certain topic. After that, text about acting comes in second place, with 19% of the number of unique words, and news comes in third place with approximately 7% of the total number of unique words. Following that, the corpus is distributed similarly between the rest of the 22 category, with music (0.02%) and history (0.03%) as the categories with smallest number of unique words.

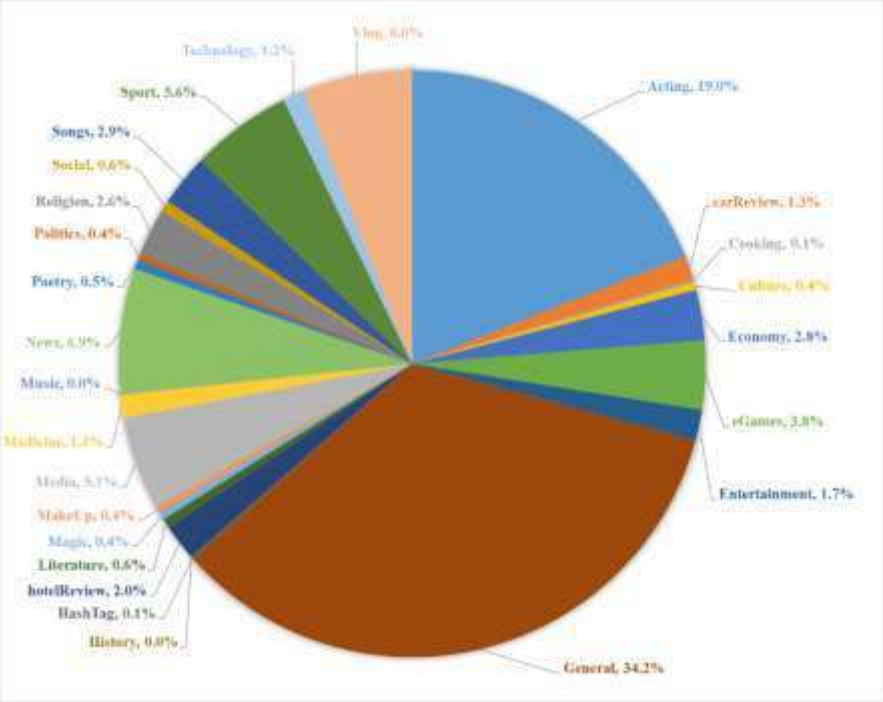


Figure 31. KSUSC designed distribution across categories

Although these statistics show the percentage of unique words in each domain, it must be clarified that these domains are based on the source text and not the vocabulary itself. In other words, if a word appears in sports-related text, this does not mean that the word does not appear in news-related text; rather, this means that it will not appear again in the same domain. Therefore, it was important to characterize the corpus with different metadata and archive it for further use.

2.4.2 Text Metadata

Any corpus in the literature has to be categorized and described with metadata. To facilitate further improvement in the future, it was decided to record as much information as possible about the corpus collected. In particular, KSUSC metadata included the source name, domain, file name, URL of the sources (or query/region), year of the text, and the total number of words, the number of sentences, and the number of unique words. These criteria will allow researchers to restrict their data to a specific source, period of time, domain, or even a query or region. These clear design criteria and metadata will be helpful in the corpus compilation process and when validating the accuracy of a certain task that is applied across the corpus.

2.4.3 Challenges and Difficulties

The process of collecting SD text revealed different challenges, and the construction of the KSUSC presented many difficulties. These challenges can be summarized as follows:

- It was difficult to distinguish SD text from other Gulf DA languages, and Saudi dialect experts had to be consulted.
- The social media text was very dirty and needed intensive incremental cleaning.
- No common SD lexicons were found; thus, they had to be created from scratch.
- Social media platforms had many restrictions when collecting the data, which extended the collection process to obtain enough text for the corpus.
- Some domains did not exist in current sources.
- Not all preexisting corpora were available or free to access.

Because of these challenges, the source of each text was included in the design criteria, but some domains and years were not fully covered.

2.4.4 Copyrights

The designed corpus was built from sources available online, and some have an active copyright, such as newspapers, magazines, books, tweets, and websites. Thus, the following actions have been taken: (1) bibliographic information about the corpus content is provided; (2) previews of the full text are restricted and are not available to the public; and (3) the collected text is not distributed and is locally used for research purposes. The corpus will be used according to the previously stated restrictions, and because it is intended for research purposes, it is consistent with the current Saudi copyright law [54].

3. Phase 3.2: Arab Word Vector Estimation

Nowadays, Internet based applications are involved in many aspects of modern life. Several important services are being provided to internet users such as E-Learning, Online gaming, E-government applications, social communications and news monitoring. For many of these applications, watching videos is a vital component for information delivery, especially in the cases of E-learning and news watching. With billions of internet users in the world, and tens of millions of them in the Middle East, there is always a need for better and more efficient videos retrieval techniques.

With the recent developments and intensive use of the Internet, word vector estimation became limited and new discoveries have been found with respect to sentence estimation. It was found that the most common criterion used to retrieve videos is key-sentences. Given a set of available videos, millions of them for practical applications, the user can query the system by entering a sentence to describe the content of the video to be viewed. There are some problems that can make this process inefficient. One of such problems is that the same video can be retrieved using several sentences. i.e., the user can use a different query other than the key sentences that are associated with the video to describe it. For any efficient retrieval process, the system should be able to recognize the correlation between similar, but different, queries.

3.1 Literature Review

In this section, BERT and BERT-Like models available in the literature are discussed and its use for sentence embeddings is also presented.

3.1.1 BERT-Like Models

Natural Language Processing (NLP) has gained a significant boost in the last few years. This is due to the advances happened in machine learning and in deep learning in particular. Recent achievements in NLP are using new technique called BERT [55] and some of similar models [56-58]. BERT and BERT-Like models [56-58] are generally self-supervised machine learning techniques that make use of the huge amounts of unlabeled text data available on the internet. The following is a summary of some of these techniques' methodologies.

BERT: BERT is a relatively new language representation model, which stands for Bidirectional Encoder Representations from Transformers. The BERT process corrupts the input by replacing some tokens with [MASK] and then train a model to reconstruct the original tokens. Unlike other relatively older language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications. BERT is conceptually simple and empirically powerful.

RoBERTa: RoBERTa is a replication study of BERT pre-training that carefully measures the impact of many key hyper parameters and training data size. The RoBERTa version of BERT is more computationally efficient and the model achieves state-of-the-art results at the time of its publication on several standard datasets. These results highlight the importance of previously overlooked design choices, and raise questions about the source of some of the reported improvements.

ELECTRA: Masked language modeling (MLM) pre-training methods such as BERT and RoBERTa corrupt the input by replacing some tokens with [MASK] and then train a model to reconstruct the original tokens. While they produce good results when transferred to downstream NLP tasks, they generally require high computation to be effective. As an alternative, the Authors of ELECTRA proposed a more sample-efficient pre-training process. Instead of masking the input, ELECTRA approach corrupts it by replacing some tokens with plausible alternatives sampled from a small generator network. Then, instead of training a model that predicts the original identities of the corrupted tokens, ELECTRA trains a discriminative model that predicts whether each token in the corrupted input was replaced by a generator sample or not. Thorough experiments demonstrate this new pre-training task is more efficient than MLM because the task is defined over all input tokens rather than just the small subset that was masked out.

Although the previously generated BERT-based models are multilingual, but the performance of these models are lower than the models that target a specific language. Many researchers with different lingual background proposed BERT-based models for several specific languages [59][60]. For Arabic language, the authors of [61] presented a BERT-based language model called ArabicBERT. Also, in [62] and [63], the Authors proposed two language models, AraBERT and AraELECTRA, for Arabic language that are based on BERT and ELECTRA respectively. The following are a summary of the methodologies used for these models.

ArabicBERT: ArabicBERT was the first pre-trained BERT model for Arabic [61]. Four Arabic BERT language models were trained from scratch and made publicly available for use. It had used a corpus that consists of the unshuffled version of OSCAR data [64] and a relatively recent data dump from Wikipedia, which sums up to 8.2B words and a vocabulary set of 32,000 Word pieces. The corpus and the vocabulary set were not restricted to MSA; they contained some DA too, which boosted models performance in terms of data from social media platforms.

AraBERT: In [62], the authors of AraBERT pre-trained BERT specifically for Arabic language in the pursuit of achieving the same success that BERT did for English language. The performance of AraBERT is compared to multilingual BERT from Google and other state-of-the-art approaches. The results showed that the newly developed AraBERT achieved state-of-the-art performance on the mostly tested Arabic NLP tasks. The pre-trained AraBERT models are publicly available on github.com/aub-mind/araBERT hoping to encourage research and applications for Arabic NLP.

AraELECTRA: The Authors of [63] developed an Arabic language representation model, which they named ARAELECTRA. Their model is pre-trained using the replaced token detection objective on large Arabic text corpora. They evaluated their model on two Arabic reading comprehension tasks, and showed that ARAELECTRA outperform current state-of-the-art Arabic language representation models that rely only on pre-training via masked language modeling given the same pre-training data and with even a smaller model size.

3.1.2 Sentence Embeddings

The most effective method to check the correlation between several sentences is based on what is called sentence embeddings. Sentence embeddings are vectors in hyperspace that describe the semantics of the sentence in the sense that the sentences that have similar semantic meaning will have close vectors in the hyperspace. To make use of the recent advances in NLP to increase the efficiency of video retrieval systems, we need to be able to generate robust sentence embeddings from users' queries and video key sentences for a better matching. BERT and BERT-like models have been built essentially for modeling natural languages. Some attempts have been conducted to use BERT-like model to extract embeddings for sentences. However, the resulting sentence embeddings were not efficient enough for practical usage [65].

In [65], the authors presented a method to utilize BERT models to generate robust sentence embeddings. While BERT and RoBERTa have set a new state-of-the-art performance on sentence-pair regression tasks like semantic textual similarity (STS), however, they require that both sentences fed into the network; which causes a massive computational overhead. For example, finding the most similar pair in a collection of 10,000 sentences requires about 50 million inference computations (around 65 hours) with BERT. The construction of BERT makes it unsuitable for semantic similarity search. In [65], the authors proposed Sentence-BERT (SBERT), a modification of the pre-trained BERT network that use siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. This can be done effectively by fine-tuning the original BERT model using a dataset like SNLI [66]. Figure 32 & Figure 33 illustrate SBERT architectures for training and inference respectively. This reduces the effort for finding the most similar pair from 65 hours with BERT / RoBERTa to about 5 seconds with SBERT, while maintaining the accuracy from BERT. The authors evaluated SBERT and SRoBERTa on common STS tasks and transfer learning tasks, where it outperforms other state-of-the-art sentence embeddings methods.

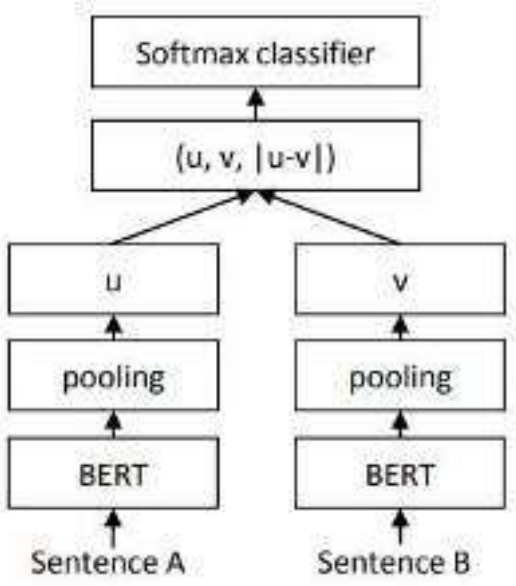


Figure 32. SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

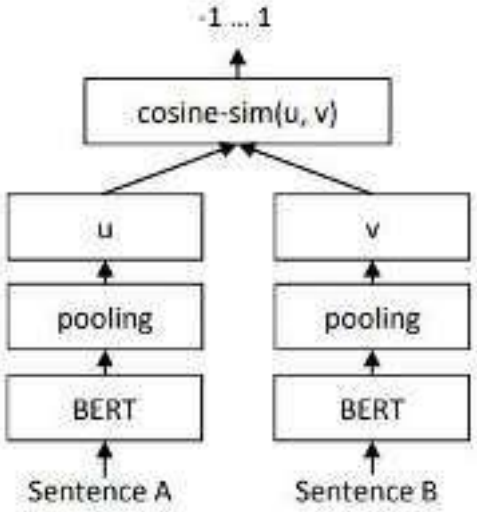


Figure 33. SBERT architecture at inference, for example, to compute similarity scores.

In [67], the authors presented an easy and efficient method to extend existing sentence embedding models to new languages. This allows creating multilingual versions from previously monolingual models. The training is based on the idea that a translated sentence should be mapped to the same location in the vector space as the original sentence. They used the monolingual model to generate sentence embeddings for the source language and then train a new system on translated

sentences to mimic the original model. They have tried several recipes and generated several model. We will refer to the best one as SBERT-paraphrase. Compared to other methods for training multilingual sentence embeddings, this approach has several advantages: it is easy to extend existing models with relatively few samples to new languages, it is easier to ensure desired properties for the vector space, and the hardware requirements for training are lower. They demonstrated the effectiveness of their approach for 50+ languages, including Arabic, from various language families.

In [68], the authors adapted multilingual BERT to produce language-agnostic sentence embeddings model, called LaBSE that supports 109 languages including Arabic. Their model combines masked language model and translation language model [69] pre-training with a translation ranking task using bi-directional dual encoders. Their sentence embeddings establish new state-of-the-art results on BUCC [70] bitext retrieval.

3.2 Empirical Study

We have applied SBERT on three datasets, ArabicBERT, AraBERT and AraELECTRA. We have also evaluated another two models: SBERT-paraphrase and LaBSE. To compare the results of all models, we have checked each model based on a semantic sentence similarity search. Using three different sentence queries, each model had to retrieve the most similar five candidate sentences out of around 10000 sentences. The following are examples of the results of each model.

3.2.1 ArabicBERT Results

Query: رجل يأكل المكرونة.
Top 5 most similar sentences in corpus:

(Score: 0.8249) رجل يأكل طعام.
(Score: 0.7945) رجل يأكل قطعة من الخبز.
(Score: 0.7492) بالبحرين.
(Score: 0.7377) البطالة بشكل فعلى.
(Score: 0.7305) بالتخلى عن الحكم.
==========

Query: شخص ما يلبس كغوريللا يعزف على الطبول.
Top 5 most similar sentences in corpus:

(Score: 0.7502) قرد يعزف على الطبول.

(Score: 0.7008) الرئيس السلوفاكى يصل الى شيآن.
(Score: 0.6667) لى روى هوان يصل الى ماكاو.
(Score: 0.6584) التقى الجانبان فى تشيكرز المقر الرسمى الريفى لبلير.
(Score: 0.6534) بليكس والبرادعي يصل الى بغداد موسع.

==========

Query: فهد يطارد فريسة في حقل.
Top 5 most similar sentences in corpus:

(Score: 0.8670) فهد يجري خلف فريسته.
(Score: 0.7497) الفيصل العرب لا يسعون لإقصاء الرئيس العراقى.
(Score: 0.7455) وترفض خطة إسقاط صدام.
(Score: 0.7453) أبوظبى اكد عبيد بن سيف الناصرى وزير النفط والثروة.
(Score: 0.7446) سفير عنيد يقول ان صدام لن يرحل عن العراق.

==========

3.2.2 AraBERT Results

Query: رجل يأكل المكرونة.
Top 5 most similar sentences in corpus:

(Score: 0.8357) تاريخية الى الامام.
(Score: 0.8287) درجة الحرارة هناك.
(Score: 0.8285) التايمز.
(Score: 0.8285) التايمز.
(Score: 0.8285) التايمز.

==========

Query: شخص ما يلبس كغوريللا يعزف على الطبول.
Top 5 most similar sentences in corpus:

(Score: 0.8719) قرد يعزف على الطبول.
(Score: 0.7936) من الوزراء والنواب فى مقاطعة اونتيريو.
(Score: 0.7926) الدولى الوقت والمساحة لحل الازمة.
(Score: 0.7873) مسؤول صحى الاردن خال من مرض الملاريا.
(Score: 0.7845) انطلاقة الكفاح المسلح.

==========

Query: فهد يطارد فريسة في حقل.
Top 5 most similar sentences in corpus:

(Score: 0.8764) فهد يجري خلف فريسته.
(Score: 0.7465) قرية صينية تشترى طائرات.
(Score: 0.7394) اكبر مركز ازهار فى آسيا.
(Score: 0.7390) رجل يأكل طعام.

(Score: 0.7370) انثى اسد كينية تتبنى عجلا اخر.
==========

3.2.3 AraELECTRA Results

Query: رجل يأكل المكرونة.
Top 5 most similar sentences in corpus:

(Score: 0.9500) رجل يأكل قطعة من الخبز.
(Score: 0.9468) نهاية الوظيفة الابدية فى الصين.
(Score: 0.9455) رجل يركب حصان أبيض على الأرضية.
(Score: 0.9413) رجل يركب حصان.
(Score: 0.9413) اجراء اول تفتيش لمسكن مواطن عراقي.
==========

Query: شخص ما يلبس كغوريللا يعزف على الطبول.
Top 5 most similar sentences in corpus:

(Score: 0.9464) زنزبار تنعش صناعة القرنفل.
(Score: 0.9441) اول فول صويا هجين فى العالم.
(Score: 0.9271) الصين قادرة على بناء محطات كهرنووية كبيرة.
(Score: 0.9229) طائفة دينية تزعم ميلاد ثانى طفل مستنسخ فى العالم.
(Score: 0.9186) ان تصريحات كويزومى سخيفة منطقيا.
==========

Query: فهد يطارد فريسة في حقل.
Top 5 most similar sentences in corpus:

(Score: 0.9680) زلزال متوسط يهز كولومبيا.
(Score: 0.9591) ضبط احياء ضارة فى مقاطعة صينية.
(Score: 0.9583) عزيز يتوقع رفض الدول المجاورة للحرب.
(Score: 0.9572) قرية صينية تشترى طائرات.
(Score: 0.9540) أربعة الاستثمارات الاجنبية فى مدينة ساحلية.
==========

3.2.4 SBERT-paraphrase Results

Query: رجل يأكل المكرونة.
Top 5 most similar sentences in corpus:

(Score: 0.8725) رجل يأكل طعام.
(Score: 0.6890) رجل يأكل قطعة من الخبز.
(Score: 0.4892) رجل يركب حصان.
(Score: 0.4572) رجل يركب حصان أبيض على الأرضية.
(Score: 0.3671) مجموعة كونغولية متمردة تنفى تناولها لحوم البشر.
==========

Query: شخص ما يلبس كغوريللا يعزف على الطبول.
Top 5 most similar sentences in corpus:

(Score: 0.6012) قرد يعزف على الطبول.
(Score: 0.5324) امرأة تعزف على الكمان.
(Score: 0.3234) وموسيقى الراب على الشباب.
(Score: 0.2967) رجل يأكل قطعة من الخبز.
(Score: 0.2838) لاعب خط وسط صينى ينضم للدورى الالمانى.
==========

Query: فهد يطارد فريسة في حقل.
Top 5 most similar sentences in corpus:

(Score: 0.7528) فهد يجري خلف فريسته.
(Score: 0.4827) الحيوان والنبات من الخارج.
(Score: 0.4233) فرار زعيم من المتمردين الى شمال اوغندا.
(Score: 0.4043) توطين الرعاة فى التبت.
(Score: 0.4001) انسداد نهر النيل فى غرب اوغندا.

3.2.5 LaBSE Results

Query: رجل يأكل المكرونة.
Top 5 most similar sentences in corpus:

(Score: 0.8879) رجل يأكل قطعة من الخبز.
(Score: 0.8353) رجل يأكل طعام.
(Score: 0.5888) رجل يركب حصان.
(Score: 0.5068) رجل يركب حصان أبيض على الأرضية.
(Score: 0.4419) امرأة تعزف على الكمان.
==========

Query: شخص ما يلبس كغوريللا يعزف على الطبول.
Top 5 most similar sentences in corpus:

(Score: 0.5975) امرأة تعزف على الكمان.
(Score: 0.5371) قرد يعزف على الطبول.
(Score: 0.3970) رجل يأكل قطعة من الخبز.
(Score: 0.3894) رجل يركب حصان.
(Score: 0.3825) رجل يركب حصان أبيض على الأرضية.
==========

Query: فهد يطارد فريسة في حقل.
Top 5 most similar sentences in corpus:

(Score: 0.6934) فهد يجري خلف فريسته.
(Score: 0.4701) مقتل فلسطينى فى مخيم عايدة فى بيت لحم.
(Score: 0.4578) توطين الرعاة فى التبت.
(Score: 0.4243) مقتل كهل فلسطينى بعد ان دهسه مستوطن بسيارته.
(Score: 0.4212) رجلين يدفعان عربات خلال الغابة.
==========

It is clear from the results above, that ArabicBERT-based model along with SBERT-paraphrase and LaBSE have the best performance in this test, followed by AraBERT-based model.

In another experiment, we have evaluated ArabicBERT-based model, AraBERT-based model and AraELECTRA-based model on a dataset of Arabic pairs of sentences [70]. It consists of 250 pairs and for each pair; a manual gold score for the similarity between the two sentences has been set. In Table 10, the results of the models on the mentioned dataset is given. For each model, Spearman's rank correlation ρ between the computed score and the gold score is calculated based on the recommendation of [71]. The evaluated models have been compared with the results of the models introduced in [66-67].

Table 10. Spearman rank correlation ρ (x 100) between the cosine similarity of sentence representations and the manual scores for Arabic dataset in [70].

	The Model	Spearman rank correlation ρ (x 100)
1	ArabicBERT	65.3
2	AraBERT	51.5
3	AraELECTRA	36.5
4	mBERT mean [68]	50.9
5	XLM-R mean [68]	25.7
6	mBERT-nli-stsb [68]	65.3
7	XLM-R-nli-stsb [68]	64.4
8	mBERT ← SBERT-nli-stsb [68]	78.8
9	DistilmBERT ← SBERT-nli-stsb [68]	77.7
10	XLM-R ← SBERT-nli-stsb [68]	79.9
11	XLM-R ← SBERT-paraphrases [68]	79.6
12	LaBSE [67]	69.1

Table 10 shows that the ArabicBERT-based model is better than AraBERT-based and AraELECTRA-based ones; and this is consistent with the results of the previously illustrated examples. While the ArabicBERT-based model has a good performance in general, the models that are based on knowledge distillation (models numbered 8, 9, 10, and 11 [66]) have better performance. However, in general, we need more robust test set to be used for evaluating the effectiveness of the models, and this will be introduced in the next year plan.

4. Conclusion

Designing and building corpora is a time consuming process that is challenging and important at the same time. Collecting dialect language increases the challenge further and in spite of all of the developments in Arabic corpora, SD corpus still in need for further contribution. This report surveyed 33 Arabic corpora to find out that, in spite of all of the developments in Arabic corpora, SD corpus still in need for further contribution. Thus, +1B words of text were collected from more than five resources. These resources included existing corpora and new text collected from websites and social media platforms to include past/recent vocabularies.

Following that, the report introduced a new pre-processing system that is incremental and scalable to new data sources. The system validated the collected data and eliminated irrelevant characters and incomplete text. The incremental propriety introduced in this system can scale to other languages in DA. As a result of this system, it was possible to design and build a new KSUSC corpus which is large in size, diverse, and contemporary. KSUSC corpus included MSA and SD, covering 25 different domains. It is a large size corpus with +1B words, +161M sentences, and +14M unique words. The SD vocabularies in KSUSC was around 50% of the total number of unique words in the corpus.

In addition to the SD corpus collected so far, this report showed different semantic models for sentence embeddings. From the empirical test, the models performance was compared with different data sets and tasks to find that SBERT produce the best results. For better evaluation of SBERT based sentence embeddings models, we are working on translating a dataset of around 1300 pair of English sentences. This set will be used to check the quality of English versions of sentence embedding models that is based on SBERT. This version is a collection of standardized sentences that have been manually assessed. For each pair of sentences, a manual score have been calculated to measure the similarity between the two sentences. The set will be translated to modern standard Arabic, Egyptian colloquial Arabic, and Saudi colloquial Arabic versions. Moreover, the collected Saudi corpus will be used to refine the Arabic corpora that we have. The new refined version will be used by SBERT methodology to generate a sentence embeddings model for Saudi dialect, and will be tested on the upcoming translated version of the dataset.

5. Reference

[1] I. A. El-Khair, "Abu El-Khair Corpus: A Modern Standard Arabic Corpus," International Journal of Recent Trends in Engineering & Research (IJRTER), vol. 2, no. 11, 2016.

[2] A. O. Al-Thubaity, "A 700M+ Arabic corpus: KACST Arabic corpus design and construction," Language Resources and Evaluation, vol. 49, no. 3, pp. 721–751, Sep. 2015, doi: 10.1007/s10579-014-9284-1.

[3] U. Nations, "Official Languages." https://www.un.org/en/sections/about-un/official-languages/index.html#:~:text=There%20are%20six%20official%20languages,%2C%20French%2C%20Russian%20and%20Spanish.

[4] M. Alruily, "Issues of dialectal Saudi twitter corpus," International Arab Journal of Information Technology, vol. 17, no. 3, pp. 367–374, 2020, doi: 10.34028/iajit/17/3/10.

[5] N. Habash, R. Eskander, and A. Hawwari, "A morphological analyzer for Egyptian Arabic," 2012, pp. 1–9.

[6] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, "AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets," Procedia Computer Science, vol. 117, pp. 63–72, 2017, doi: 10.1016/j.procs.2017.10.094.

[7] N. Y. Habash, "Introduction to Arabic natural language processing," Synthesis Lectures on Human Language Technologies, vol. 3, no. 1, pp. 1–187, 2010.

[8] A. Farghaly and K. Shaalan, "Arabic Natural Language Processing: Challenges and Solutions," ACM Transactions on Asian Language Information Processing, vol. 8, no. 4, Dec. 2009, doi: 10.1145/1644879.1644881.

[9] M. Sawalha, F. Alshargi, A. Alshdaifat, S. Yagi, and M. A. Qudah, "Construction and Annotation of the Jordan Comprehensive Contemporary Arabic Corpus (JCCA)," 2019, pp. 148–157.

[10] N. Al-Twairesh et al., "SUAR: Towards Building a Corpus for the Saudi Dialect," Procedia Computer Science, vol. 142, pp. 72–82, 2018, doi: 10.1016/j.procs.2018.10.462.

[11] H. Mubarak and K. Darwish, "Using Twitter to Collect a Multi-Dialectal Corpus of Arabic," in Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Stroudsburg, PA, USA, 2014, pp. 1–7, doi: 10.3115/v1/W14-3601.

[12] M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki, "The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus," NEMLAR Conference on Arabic Language Resources and Tools, no. September 2016, pp. 102–109, 2004.

[13] S. Khalifa, N. Habash, D. Abdulrahim, and S. Hassan, "A large scale corpus of Gulf Arabic," Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, pp. 4282–4289, 2016.

[14] S. Khalifa, N. Habash, F. Eryani, O. Obeid, D. Abdulrahim, and M. A. Kaabi, "A morphologically annotated corpus of Emirati Arabic," LREC 2018 - 11th International Conference on Language Resources and Evaluation, pp. 3839–3846, 2019.

[15] M. Jarrar, N. Habash, F. Alrimawi, D. Akra, and N. Zalmout, "Curras: an annotated corpus for the Palestinian Arabic dialect," Language Resources and Evaluation, vol. 51, no. 3, pp. 745–775, 2017, doi: 10.1007/s10579-016-9370-7.

[16] R. Al-Sabbagh and R. Girju, "YADAC: Yet another dialectal Arabic corpus," Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, pp. 2882–2889, 2012.

[17] M. Diab, N. Habash, O. Rambow, M. Altantawy, and Y. Benajiba, "COLABA: Arabic dialect annotation and processing," LREC Workshop on Semitic Language Processing, no. January 2016, pp. 66–74, 2010.

[18] R. Al-Sabbagh and R. Girju, "A supervised POS tagger for written arabic social networking corpora," 11th Conference on Natural Language Processing, KONVENS 2012: Empirical Methods in Natural Language Processing - Proceedings of the Conference on Natural Language Processing 2012, vol. 5, no. September 2012, pp. 39–52, 2012.

[19] M. Abdul-Mageed and M. Diab, "SANA: A large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis," Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, pp. 1162–1169, 2014.

[20] R. Bouchlaghem, A. Elkhlifi, and R. Faiz, "Tunisian dialect Wordnet creation and enrichment using web resources and other Wordnets," in Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Stroudsburg, PA, USA, 2014, pp. 104–113, doi: 10.3115/v1/W14-3613.

[21] M. Abdul-Mageed and M. Diab, "AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis," Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, pp. 3907–3914, 2012.

[22] A. Assiri, A. Emam, and H. Al-Dossari, "Saudi Twitter Corpus for Sentiment Analysis," International Journal of Computer and Information Engineering, vol. 10, no. 2, pp. 272–275, 2016.

[23] R. Baly et al., "Comparative Evaluation of Sentiment Analysis Methods Across Arabic Dialects," Procedia Computer Science, vol. 117, pp. 266–273, Jan. 2017, doi: 10.1016/j.procs.2017.10.118.

[24] Y. Yamamoto, "Twitter4j-a java library for the twitter api." 2014.

[25] O. Einea, A. Elnagar, and R. Al Debsi, "SANAD: Single-label Arabic News Articles Dataset for automatic text categorization," Data in Brief, vol. 25, p. 104076, 2019, doi: 10.1016/j.dib.2019.104076.

[26] I. Abu El-khair, "1.5 billion words Arabic Corpus," arXiv, 2016.

[27] K. Meskaldji, S. Chikhi, and I. Bensalem, "A New Multi Varied Arabic Corpus," in 2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS), Oct. 2018, pp. 1–5, doi: 10.1109/PAIS.2018.8598524.

[28] M. Alrabiah, A. Al-Salman, and E. Atwell, "The design and construction of the 50 million words KSUCCA," 2013, pp. 5–8.

[29] A. Elnagar and O. Einea, "BRAD 1.0: Book reviews in Arabic dataset," in 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), Dec. 2016, pp. 1–8, doi: 10.1109/AICCSA.2016.7945800.

[30] M. K. Saad and W. M. Ashour, "Osac: Open source arabic corpora," Osac: Open source arabic corpora, vol. 10, 2010.

[31] T. Zerrouki and A. Balla, "Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems," Data in brief, vol. 11, p. 147, 2017.

[32] A. Chouigui, O. B. Khiroun, and B. Elayeb, "ANT Corpus: An Arabic News Text Collection for Textual Classification," in 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), Nov. 2017, pp. 135–142, doi: 10.1109/AICCSA.2017.22.

[33] J. Tiedemann, "Parallel Data, Tools and Interfaces in OPUS," Istanbul, 2012, vol. 2012, pp. 2214–2218.

[34] A. Eisele and Y. Chen, "MultiUN: A Multilingual Corpus from United Nation Documents.," presented at the Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), 2010.

[35] P. Lison and J. Tiedemann, "Opensubtitles 2016: Extracting large parallel corpora from movie and tv subtitles," 2016.

[36] A. Elnagar, Y. S. Khalifa, and A. Einea, "Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications," in Intelligent Natural Language Processing: Trends and Applications, K. Shaalan, A. E. Hassanien, and F. Tolba, Eds. Cham: Springer International Publishing, 2018, pp. 35–52.

[37] M. El-Haj, "Habibi-a multi Dialect multi National Arabic Song Lyrics Corpus," France, 2020.

[38] S. N. Alyami and S. O. Olatunji, "Application of Support Vector Machine for Arabic Sentiment Classification Using Twitter-Based Dataset," Journal of Information & Knowledge Management, vol. 19, no. 01, p. 2040018, 2020.

[39] H. Bouamor et al., "The madar Arabic dialect corpus and lexicon," LREC 2018 - 11th International Conference on Language Resources and Evaluation, pp. 3387–3396, 2019.

[40] A. Pasha et al., "MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic," Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, no. May, pp. 1094–1101, 2014.

[41] N. Habash and O. Rambow, "Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop," 2005, pp. 573–580.

[42] N. Habash, O. Rambow, and R. Roth, "MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization," 2009, vol. 41, p. 62.

[43] M. Diab, K. Hacioglu, and D. Jurafsky, "Automated methods for processing arabic text: from tokenization to base phrase chunking," Arabic Computational Morphology: Knowledge-based and Empirical Methods. Kluwer/Springer, 2007.

[44] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A Fast and Furious Segmenter for Arabic," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2016, no. June, pp. 11–16, doi: 10.18653/v1/N16-3003.

[45] O. Obeid et al., “CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing,” in Proceedings of The 12th Language Resources and Evaluation Conference, 2020, no. May, pp. 7022–7032.

[46] M. Althobaiti, U. Kruschwitz, and M. Poesio, “AraNLP: A Java-based library for the processing of Arabic text,” in Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, 2014, pp. 4134–4138.

[47] D. A. Said, N. M. Wanas, N. M. Darwish, and N. H. Hegazy, “A Study of Text Preprocessing Tools for Arabic Text Categorization,” in The Second International Conference on Arabic Language Resources and Tools, 2009, no. January 2009, pp. 230–236.

[48] M. Attia, “A large-scale computational processor of the Arabic morphology, and applications,” Faculty of engineering, Cairo university, Egypt, 2000.

[49] K. M. Darwish, “Probabilistic methods for searching OCR-degraded Arabic text.,” 2004.

[50] N. Habash and O. Rambow, “MAGEAD: a morphological analyzer and generator for the Arabic dialects,” 2006, pp. 681–688.

[51] YouTube, “CommentThreads: YouTube Data API,” 2020. https://developers.google.com/youtube/v3/docs/commentThreads.

[52] FaceBook, “Graph API: FaceBook for Developers,” 2020. https://developers.facebook.com/docs/graph-api/.

[53] FindMyFBID, “Find your Facebook ID.” https://findmyfbid.com/.

[54] SAIP, “Copyright Law Issued by Royal Decree No. M / 41 dated 2/7/1424 AH Amended by the Council of Ministers Resolution No. (536) dated 19/10/1439 AH,” Saudi Authority for Intellectual Property. https://www.saip.gov.sa/wp-content/uploads/2019/10/Copyright-Law.pdf.

[55] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[56] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

[57] Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. "Albert: A lite bert for self-supervised learning of language representations." arXiv preprint arXiv:1909.11942 (2019).

[58] Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. "Electra: Pre-training text encoders as discriminators rather than generators." arXiv preprint arXiv:2003.10555 (2020).

[59] Polignano, Marco, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, and Valerio Basile. "Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets." In 6th Italian Conference on Computational Linguistics, CLiC-it 2019, vol. 2481, pp. 1-6. CEUR, 2019.

[60] Masala, Mihai, Stefan Ruseti, and Mihai Dascalu. "RoBERT–A Romanian BERT Model." In Proceedings of the 28th International Conference on Computational Linguistics, pp. 6626-6637. 2020.

[61] Safaya, Ali, Moutasem Abdullatif, and Deniz Yuret. "Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media." In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 2054-2059. 2020.

[62] Baly, Fady, and Hazem Hajj. "AraBERT: Transformer-based model for Arabic language understanding." In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pp. 9-15. 2020.

[63] Antoun, Wissam, Fady Baly, and Hazem Hajj. "AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding." arXiv preprint arXiv:2012.15516 (2020).

[64] Pedro Javier Ortiz Su´arez, Laurent Romary, and Benoˆıt Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1703–1714, Online, July. Association for Computational Linguistics.

[65] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).

[66] Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning Distributed Representations of Sentences from Unlabelled Data. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1367–1377, San Diego, California. Association for Computational Linguistics.

[67] Reimers, Nils, and Iryna Gurevych. "Making monolingual sentence embeddings multilingual using knowledge distillation." arXiv preprint arXiv:2004.09813 (2020).

[68] Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. "Language-agnostic bert sentence embedding." arXiv preprint arXiv:2007.01852 (2020).

[69] Lample, Guillaume, and Alexis Conneau. "Cross-lingual language model pretraining." arXiv preprint arXiv:1901.07291 (2019).

[70] Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France. European Language Resources Association (ELRA).

[71] Cer, Daniel, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation." arXiv preprint arXiv:1708.00055 (2017).

[72] Reimers, Nils, Philip Beyer, and Iryna Gurevych. "Task-oriented intrinsic evaluation of semantic textual similarity." In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 87-96. 2016.

Appendix A – KSUSC Detailed Statistics

Source	Language	Category	Time	No. of Unique Words	No. of Sentences	No. of Words
Pre-Existing Dataset	SD	General	2017	220,643	381,709	3,444,749
		General	2018	5,250	3,953	21,634
		Poetry	2020	62,595	99,125	477,770
		Politics	2020	22,810	7,417	95,481
		Social	2018	24,420	10,358	103,327
	MSA	Culture	2017	52,736	11,350	345,937
		Technology	2017	57,504	24,532	680,140
		Sport	2017	75,923	44,173	1,460,770
		News	2017	144,306	158,439	4,560,388
		Economy	2017	49,535	26,684	827,246
		hotelReview	2016	282,899	512,512	11,029,537
		Acting	2020	2,543,126	64,164,808	373,045,451
		General	2010	131,984	148,668	2,049,168
		General	2013	1,734,837	69,699,319	351,853,997
		General	2018	570,132	10,402,957	234,370,291
Website	SD	Economy	2020	97,791	255,565	2,476,153
	MSA	Medicine	2018	103,682	437,011	8,886,837
		News	2018	226,147	112,068	2,251,042
		General	2011	1,433,804	100,113	21,044,828
YouTube	SD	Acting	2020	141,894	214,421	1,926,650
		carReview	2020	190,485	251,437	3,858,461
		Economy	2020	45,452	18,655	296,064
		eGames	2020	495,436	2,401,397	18,667,161
		Entertainment	2020	238,918	490,526	4,457,690
		Magic	2020	58,450	65,064	555,379
		MakeUp	2020	55,393	37,453	441,649
		Media	2020	644,047	1,383,250	17,070,306
		General	2020	110,800	106,465	995,822
		News	2020	158,247	108,220	1,400,449
		Poetry	2020	6,798	2,067	17,859
		Politics	2020	15,325	4,748	62,413
		Religion	2020	72,616	82,043	891,185
		Songs	2020	405,908	1,158,668	10,412,711
		Sport	2020	220,248	349,165	3,975,687
		Technology	2020	94,785	150,694	1,840,975
		Vlog	2020	856,894	3,850,130	35,493,114
	Mixed	Religion	2020	198,810	232,131	4,016,044
Twitter	SD	Acting	2020	19,374	6,491	57,696
		Economy	2020	14,473	8,528	94,534
		eGames	2020	45,464	58,468	626,587
		HashTag	2020	20,124	9,630	82,683
		History	2020	3,606	606	6,730
		Literature	2020	3,527	1,477	8,864
		Media	2020	57,556	58,728	721,909
		Medicine	2020	8,604	2,156	22,760
		General	2020	301,142	501,076	4,853,438
		Music	2020	2,887	1,089	13,012
		News	2020	134,916	178,791	2,124,959
		Poetry	2020	4,748	1,216	9,266

Source	Language	Category	Time	No. of Unique Words	No. of Sentences	No. of Words
		Politics	2020	12,669	9,560	131,335
		Religion	2020	6,111	4,256	49,634
		Social	2020	63,760	93,446	1,208,296
		Sport	2020	56,051	67,863	829,431
		Technology	2020	14,376	6,044	53,268
	Mixed	Economy	2020	2,369	1,342	20,056
		Hashtag	2020	904	307	2,132
		Media	2020	4,171	1,340	10,408
		Medicine	2020	5,570	1,905	13,972
		News	2020	48,891	27,553	336,968
		Religion	2020	3,919	852	8,514
		Sport	2020	9,715	5,728	55,562
Facebook	MSA	Cooking	2018	20,852	13,412	112,411
		Economy	2018	98,465	248,226	10,666,947
		Literature	2018	76,101	18,997	576,846
		News	2018	111,391	63,026	1,513,749
		Sport	2018	315,594	783,451	7,539,048
	Mixed	Acting	2018	6,038	1,289	20,498
		Culture	2018	7,313	2,394	41,547
		Economy	2018	87,463	98,739	1,422,899
		Media	2018	23,322	22,993	310,189
		Medicine	2018	64,813	71,616	1,008,600
		General	2018	362,386	944,522	10,656,985
		News	2018	159,900	188,972	2,274,879
		Religion	2018	89,348	606,274	8,582,885
		Sport	2018	117,848	174,813	1,667,435
		Technology	2018	4,356	1,196	15,303
Total				**14,240,747**	**161,795,667**	**1,183,156,600**

Phase#4

Developing video content recognition system

Rajab 1442 – Feb 2021

Table of Contents

1. Introduction

Video content recognition is a comprehensive problem, in which multiple components such as static scenes and objects and dynamic events and human actions should be addressed in integrated ways. Unfortunately, most of the research efforts in video understanding in the past years touched isolated sub-aspects of the problem instead of providing comprehensive solutions to tackle the actual problem. Some of those efforts focused on human activity recognition and others focused on object detection or recognition and so on. Even though several methods have been proposed for action recognition in literature, action recognition is still a challenging task due to several reasons such as the optimal way for deriving the temporal features and the computational costs. Action recognition can be improved either by improving the feature extraction which in turn requires new model components to be developed, or by combining the current models with some neural network component for providing better fused features.

Recently, the research pioneers in AI and computer vision such as Google and Facebook launched the new direction for comprehensive video understanding, which is still in its early stages. A comprehensive video understanding dataset called holistic video understanding (HVU) was recently published to support the new direction of research. It is structured in a hierarchal way, where it has six main categories of visual components, actions, objects, events, scenes, attributes, and concepts. Under each category there are large number of classes. This dataset is publicly available in two versions the full version and a mini version.

In this work we address the problem of video understanding in two ways. In one hand we try to enhance the performance of action recognition techniques on well-known benchmarks such as Kinetics dataset via

- Employing a variational feature learning on top of action recognition models for better discriminate inter/intra action categories.
- Feature fusion of the action recognition with the classification features of the corresponding frames

while on the other hand, we try to generalize the solution toward a comprehensive video understanding.

Different architectures were developed and evaluated separately on two datasets, Kinetics-400 and HVU dataset.

2. Literature Review

2.1 Literature Review of Action Recognition

Action recognition has been researched in many papers under various titles such as video analysis or understanding, human action or activity recognition and video events captioning. Most of these research works focus on developing robust techniques to encode the evolution of the events over a sequence of frames and then, predicting a label from a discrete set of action categories as in [1, 2]. The drawback of these works is that there is no detailed description for the action in the output. To overcome this lack of details in the detected action, other subsequent works explore explaining video semantics using sentence descriptions as in [3, 4]. Typically, a short video clip contains multiple overlapped actions, with high variation in their durations. Some of the actions might span across the entire clip, while others might take place in few seconds. Temporal proposal is the most common technique investigated in different ways for action detection and localization. It is inspired by the object proposal techniques, which are the state-of-the-art in object detection and image captioning. For action proposal, a sliding window of varying size is typically used to generate multiple segment proposals of different volumes through the input video sequence. The drawback of this technique is the high computational cost of applying sliding windows of multiple sizes. An important enhancement for the temporal proposals is the deep action proposals (DAPs) [5]. Instead of applying windows of different sizes on the input, the authors encode a stream of visual observations via 3DCNN followed by LSTM sequence encoder into a sequence of discriminative hidden states. A linear combination of the last state in the sequence encoder is then used to generate multiple proposals of different sizes as well as confidence values of including action within the temporal extent of each proposed segment.

Even though the mentioned techniques might perform well with a single action per shot or at most with sequentially ordered actions, it fail in the identification and description of multiple overlapped actions. Dense-captioning events in videos is a new approach, that tries localizing and describing all the events in the video. For instance, the work of Krishna et al. [6] extends the

DAPs technique [5] of temporal proposal generation to detect events in short as well as long video segments. To achieve that the 3DCNN features of the input sequence are sampled at different strides to capture events of different lengths. Furthermore, in the event captioning phase the context from the surrounding proposals is investigated as it is observed that events in a given video are almost highly correlated.

2.2 Literature Review of Video Datasets

To build the models for recognizing the video content we need to have a video dataset that will include general videos and local videos. For the local videos part, we have to build it ourselves. For the general part we will use one or more of the datasets used by researchers in the field, so we can compare the video recognition techniques that we will use with the techniques used by other researcher and be able to publish our work.

Several benchmark datasets are available for video analysis and understanding researches such as MDB51 [7], UCF101 [8], YouTube 8M, Activity Net, TRECVID, HVU. We looked at many datasets and Table 1 present some of the major datasets and a comparison between them. From this evaluation we selected 4 datasets: TRECVID, ActivityNet, Kinetics, and HVU. They were the best suitable to our research. In the following we will describe these 4 datasets and the dataset YouTube-8M due its importance.

YouTube-8M

YouTube-8M is a huge multilabel video classification dataset. It consists of nearly 8M YouTube video IDs with a total time duration of 350K hours [9]. It is annotated by more than 3800 vocabularies of visual entities with 3 labels per video in average. Figure 1 shows the top 200 entities in YouTube-8M. The first round of annotation process was machine-generated, which was followed by other verification rounds. The verification rounds used different strategies, including asking human raters whether the labels are visually recognizable or not. The dataset is available for download in the form of precomputed audio-visual features from billions of frames and audio segments to reduce the required storage size and enable training starter model on this dataset on limited computation resources.

YouTube-8M Segments dataset is an extension for YouTube-8M dataset, with human-verified segment-level annotations [9]. The occurrence times of entities are localized in this version. It

contains verified labels for 237K segments on 1000 classes from the validation set of the YouTube-8M dataset, with an average of 5 segments per video.

Figure 1. The top 200 entities in You Tube-8M dataset. Font size is proportional to the number of videos labeled with the entity [10].

Table 1: Comparison between the datasets

	Dataset	Usage	Collected From	Type (Video/Images)	# of Classes	Type Features	# Videos	# Video SHOTS	Nbr Hours	Min Video Len
1	Trecvid 2010	Licensed + Free	IACC.1.tv10.training		130	raw vid	3127	118205	198	211
2	Trecvid 2011	Licensed + Free	2010 train + 2010 test		346	raw vid	3127+8358	118205+144757	198+220	11
3	Trecvid 2012		2011 train + 2011 test		346	raw vid	3127+8216	118205+137327	198+218	11
4	Trecvid 2013		2012 train + 2012 test		346	raw vid	3127+8263	118205+145634	198+221	11
5	Trecvid 2014		2013 train		346	raw vid	2407	110947	199	11
6	Trecvid 2015		2013 train		346	raw vid	2407	110947	199	11
7	YouTube 8M	Free	youtube	features	3862	tensorflow	6.1Millions		350000Hours	
8	MSR-VTT	Free	commercial vid eng	videos		raw vid	10K clips		42.1Hours	
9	Activity Net	Free		videos	432	raw vid	15K clips			
10	MRVD corpus	Free		videos		raw vid				
11	CamVID				32	raw vid			10 min	
12	NYU Dataset			vid /Img	1000cl/23 scenes	raw vid				
13	Wild8		documentary series	vid /Img	8	raw vid	100seq			
14	Tyoutube Obj Dataset				10	raw vid	126vid			30sec
15	SUNY		xigh.org	vid /Img	24	raw vid	8			
16	VSB100					raw vid	100			
17	SegTrack V2 Dataset					raw vid	14			
18	Cityscapes Dataset				30	raw vid				
19	HVU	free	youtube	videos	419	raw vid	572k			

Max Video Len	Annotation Level	# Annotations	# Annotators	# of Annotations / per video	Avg # of Annotations / per video	Bounding Boxes	Multi Language	Metrics Used	
248	Frame shot(Best frame)	per video shot		1 per shot		Yes but not for SIN	YES	MAP	
211	Frame shot(Best frame)	per video shot		1 per shot		Yes but not for SIN	YES	MAP	
211	Frame shot(Best frame)	per video shot		1 per shot		Yes but not for SIN	YES		
211	Frame shot(Best frame)	per video shot		1 per shot		Yes but not for SIN	YES		
211	Frame shot(Best frame)	per video shot		1 per shot		Yes but not for SIN	YES		
211	Frame shot(Best frame)	per video shot		1 per shot		Yes but not for SIN	YES		
	action per video			1 to 23 sentences	3.01	-	Yes		
	per video	200K	1327	20 sentences		No			
		52.1K				157K per split			
	sentence per video	120K				No	bilingual		
	pixel level	700 frames							
	pixel level	33 seq							
3min	BB				1class / 10 frames	frame Level			
	pixel level								
	pixel level		4						
	pixel level	One per Vid							
	pixel level								
10s	per video					no	no		

TRECVID

The TREC conference series is sponsored by the National Institute of Standards and Technology (NIST) with additional support from other U.S. government agencies. The goal of the conference series is to encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. In 2001 and 2002 the TREC series sponsored a video "track" devoted to research in automatic segmentation, indexing, and content-based retrieval of digital video. Beginning in 2003, this track became an independent evaluation (TRECVID) with a workshop taking place just before TREC. [11]

The TRECVID datasets are a set of collected videos from diverse sources (Internet Archives HAVIC, Vine, BBC EastEnders, CNN, CCTV, LBC, NBC, etc. ...), and have diverse lengths and sizes.

Video recording quality also differs between one video and another, some videos are from the 1970's, as presented in Figure 2.

The concepts were defined and assigned not to the whole video, but to representative Key Frames called (RKF), thus one video might contain one or more RKF, and each RKF contains only one concept, even if the frame does contain other information, but the most dominant concept is singularly selected.

In each Semantic TRECVID task, the organizers were focusing on the RKF classification. Given a set of RKF frames from videos in the training and validation sets, competitors were asked to classify the test RKF frames to the right concept, as shown in Figure 3.

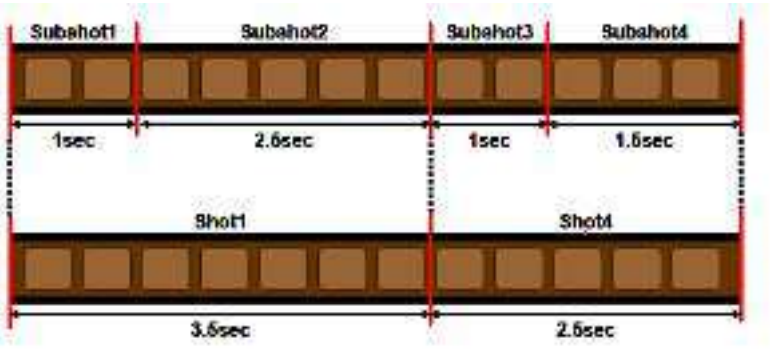


Figure 3. Visualization of the composition of shots and subshots in TRECVID data [35].

In Figure 4, a list of the TRECVID tasks per year and over years is presented, the different tasks appear and disappear over the years, depending on the research directions of the TRECVID team.

In some tasks, just some

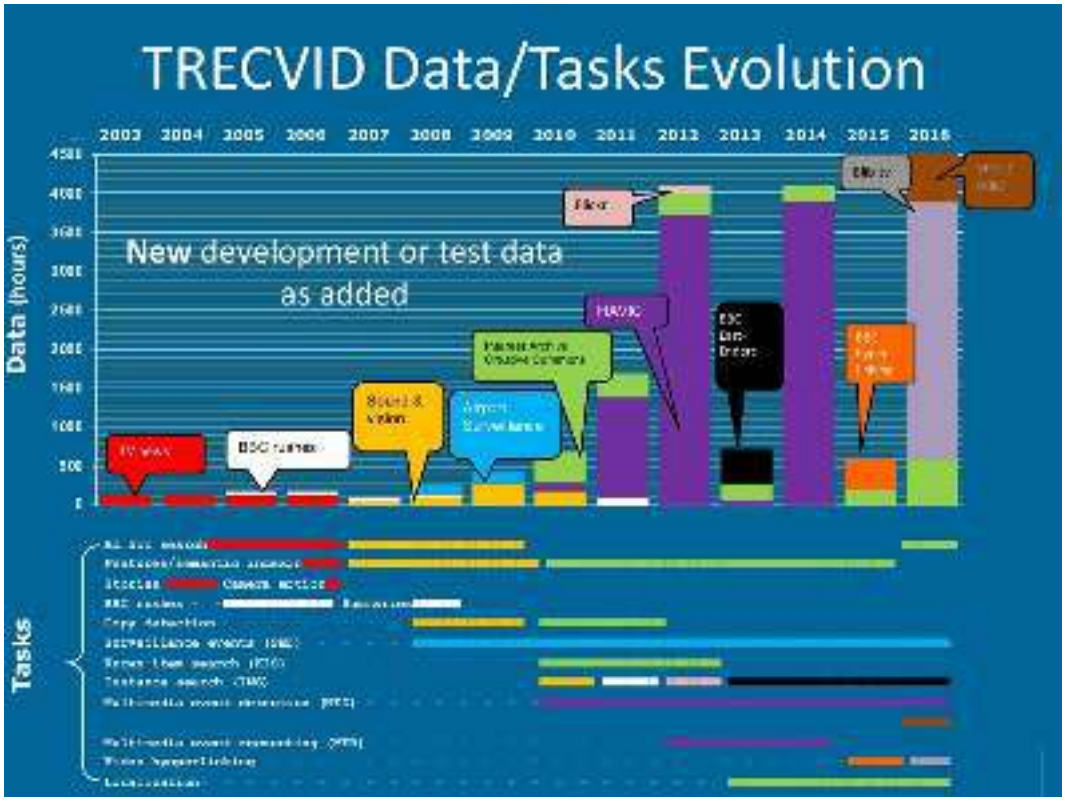


Figure 4. TRECVID list of tasks over the years (2003-2016) [36]

selected concepts, were used for the competition, even if the competitors gave the results for all the required RKF.

Most of the videos had copyrights, and require signing applications, beside the HVAC Internet Archives that were mostly free for download and few videos were removed over time for copyright issues. Here is a list of the Internet Archives (IACC) [12] [13], as listed in Table 2.

Table 2: Internet Archives Videos used in TRECVID over the years 2010-2015

Collection (slice)	Total Duration (h)	video files	min/mean/max duration (s)	Video shots	Mean duration (s)	used for training	used for test
IACC.1.tv10.training	198	3127	211/228/248	118205	6.04	2010-2015	-
IACC.1.A	220	8358	11/95/211	144757	5.48	2011-2015	2010
IACC.1.B	218	8216	11/96/211	137327	5.72	2012-2015	2011
IACC.1.C	221	8263	11/96/211	145634	5.46	2013-2015	2012
IACC.2.A	199	2407	10/297/387	110947	6.46	-	2013
IACC.2.B	197	2368	10/299/387	106611	6.65	-	2013-2014
IACC.2.C	199	2395	10/298/387	113046	6.32	-	2013-2015
Total	**1452**	**35134**	**10/149/387**	**876527**	**5.97**		

LSCOM

Semantic concept detection represents a key requirement in accessing large collections of digital images/videos. Automatic detection of presence of a large number of semantic concepts,

such as “person,” or “waterfront,” or “explosion”, allows intuitive indexing and retrieval of visual content at the semantic level. Development of effective concept detectors and systematic evaluation methods has become an active research topic in recent years. For example, a major video retrieval benchmarking event, NIST TRECVID [11], has contributed to this emerging area through

-the provision of large sets of common data

-the organization of common benchmark tasks to perform over this data. [14] [15]

The LSCOM concepts selection went through phases over the years, and each couple of years, new concepts were added. The TRECVID concepts were inspired from the LSCOM 500. The last LSCOM version 1.0 contained nearly 856 concepts. Figure 5 shows some concepts from the broadcast news, also called LSCOM Lite.

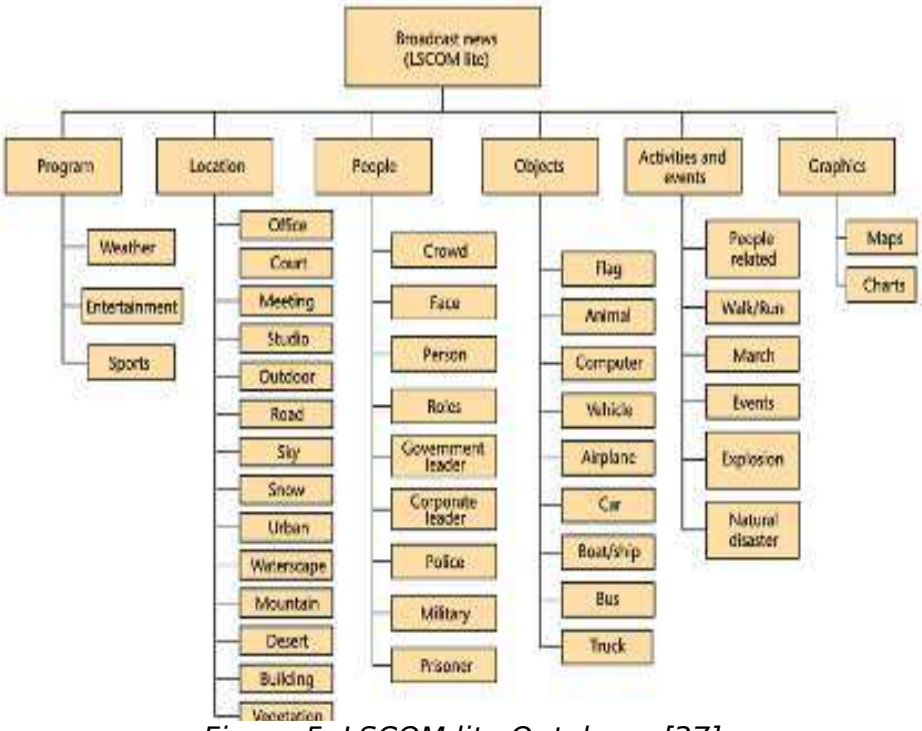


Figure 5. LSCOM lite Ontology. [37]

ActivityNet Captions

ActivityNet Captions is another large-scale dataset for dense-captioning events [16]. It contains 20k videos, with a total time of 849 hours and 100k total descriptions. Each description is a

textual sentence describing an event that occurs in a specific segment in a video. The average length of each sentence is 13.48 words. Each video is described by a paragraph of 3.65 sentences in average. The average length of segments is 36 seconds, while the average length of videos is 180 seconds. This benchmark is also available for researchers in the form of YouTube video IDs, segments' time stamps and textual descriptions.

ActivityNet Entities

The ActivityNet-Entities is based on the video description dataset ActivityNet Captions and augments it with 158k bounding box annotations, each grounding annotation is a noun phrase (NP). The noun phrases in these annotations are based on ActivityNet Captions, which are linked to videos in ActivityNet 1.3.

This dataset consists of 10k, 2.5k, and 2.5k videos for training, validation, and testing, respectively. It also encompasses 35k, 8.6k, and 8.5k event segments and sentence descriptions in addition to 105k, 26.5k, and 26.1k bounding box annotations on each split [17].

An advantage of this dataset is the multi annotation of videos, each video segment has a token that describes the action in the segment as well as bounding boxes for some of the objects in that segment. Table 3 illustrates the annotations in sample frames from three consecutive segments in a video.

Table 3. Annotation samples from ActivityNet entities dataset

	Object bounding boxes	**Action annotation**
First segment		"**A woman is shown riding a camel past pyramid in Egypt**"
Second segment		"**The camel walks as the woman leans forward**"

Third segment		**"A woman is shown riding a camel past pyramid in Egypt"**

Kinetics

It is a large-scale dataset of up to 650,000 video clips. The dataset is available in three versions, Kinetics 400, Kinetics 600, and Kinetics 700. Depending on the dataset version, the collection of videos covers 400, 600, or 700 action classes [19]. This dataset is mostly focused on human actions, which include human actions such as laughing, human-human actions such as shaking hands, and human-object actions such as washing dishes [20].

The dataset is not hierarchically organized even though it contains several parent-child groupings. The dataset consists of three folds, one for training with 250–1000 videos per class, other fold for validation with 50 videos per class and another fold for testing with 100 videos per class [20].

Four stages were used to create the Kinetics dataset. The first one is searching the video clips in YouTube corpus by matching video titles with the Kinetics actions list [20]. In the second stage, a temporal positioning for the action in the videos was performed by applying a pretrained image classifier on the video frames. Then a manual labeling was performed, in the third stage, to verify that the corresponding actions are actually occurring during the timely stamped clips, where Amazon's Mechanical Turk (AMT) service was used for this purpose. All the videos added to the dataset have at least three positive responses out of five from the AMT workers. In the final stage intensive work was performed to remove duplicated and noisy videos [20].

HVU

Contrary to the previous research work which were focusing on highly specific video understanding task, Holistic Video Understanding (HVU) dataset is aimed to describe the entire content of a video by reframing the comprehensive problem of video understanding as a multi

label and multitask recognition in a dynamic scene [18]. For that reason, the HVU is organized hierarchically based on the semantic taxonomy. HVU contains 572k videos with a total of 9 million annotations for training, validation and test set spanning over 3457 labels. It encompasses various categories of semantics that capture real world scenarios. These categories include objects, scenes, events, attributes actions and concepts. The total number of HVU labels is distributed among these categories as follows:

- Objects include 1917 labels,
- Scenes include 282 labels,
- Events include 77 labels,
- Attributes include 106 labels,
- Actions include 882 labels,
- Concepts include 193 labels.

A detailed statistic of the HVU training dataset is given in Table 4. This dataset is supported by rich annotations of 2112 annotations per label in average, which is a reasonable amount of training.

Table 4. Statistics of the HVU training set for different categories [38]

Task Category	Scene	Object	Action	Event	Attribute	Concept	Total
#Labels	282	1917	882	77	106	193	3457
#Annotations	733,332	3,717,455	1,005,954	450,776	380,921	904,514	7,192,952
#Videos	242,908	469,141	454,592	224,940	285,811	371,438	475,797

Manually annotating a huge dataset such as HVU with multiple semantic categories is a very challenging task. It has two main difficulties:

1- It is error prone as a human cannot pay attention to every detail occurring in the video, which leads to high rate of mislabeling.

2- The annotation of large-scale video is very time consuming due to the time duration and the number of videos.

HVU was annotated in a semi-automatic way of two rounds to overcome these difficulties. In the first round, rough video annotation for 30 tags was performed via utilizing Google Vision AI and Sensifai Video Tagging API. The second round was a human verification, which was performed by three teams in multiple stages. These teams were responsible for building the semantic taxonomy based on the definition of the predicted tags in the first round. To achieve that, the teams were asked to classify the predicted tags in the first round into the six semantic categories. They were also responsible for adding any possible missed tags as well as removing any mislabeled noisy tags. This incorporation of the machine generated tags and human verification makes the HVU the most diverse dataset with clean annotations. As illustrated in Figure 6 there is nearly a balanced distribution for the samples among different tags.

Figure 6. The t-SNE relationship based on label co-occurrence, without using video content *[18]*.

Dataset Selection

To select one dataset out of the four selected we put the following conditions:

1) Include as much as possible of our concepts' classes.
2) Include as much as possible items in each class.
3) Variation of items in each class.
4) Easy to localize (by adding videos to it)
5) Easy to Arabize (by Arabizing its notation)

6) Large scale (in terms of number of hours)
7) Multilabel per video
8) Easy to build a system based on it (not long videos)
9) Most of it useful in our system (i.e. should cater to the local use)
10) Used by others and has published results using it
11) Has models available for it in the net.

Table 5 present a comparison between the four datasets (ActivityNet has two versions ActivityNet-Captions and ActivityNet-Entities).

By comparing the four datasets and according to our above conditions of selection we selected HVU to start with, because it is the newest one and built for similar research, has many labels, and has the largest number of videos, and has the largest number of classes. In addition, we can claim that from the high number of annotations viewpoint, the HVU dataset is considered as an extremely valuable and unique dataset for video content understanding.

HVU developers have issued two versions of the dataset a Full HVU version (train, test, val) and a Mini-HVU version (train, test, val.), nevertheless the number of videos was still immense and had to be reduced to make a pilot study. Table 6 shows the number of HVU videos in the full and mini versions as per the HVU website.

Table 5. Summarization of four different datasets

Kinetics	2017	actions	10 seconds	Whole video	Yes	N/A

Kinetics		TRECVID	ActivityNet-Captions	ActivityNet-Entities	HVU
Textual	**Year published**	2010-20XX	2017	2018	2019
Yes	**Categories included**	Concepts including Events, actions, scenes, objects	Events	classes, annotations and objects	6 main categories: scene, object, action, event, attribute, and concept
Yes	**Max, Min, Avg Length of each video**	Training:211/228/248sec; Testing 11/95/211sec	Max: 4 minutes and Avg. 36 seconds	not mentioned	the duration of the videos is different with a maximum of 10 seconds length
Yes	**Labeling for whole video or segments**	Only Some RKF of each video shot is labelled	Segment-based labeling	selected segment	trimmed video clips
Up to 650k videos	**Are all videos in DB annotated**	Yes	Yes	yes	YES
	Multi-labeling	Many RKF contain diverse concepts at the same time	NA	yes	YES (multi-task, multi-label video benchmark dataset with comprehensive tasks and annotations"

	TRECVID	ActivityNet-Captions	ActivityNet-Entities	HVU
Labeling Format	Textual	Textual dense description	Action textual description and object bounding box	Multi labelled for different semantic concepts
Used in papers	Yes	Yes	Grounded Video Description	only in this paper: https://arxiv.org/pdf/1904.11451v2.pdf
Used in competitions	Yes	Yes	Task A – ActivityNet Entities Object Localization	YES
Availability of models using the DB	Semantic Indexing No; other tasks	There is some	no	there is one model called HatNet but the code not available
Total number of hours	Training: 198 Testing 1200	849	648	HVU consists of 572k videos

Table 6. Number of HVU videos in the full and mini HVU

Train	# Videos in Full	# Videos in Mini
	481417	129627

Val	31171	10057
Total	515588	139684

Selection of a small subset of HVU

HVU is a quite big dataset, with specific classes as per the HVU initiators. The defined categories/tags do not specifically fit to any semantic problem but covers a wide range of possibilities. In order to make a pilot study or proof the concept for our project for the Saudi local videos, we will select a restricted set of categories, consequently reducing thereby the number of videos.

We decided to select very small number of videos to build a pilot system. This will allow us a jumpstart in selecting and developing the best techniques plus this work will help us in building the local video part of the database. The selection included actions, objects, scenes, concepts, and events.

We did this by first analyzing the actions part of the HVU. From this analysis we arrived at:

News: 16 actions

Sport: 210 actions

Human Body Actions: 271 actions

Education: 4 actions

Entertainment: 100 actions

General: 114 actions

Fashion: 16 actions

Commercial: 76 actions

Not Clear: 4 actions

We decided to concentrate on one or two domains to build a system that will show that our framework will accomplish the desired goal of the project. We decided that news and sport are

the best two candidates. From this analysis we can see that sport has a large number of videos in HVU and hence is really an excellent candidate.
We selected the HVU tags (actions, objects, scenes, concepts, and events.) that are well known in Saudi Arabia and the Arabic world and hence will be beneficial to our research.

We went into many iterations of selecting the tags then reducing number of selected tags until we reached 84 tags from the complete set. Unfortunately, some of these tags were not in the mini version so the number of selected tags was reduced to 75. The number of the selected tags for each HVU category is presented in the Table 7 below and the selected tags are presented in Table 8. From Table 7 we can see that the selected tags constitute 2.4% of the HVU dataset.

Table 7. Number of selected tags for each category of HVU

Categories	HVU	Our Selection
action	739	22
object	1678	32
attribute	117	0
scene	248	13
concept	291	5
event	69	3
	3142	**75**

When downloading we found that some of the videos listed in the HVU were removed from YouTube. Table 9 present the number of HVU videos, selected videos, downloaded videos and removed videos.

Table 8. Our selected tags sorted by Category

Tag	Category	Tag	Category	Tag	Category
applauding	action	aircraft	object	crowd	concept
basketball_moves	action	athlete	object	emergency	concept
bowling	action	car	object	news	concept
boxing	action	door	object	police	concept
gymnastics	action	drink	object	team	concept
jumping_jacks	action	face	object	demonstration	event
kicking_soccer_ball	action	fence	object	disaster	event
news_anchoring	action	fire	object	news_conference	event
playing_basketball	action	flag	object	basketball_court	Scene
playing_volleyball	action	grass	object	beach	scene
presenting_weather_forecast	action	hand	object	building	scene
public_speaking	action	helmet	object	field	scene
running	action	man	object	forest	scene
shooting_goal_soccer_	action	microphone	object	house	scene
snorkeling	action	newscaster	object	mountain	scene
speaker	action	newspaper	object	park	scene
sports_training	action	pedestrian	object	restaurant	scene
strength_training	action	player	object	room	scene
swimming	action	red_carpet	object	stadium	scene
water_skiing	action	screen	object	street	scene
windsurfing	action	shirt	object	swimming_pool	scene
wrestling	action	sign	object		
		soccer	object		
		soccer_ball	object		
		soil	object		
		sportswear	object		
		swimmer	object		
		television_reporter	object		
		umbrella	object		
		wall	object		
		water	object		
		woman	object		

Table 9: Number of HVU videos, selected videos, downloaded videos and removed videos

	HVU Videos	**Selected Number of Videos**	**Downloaded Videos**	**Videos removed from YouTube**
Train	481417	105648	98382	7266
Val	31171	8070	7526	544
Total	515588	113178	105944	7810

We used Kinetics-400 dataset instead HVU in some of our work, due to some issues faced with the HVU dataset, as there are no published comparisons on this dataset. Other issues were found in many video samples such as missed, damaged and very short videos. All these issues were not

clearly reported in the available documentation of the dataset and furthermore, we could not get any response from the owners of that dataset.

3. Proposed Methods and Experiments On Kinetics Dataset

3.1 Variational Feature Learning

Variational feature learning have been applied in several application such as image generation [21] and object re-identification [22]. In [21], authors have applied variational auto-encoder for MNIST classification in an unsupervised way. While in [22], the authors have applied variational features for vehicle re-identification in a supervised manner. This shows the positive impact that can be made by variational features modelled as a gaussian distribution from the raw CNN features.

In respect of the CNN fusion for obtaining better feature representations, there are several methods in literatures that were proposed to fuse different features hierarchy from same model but with different layers, of from different branches. In [23], the authors have proposed a neural network module (MMTM) for leveraging the knowledge from multiple models in CNNs. CNN fusion idea has been applied widely for object detection [24]–[26] where the fused features boost the bounding box localization and the classes prediction.

Moreover, CNN fusion is applied for action recognition in [27], where the authors summed up there works in the following points:

(i) instead of fuse CNN feature at the Softmax layer, the feature fusion can be applied at a convolution layer without loss of model performance, but with a substantial saving in parameters,

(ii) It is better to fuse networks spatially at the last convolutional layer than earlier, and that additionally fusing at the class prediction layer can boost accuracy.

(iii) Pooling of abstract convolutional features over spatiotemporal neighborhoods further boosts performance.

Based on the above-mentioned studies, we try to adopt these methods but with some modifications which can help to improve the action recognition performance.

3.2 Variational Feature Learning for Action Recognition

Variational feature models have shown a remarkable boost in both the supervised learning and unsupervised learning. Thus, we have an attempt to adopt the VFL [22] module on top of action recognition models. The general idea of the Variational Feature Learning is that the gaussian distribution (Means μ and the standard deviation σ) is modeled from the raw features of the convolutional neural network. This distribution ensures the variation among the classes. Moreover, the derived features are more discriminating which keep the variation for both inter/intra classes.

As shown in Figure 7, the variation and means of the CNN features are generated by two fully connected layers. Then the classifier is used on top of the means layer. The means layer features are better normalized, compact, and more representative.

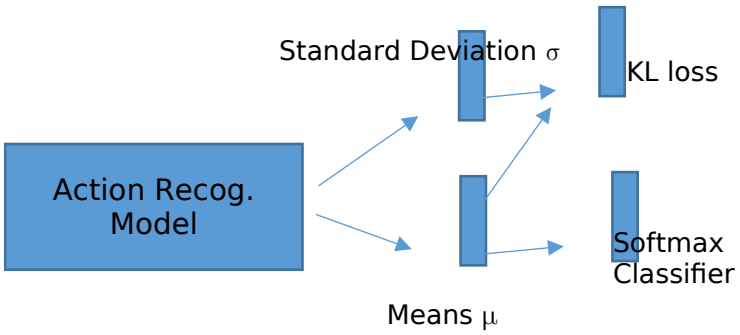


Figure 7. Variational Feature Learning for Action Recognition.

The loss function consists of two parts, first part is the classification loss of the action categories using cross-entropy (used Softmax classifier), while the second term represent the KL loss which compute the KL divergence between Gaussian distribution of the prior $N(0,1)$ and the posterior distribution $N(\mu,\sigma)$ as in equation (1).

$$L = L_{cls} + \alpha D_{KL}(N(\mu,\sigma), N(0,1)) \quad (1)$$

where α controls the contribution of the KL loss on the total loss which is empirically recommended to be between 0 and 1 based on the best results reported.

3.2.1 Feature Fusion of action recognition features with CNN classification

The proposed method in this part fuses action recognition features of any model with the extracted CNN features of the same video clip. We used image classification model trained on ImageNet [28] for extracting the features. Then, a new small model that consists of 7 1D convolutional layers is designed to extract the temporal knowledge among the sequential frames, as shown in Figure 8.

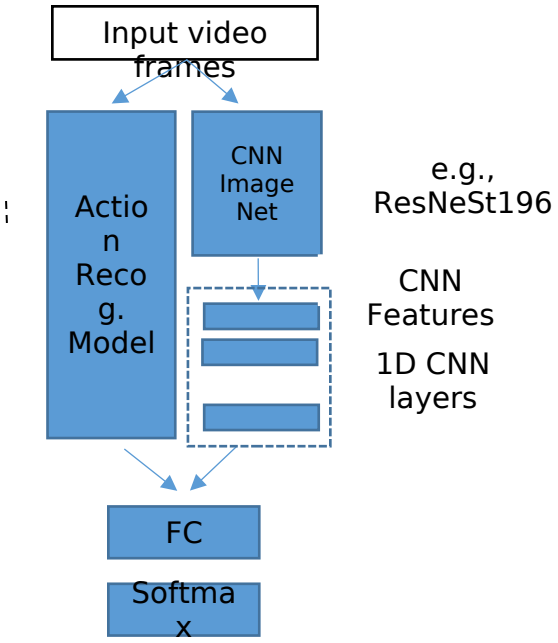


Figure 8. The proposed CNN fusion model for action recognition.

Table 10: The proposed 1D block for CNN feature classification.

No	Layer	Input Channels	Kernel Size	Stride	Output Channels

1	1D Convolution	16	7	5	32
2	1D Convolution	32	3	2	64
3	1D Convolution	64	3	2	128
4	1D Convolution	128	3	2	256
5	1D Convolution	256	3	2	512
6	1D Convolution	512	3	2	1024
7	1D Convolution	1024	3	2	2048
8	Global average pooling				
9	Flatten layer				
10	Dropout	With rate (0.5)			

A 1D convolution block consists of 7 1D convolutional layers. First layer uses kernel size of 7 and stride 5 while other layers use kernels of size 3 and stride 2. We did not use pooling layers in order to preserve the classification features. Instead of using pooling layers we keep applying stride of 2 in each 1D layer till we obtain feature map of depth 2048, then we applied global average pooling to finally obtain feature vector of size 2048. Table 10 shows the 1D block and its layers' characteristics. Input Channels in first layer in respect to number of frames of the video instance.

The output of the 1D block is concatenated with the action recognition model features and then using a fully connected layer the feature of 1D block and the action recognition model are fused resulting in a feature vector with size 1024. Softmax classifier is added on top of both branches with number of action classes.

3.2.2 Experiments and Discussion

In this part, two methods have been evaluated on Kinetics-mini with (200 classes). A cleaning process and preprocessing procedures have been conducted on Kinetics-Mini as explained in more details in the following section. The experiment of this work is done on

MXNET [30], on RTX 2018 Ti × 2. Each model instance in this section is trained for 100 epochs with total mini-batch size of 16. The SGD optimizer is used with the learning rate of 0.01, reduced in the 40th and the 80th epochs by multiplying it with 0.1. The input size of the model is 224×224, and depth of 16 (16 frames to represents temporal dimension of the video).

3.2.3 Datasets

We have used Kinetics-mini for conducting our experiments due to two main reasons: firstly, Kinetics-mini contains 200 classes with many videos for each action, about 400 videos per class. Secondly, Kinetics-mini is easier to be used for training and evaluating models compared to using the full kinetics dataset which is time consuming. Figure 11 shows statistic of the Kinetics-mini dataset.

At the time of proposing Kinetics-mini proposed in [29], each action has 400 video in the training set, while each action in validation set contains about 25 videos. However, at the time of using this dataset in our experiments many videos are not available in YouTube.

Table 11: reports the statistics of the kinetics-mini dataset utilized in this work.

Subset	Proposed	Extracted	Size (raw videos)	# of Decoded Frames	Size of Decoded frames
Train	80,000	65,961	32.8 GB	16,518,138	660.4 GB
Val	5,000	4,989	3.3 GB	1,549,222	63.44 GB

3.2.4 Results of Variational Feature Learning for Action Recognition

To evaluate the performance of the proposed VFL for action recognition we trained four instances of ResNet-18 model on the preprocessed Kinetics-mini-200. One instance is trained without the proposed VFL, while the three other instances were trained after plugging the proposed VFL module on top of ResNet-18. Table 12 shows the reported results of these instances respectively.

Table 12: The reported performance of ResNet18 on Kinetics-mini 200 classes.

No.	Model	Total batch size	Dataset	Top1	Top2
1	Resnet18_v1b_kinetics200	16	Kinetics-mini 200	66.96	86.45

2	Resnet18_v1b_kinetics200 vfl FC	16	Kinetics-mini 200	65.02	85.69
3	Resnet18_v1b_kinetics200 vfl split	16	Kinetics-mini 200	66.49	86.89
4	Resnet18_v1b_kinetics200 Softmax + KL loss	16	Kinetics-mini 200	65.37	86.22

In Table 12, we trained 4 instances of ResNet18-v1b on Kinetics-mini-200, where the first instance is the defualt model trained for action recognition with default settings. The second instance is trained by applying VFL on top of the feature map of resnet (prior to final classification layer) where, we did not add two fully connected layers, instead we just split the input features of 2048 size into two main parts one to represents the standard deviation and the seocnd represents the means of the gaussian distribution.

The third instance is the action recgnition model with the proposed VFL module by adding two fully connected layers, one generates the means of the CNN features while the second generates the variations of the gaussian distribution. The last training instance is the default action recognition model but with adding the KL loss on top of the model (in parallel with the cross-entropy loss). Based on the reported results in Table 12, the performance of VFL module does not show a significant boost in the performance.

3.2.5 Results of Feature Fusion

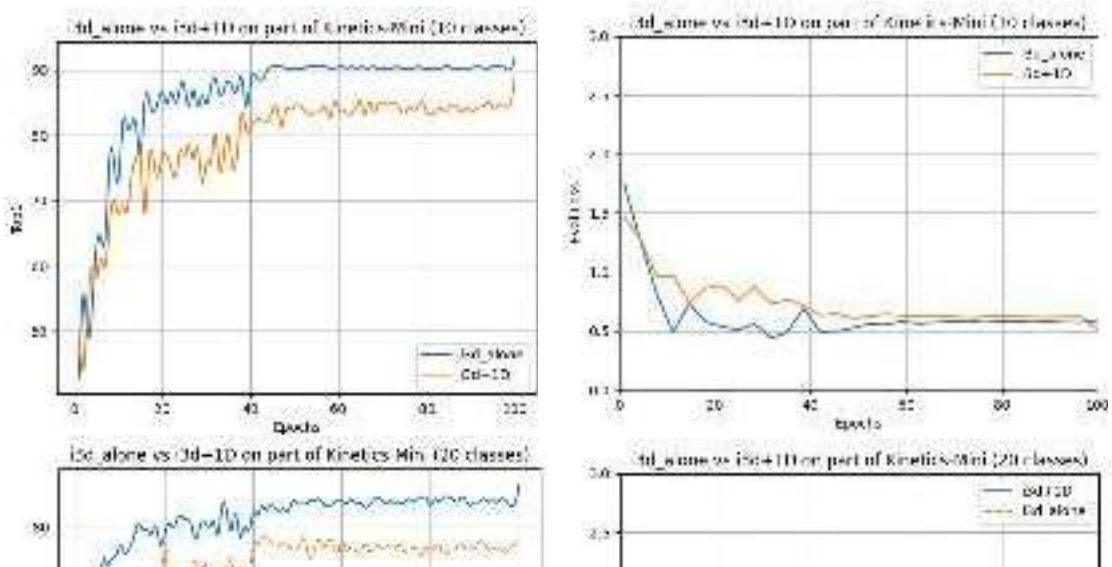
In this part, we first extracted the classification features of each frame by ResNeSt-269 that trained on ImageNet. We extracted the CNN featrues prior to the classification layer. The extracted features has 2048 vector size. These features will be fed to the proposed 1D CNN block which in turn is going to be trained syncroniously with the action recognition model I3D and the baseline ResNet-101. Overall, the input of the action recognition model is the raw frames, while the input to the 1D CNN block is the Classification features of the corrsponding frames. Due to the required large space and the time of extracting the CNN classification features of all Kinetics-mini dataset we extracted the CNN features of 10% of the total frames of Kinetics-min.

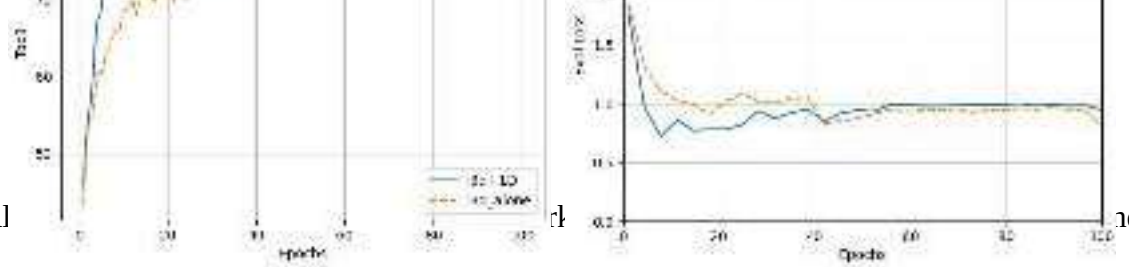
Table 13: Performance of (I3D-ResNet-101 + ResNeSt-269) on a subset of Kinetics-min.

Model I3D-ResNet101	# classes	Dataset kinetics	length	Top1	Top5
I3d-only	10	Part of Mini	16	85.26	98.66
I3d+IDcnn	10	Part of Mini	16	90.62	98.66
I3d-only	20	Part of Mini	16	78.79	96.42
I3d+IDcnn	20	Part of Mini	16	85.72	97.30
I3d-only	200	Part of Mini	16	66.36	86.25
I3d+IDcnn	200	Part of Mini	16	65.80	87.23

Table 13 reports the obtained results of the proposed model I3D+1Dcnn against the default action recognition model I3D. Both instances use the baseline ResNet-101. The results show a remarkable boost in the performance in terms of Top1 and Top5. This can be noted in Figure 9, which shows the validation Top1 accuracy (left column) and the validation loss (right column). First row shows the performance of the training on randomly selected 10 classes of Kinetics-mini-200, while second row shows the performance of the same models but on 20 randomly selected classes.

As it is notable in Table 13 and Figure 9, I3D-ResNet101 combined with the proposed 1D block shows significant progress in terms of Top1 and Top5 accuracies on 10 classes. Same model shows a superior performance on 20 classes.





Al[illegible]rk [illegible]nd 20 classes. Surprisingly, same model on 200 classes does not provides the same superior performance. However, it still shows competitive performance as reported in the lower part of Table 13.

4. Proposed Methods and Experiments on HVU Dataset

In this work, we also investigated the performance of some state-of-the-art deep learning architectures on the HVU mini dataset version. To simplify the problem, we decided to start with the most important visual component, which is the actions to work on at this phase. Furthermore, we selected only 22 common classes out of 739 classes in that category as a pilot study.

In this work we utilized the well-known ResNet-50 deep model as a backbone in two different architectures, Inflated 3D ResNet-50 (I3d-ResNet-50) and Slow-Fast-ResNet-50. The idea behind inflated 3d architectures, in general, is to expand state-of-the-art 2D Conv-Nets, which were optimized well on ImageNet to 3D Conv-Nets with their optimal parameters such that they can perform well on video action recognition. To achieve that as in [31], each 2D filter kernel of size *kxk* in the original ResNet-50 [32] is inflated to a cubic kernel of size *kxkxt*. The original weights of the *kxk* are replicated along the depth of the new kernel and then normalized by the depth value *t*.

Slow-Fast framework as illustrated in Figure 10, on the other hand, involve two Conv-Net streams [33]. A slow stream operating on low frame rate to capture spatial semantics and another fast one operating on high frame rate to capture the fine temporal resolution of the motion. The backbone of this framework can be any Conv-Net architecture. In both architectures, we modified the activation function of the output and the loss function in the appropriate way to serve our goal.

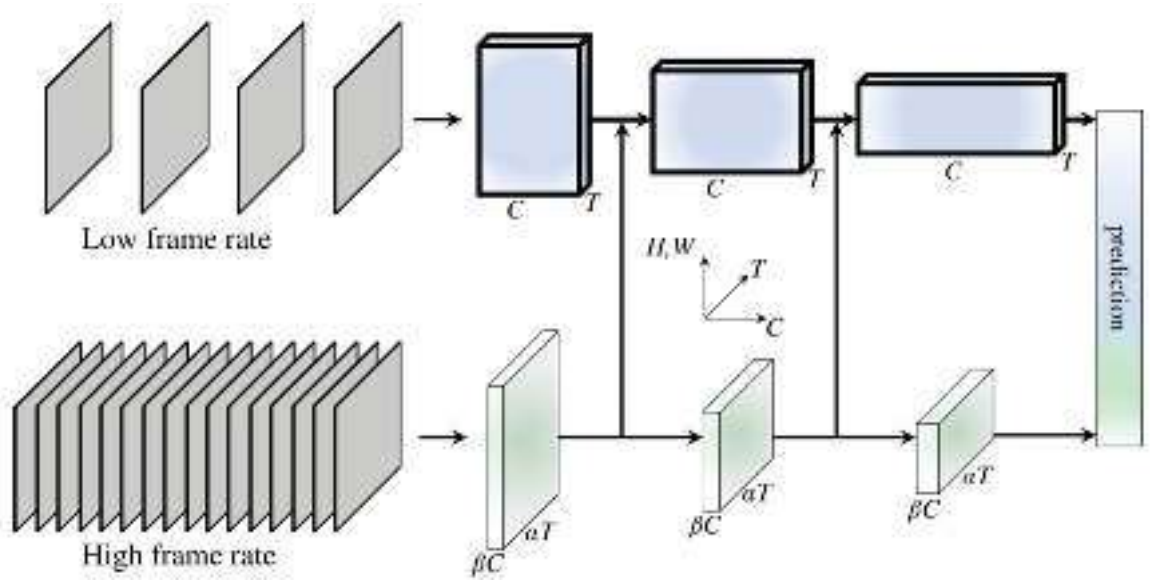


Figure 10. A SlowFast network has a low frame rate, low temporal resolution Slow pathway and a high frame rate, α× higher temporal resolution Fast pathway.

4.1 Dataset Exploration and Pre-Processing

The original HVU mini dataset is supposed to have 105648 samples for training and 8070 samples for validation. Each clip should have a duration of 10 seconds and be annotated with multiple tags of different categories such as objects, scenes, actions, events, and attributes. All this information is presented in the description csv files of the dataset. Unfortunately, by investigating this dataset, we faced different problems as there are many missed and damaged videos due to the downloading process or the unavailability of some videos on YouTube. Also, there are many videos of short duration (less than 10 seconds). All these problems are still not reported in the original csv files of the dataset, which makes it difficult to deal with the dataset before avoiding these issues. For that reason, we decided to perform a preprocessing step to explore, clean, and prepare the dataset. As a result, we obtained the following statistics:

HVU Mini Train:

Total No. of samples:	105648
Very short videos (less than 3 sec):	3313
Short videos (less than 10 secs):	25562

Missed videos:	7252
Damaged videos:	4696
Good videos with at least 3 sec:	90387
Good videos with exact 10 sec:	68138
Videos with some of 22 action tags:	10273

HVU Mini Validation:

Total No. of samples:	8070
Very short videos (less than 3 sec):	47
Short videos (less than 10 secs):	1005
Missed videos:	543
Damaged videos:	1
Good videos with at least 3 sec:	7479
Good videos with exact 10 sec:	6521
Videos with some of 22 action tags:	874

As we are currently interested in the action recognition part, we performed another step to prepare the dataset for that purpose. In this regard and as a beginning, 22 actions were selected out of the total 739 action labels reported in [18]. The selected labels are listed in the first column in Table 14.

After removing the records of damaged and missed videos from the list and ignoring the very short videos, we found that out of 90387 good videos of at least three seconds duration, in the training set, there are only 10273 videos that contain at least one of the 22 selected labels. Similarly, in the validation set, out of 7479 good videos, we ended up with 874 videos with at least one label from the 22 actions' list.

Furthermore, another step was performed on the final dataset of 22 actions to clarify the distribution of videos among the different lengths of labeling list. In other words, to know the

number of samples annotated by different numbers of tags, which is summarized in Figure 11 and Figure 12 for training and validation sets, respectively.

The occurrence frequencies of each action in the training and validation sets are also illustrated in Figure 13 and Figure 14.

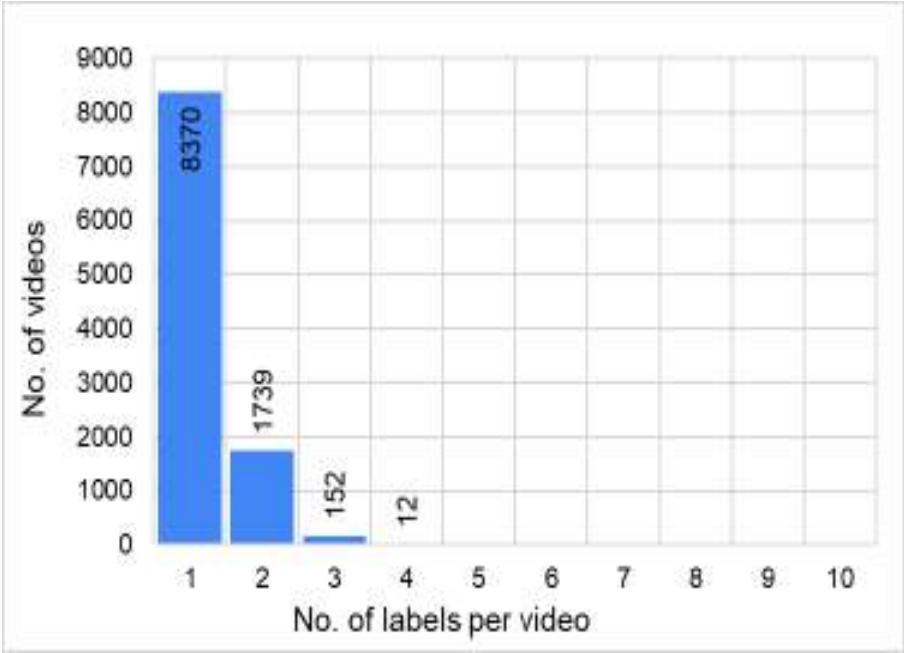


Figure 11. Distribution of training videos among labeling categories of different lengths.

Table 14: Number of samples in each class used for training and validation.

Class	**Train**	**Test**
applauding	175	25
basketball_moves	169	9
bowling	158	40
boxing	541	8
gymnastics	600	36

jumping_jacks	256	19
kicking_soccer_ball	186	21
news_anchoring	257	34
playing_basketball	338	34
playing_volleyball	95	23
presenting_weather_forecast	414	21
public_speaking	789	66
running	1143	82
shooting_goal_soccer	218	20
snorkeling	73	0
speaker	165	33
sports_training	1012	168
strength_training	2303	150
swimming	788	30
water_skiing	144	4
windsurfing	184	29
wrestling	265	22
Total	10273	874

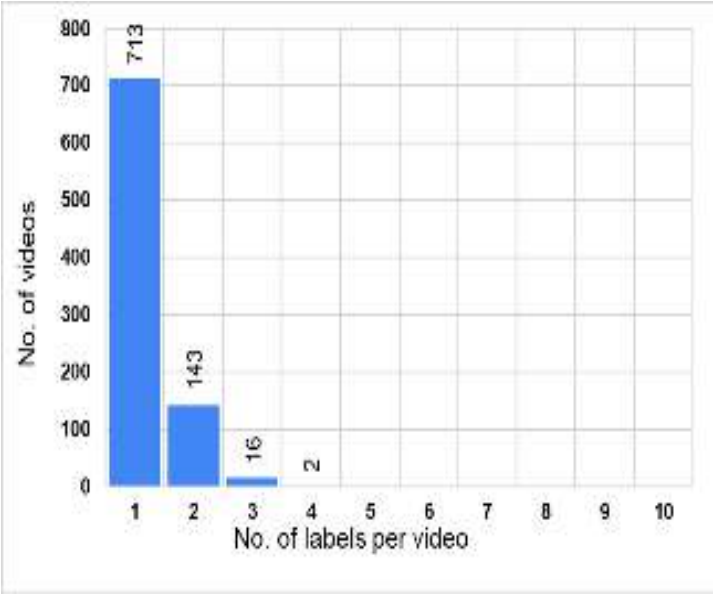


Figure 12. Distribution of validation videos among labeling categories of different lengths.

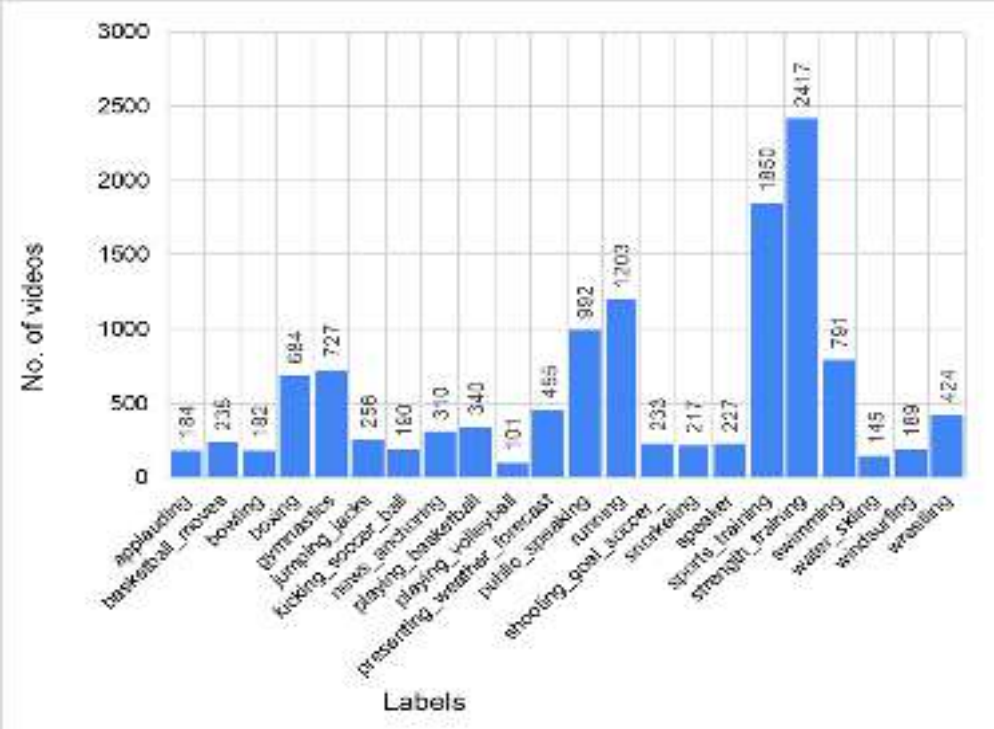


Figure 13. The occurrence frequencies of 22 tags in the training set.

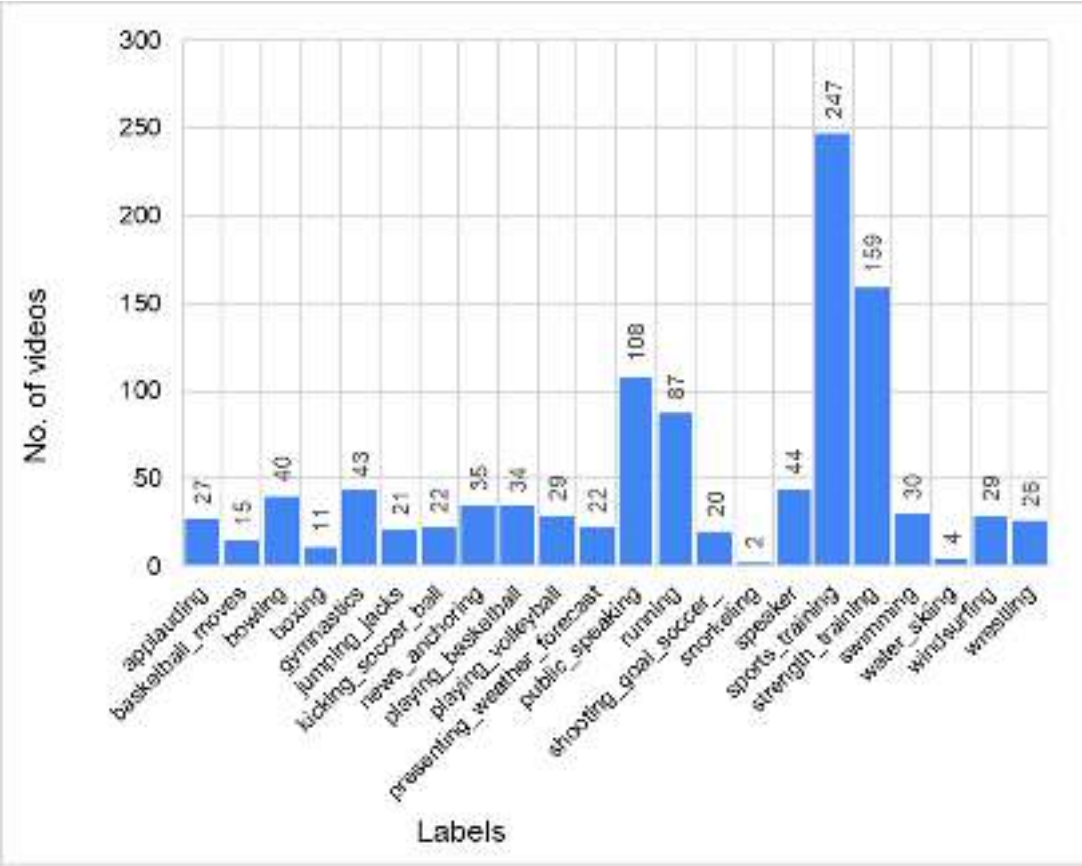


Figure 14. The occurrence frequencies of 22 tags in the validation set.

4.2 Experimental Results and Discussion

4.2.1 Multiclass Classification

To perform a multiclass action recognition, each sample in the dataset should have one and only one label. To achieve that, we considered the first action tag in the tags' list of each sample as the single label for that sample. As a result of this step, we ended up with the distribution of the training and validation samples summarized in Table 14.

A slow-fast architecture, with the well-known ResNet-50 as a backbone, was utilized in two experiments. In the first experiment, the architecture was trained from scratch on the training dataset and evaluated on the validation dataset. After 100 epochs, architecture achieved recognition accuracy of 49.2%.

In the second experiment, we utilized transfer learning by fine-tuning a pretrained version of the same architecture, which was already optimized on the Kinetics-400 dataset. After 100 epochs of fine-tuning, the architecture achieved a recognition accuracy of 75.4%.

This low performance of the architecture achieved even with utilizing a pretrained version of the architecture might be attributed to multiple issues related to the dataset and the way we used to select a single label for each sample.

We decided to repeat the work by picking only the dataset samples that originally have a single label each (the samples belong to only one label of the 22). That means the number of samples for training and testing in this case are respectively, 8370 and 713 as illustrated in the first bar in Figure 11 and Figure 12, respectively. Moreover, by performing a deeper statistical investigation on the resulted dataset, we noticed that the 22 classes were not represented in a balanced way. While some of the classes have enough samples for training, others have only few samples. To solve this issue, we had to remove three classes that do not enough samples for training. This step led to a dataset of 19 classes with the distribution illustrated in Table 15.

Table 15: Number of samples in each class in the original dataset of samples with a single label.

Class	**Train**	**Test**
applauding	161	23
basketball_moves	131	7
bowling	150	40
boxing	504	7
gymnastics	491	26
jumping_jacks	134	11
kicking_soccer_ball	167	19
news_anchoring	222	24
playing_basketball	250	26
presenting_weather_forecast	414	21
public_speaking	641	57

running	913	67
shooting_goal_soccer	188	14
sports_training	936	163
strength_training	1724	115
swimming	600	27
water_skiing	142	4
windsurfing	183	29
wrestling	257	22
Total	8208	702

After training the architecture for 50 epochs, we noticed that the recognition accuracy was enhanced to around 78%, but still there is high confusion between some pairs of classes. As it is clear in the confusion matrix in Figure 15 the highest level of confusion was between “sports_training” and “running”. By exploring many samples of these two classes, we noticed the two labels were used interchangeably to tag the same action in different samples, which led to this highly confused test results.

Based on the previous results, we decided to remove one of the confused classes, which is “sports_training”. We repeated the experiment for 18 classes with 7272 samples in the training set and 539 samples in the testing set. An initial learning rate of 1e-3 which decayed by 10% after each 10 epochs to get smoother parameters’ tuning.

The accuracy and loss metrics during the training progress were depicted in Figure 16. The evaluation on the testing dataset after training was detailed as a confusion matrix in Figure 17. The architecture achieved a recognition accuracy of 83.3%. From the learning curves, we can notice that there is still a continuous slow enhancement, if we continue for extra learning epochs.

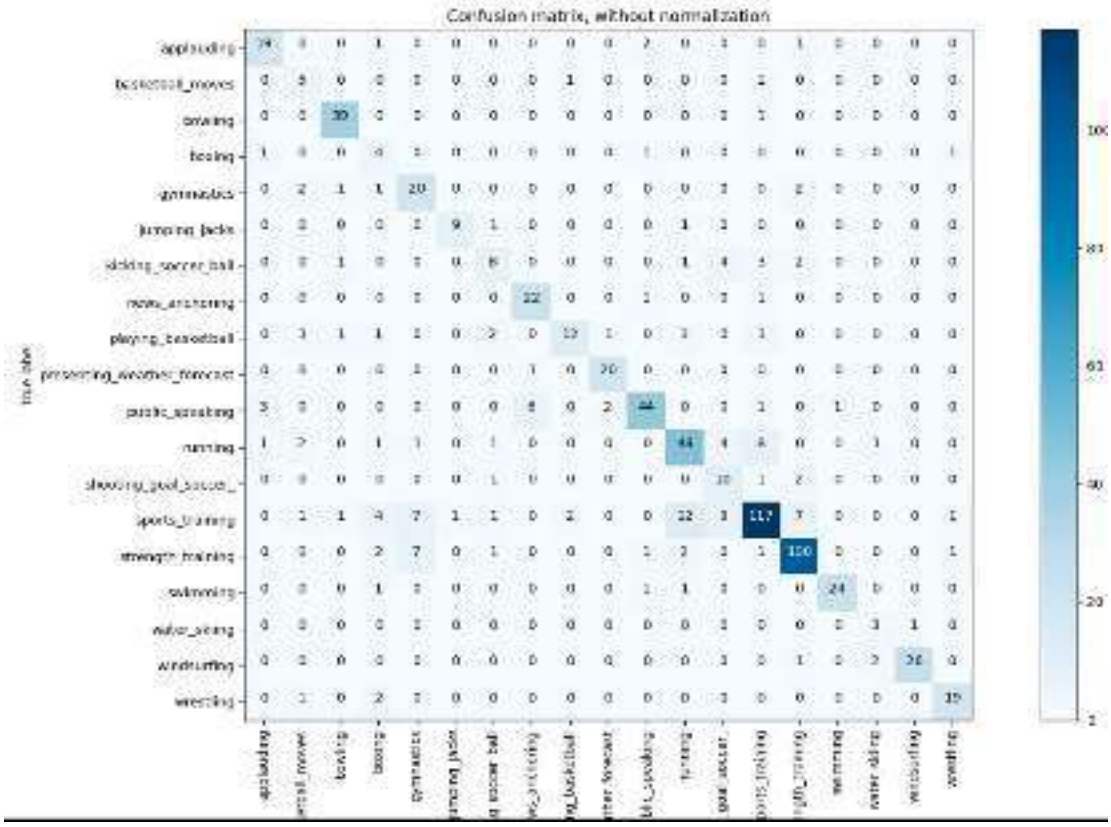


Figure 15. The confusion matrix of testing slow-fast architecture for multiclass classification on 19 classes.

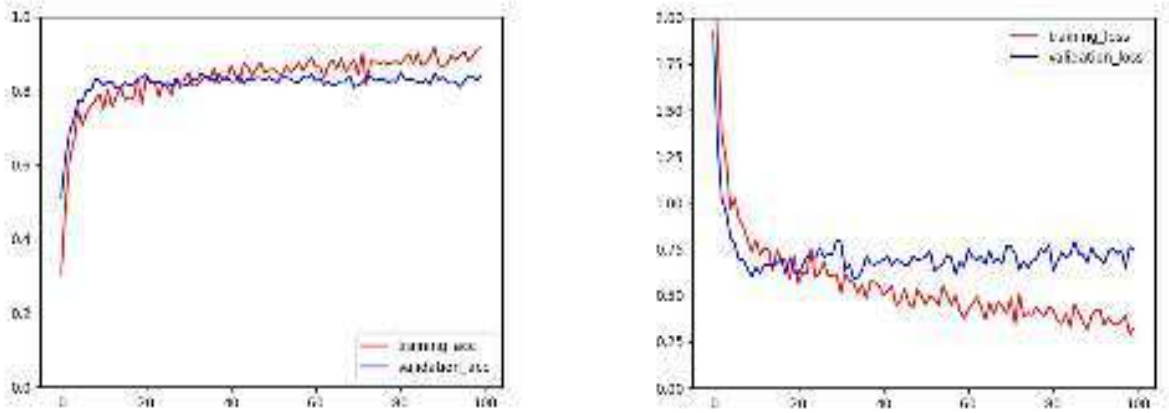


Figure 16. Training and validation accuracy and loss of the SlowFast on 18 HVU classes.

In another experiment, we repeated the same setup of the last experiment, on the I3d ResNet-50 architecture. This architecture achieved a slightly better recognition rate. As illustrated in the confusion matrix in Figure 18, an accuracy of 83.85 was achieved. The performance metrics of this architecture during training epochs are also illustrated in Figure 19.

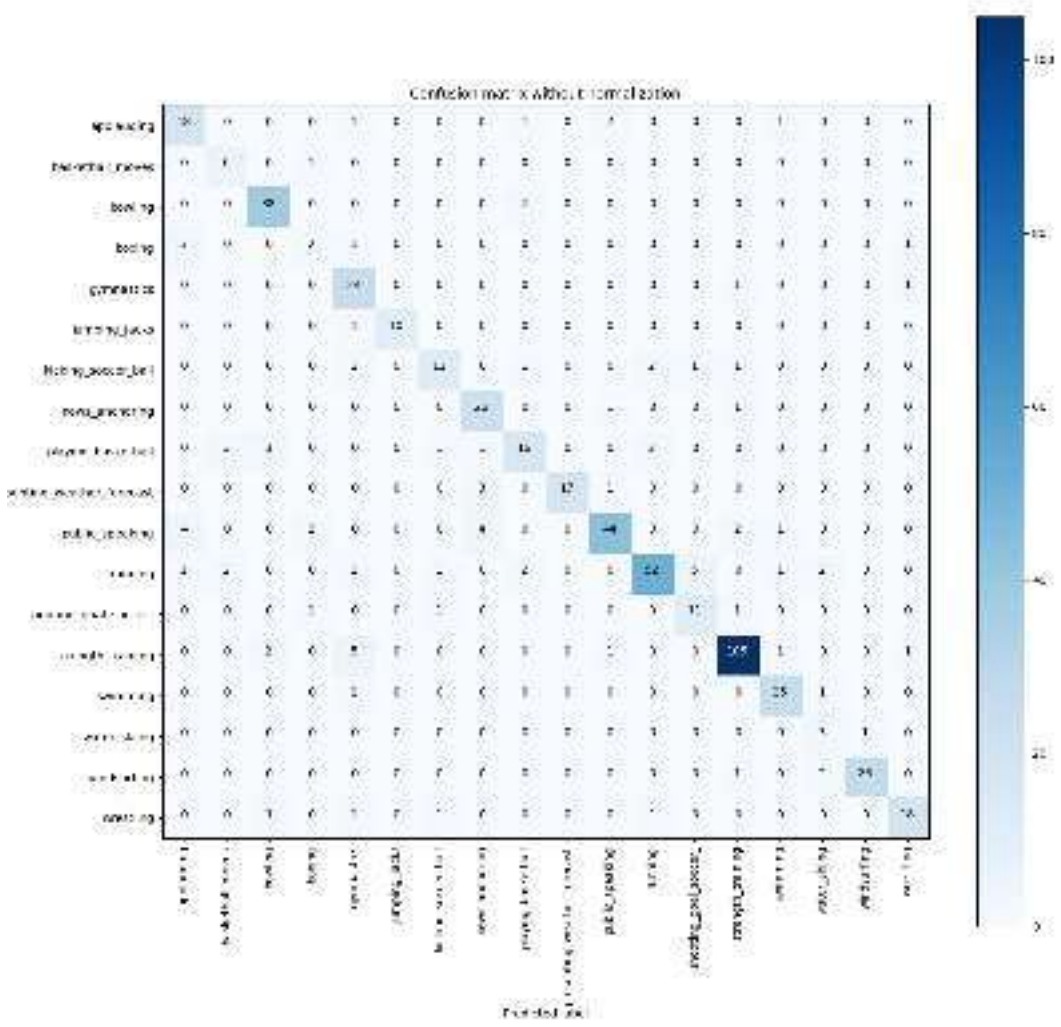


Figure 17. The confusion matrix of testing slow-fast architecture for multiclass classification on 18 classes.

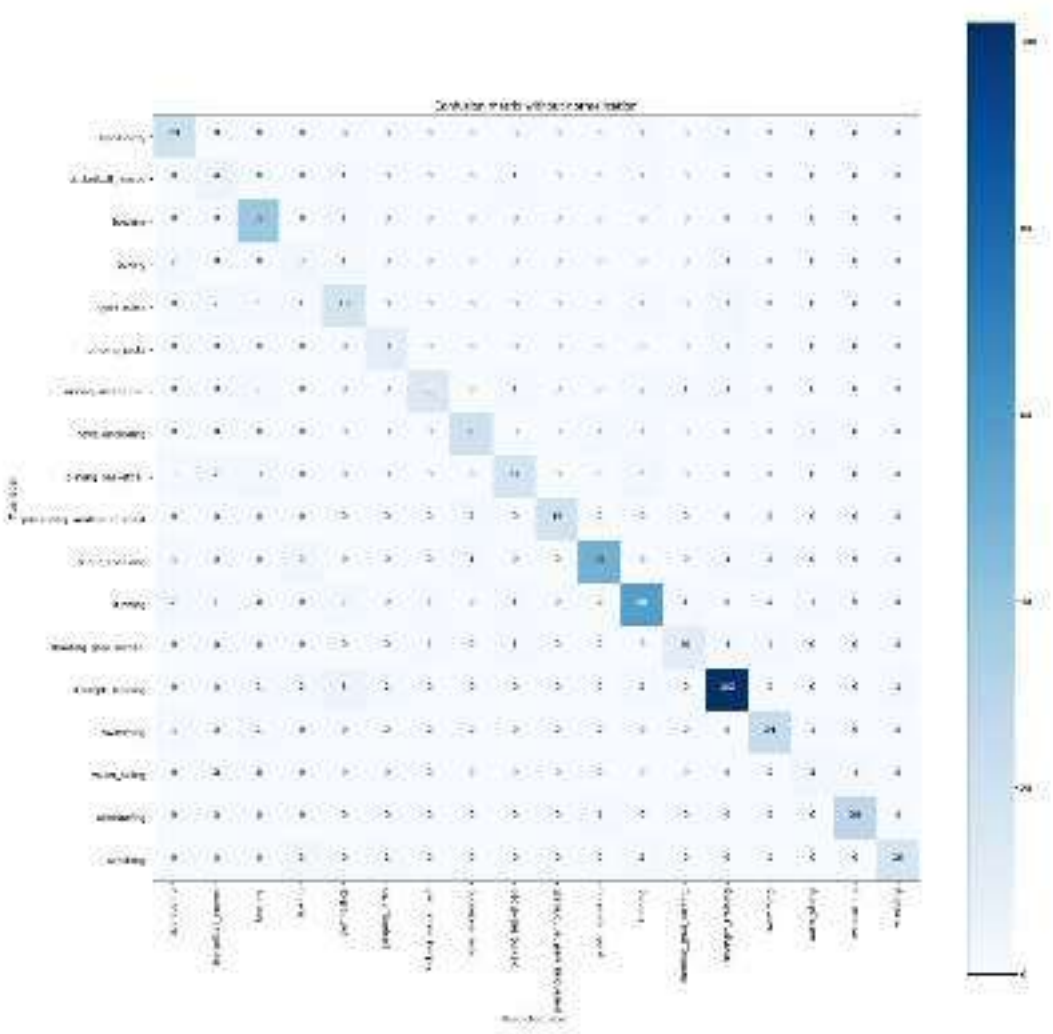


Figure 18. The confusion matrix of testing I3d-ResNet-50 architecture for multiclass classification on 18 classes.

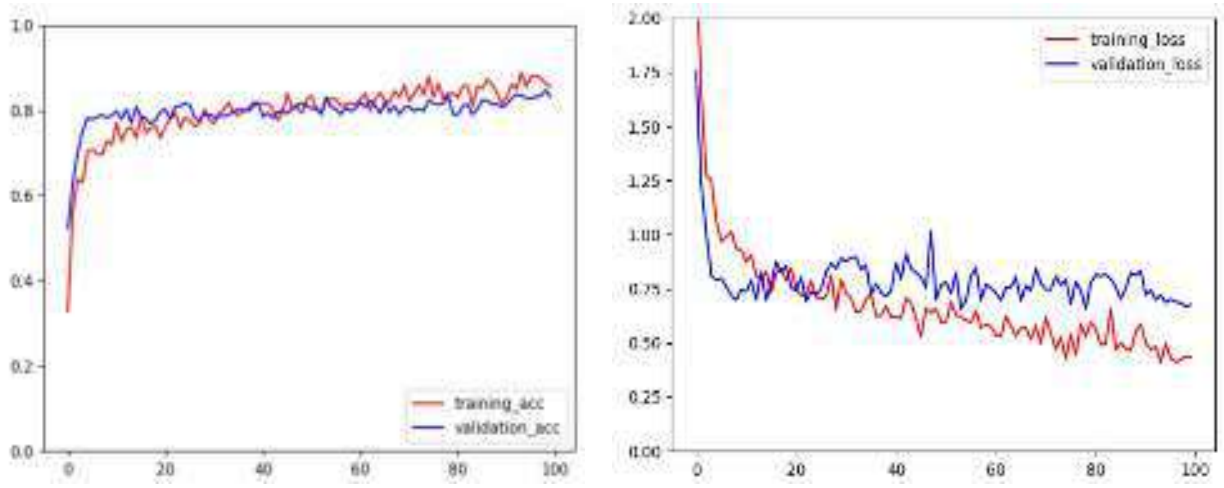


Figure 19. Training and validation accuracy and loss and loss of the I3d ResNet-50 on 18 HVU classes.

4.3.3 Multilabel Classification

For multilabel classification, we modified the slow-fast architecture of ResNet50 for this purpose. The output layer of the architecture was replaced by a dense layer of 22 neurons, which is the number of classes in our dataset. Furthermore, sigmoid function was used to activate the neurons of this layer instead of softmax and binary cross entropy loss instead of categorical entropy loss. The originally prepared dataset of 22 actions was used in this experiment. Hence there are 10273 samples for training and 874 samples for testing, which are distributed as Figure 11 and Figure 12.

For evaluation purpose, we used three metrics defined as in the following equations:

$$Jaccardindex(prediction, actual) = \frac{|prediction \cap actual|}{|prediction \cup actual|}$$

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F1\,score = \frac{2 \times Precision \times Recall}{(Precision + Recall)}$$

The architecture was trained from scratch over 100 epochs and then evaluated on the testing dataset with different threshold values. The performance of the architecture is summarized in Table 16.

It is clear from these results that the performance of the architecture is not encouraging even with very low threshold values. This bad performance is attributed to the poor representation of different labels' permutations. Most of the labels' permutations are not presented in the dataset as it is clear in the distribution in Figure 13 and Figure 14.

Table 16: The performance of slow-fast resNet50-based architecture on 22 HVU classes in multilabel scenario.

Threshold	Jaccard index	Precision	Recall	F1 score
0.9	**0**	-	**0**	-
0.8	**0**	-	**0**	-
0.7	**0**	-	**0**	-
0.6	**0**	-	**0**	-
0.5	**0**	-	**0**	-
0.4	**0**	-	**0**	-
0.3	**0**	-	**0**	-
0.2	**0.1555**	**0.18**	**0.16**	**0.17**
0.1	**0.1819**	**0.19**	**0.47**	**0.27**

5. Conclusion

This report describes the designed VFL and 1D block. It also explains a set of experiments that conducted on Kinetics-Mini-200. Generally, the VFL modules does not boost the action recognition performance, while the 1D block shows a remarkable improvement on the Top1,

Top5 performance when we train the model on few classes 10 classes and 20 classes. However, 1D block does not show same performance improvement when it trained with 200 classes.

Moreover, two state-of-the-art architectures, SlowFast and inflated 3D Conv-Net are utilized for action recognition on HVU dataset. the backbone of the both architectures is Resnet-50. With careful data preprocessing, both architectures obtain comparable results in multiclass classification. The low performance of SlowFast in multilabel classification might be attributed to the lack of balance in the dataset action tags.

References

[1] Y. Tian, R. Sukthankar and M. Shah, "Spatiotemporal deformable part models for action detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, June 2013.

[2] O. Dan, J. Verbeek and C. Schmid, "Efficient action localization with approximately normalized fisher vectors," in Proceedings of the IEEE conference on computer vision and pattern recognition, Columbus, Jun 2014.

[3] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," arXiv:1412.4729v3, 2015.

[4] Y. Pan, T. Mei, T. Yao, H. Li and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), LAS VEGAS, June 26th - July 1st, 2016.

[5] V. Escorcia et al., "DAPs: Deep Action Proposals for Action Understanding," in European Conference on Computer Vision (ECCV), Munich, Sep. 2016.

[6] R. Krishna, K. Hata, F. Ren, L. Fei-Fie and J. Carlos Niebles, "Dense-Captioning Events in Videos," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Oct. 2017.

[7] H. Kuehne, H. Jhuang, R. Stief and T. Serre, "HMDB51: A Large Video Database for Human Motion Recognition," in High Performance Computing in Science and Engineering, Stuttgart, Oct. 2012.

[8] K. Soomro, A. R. Zamir and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," arXiv:1212.0402v1 [cs.CV] , 2012.

[9] Google, "YouTube-8M," Google, [Online]. Available: https://research.google.com/youtube8m/. [Accessed 24 July 2020].

[10] S. Abu-El-Haija and et. al., "YouTube-8M: A Large-Scale Video Classification Benchmark," arXiv:1609.08675v1 [cs.CV], 2016.

[11] NIST, "TREC Video Retrieval Evaluation: TRECVID," [Online]. Available: https://trecvid.nist.gov/. [Accessed 28 8 2020].

[12] G. A. M. M. J. F. G. S. B. S. W. K. A. F. S. a. G. Q. P. Over, "P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. TRECVID 2012-an overview of the goals, tasks, data, evaluation mechanisms and metrics," 2013.

[13] T. I. W. K. a. A. F. S. O. Paul, "TRECVID 2005-an overview," 2005.

[14] C. S. F. K. L. a. H. W. ". L. a. A. H. Y. A., "LSCOM lexicon definitions and annotations (version 1.0)," 2006.

[15] A. H. a. S.-F. C. K. Lyndon, "Revision of LSCOM event/activity annotations," in in DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, 2006.

[16] R. Krishna, K. Hata, F. Ren, L. Fei-Fie and J. Carlos Niebles, "Dense-Captioning Events in Videos," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Oct. 2017.

[17] facebookresearch, "ActivityNet Entities Dataset and Challenge," 2020. [Online]. Available: https://github.com/facebookresearch/ActivityNet-Entities. [Accessed 27 8 2020].

[18] A. Diba, M. Fayyaz, V. Sharma, M. Paluri, J. Gall, R. Stiefelhagen and L. V. Gool, "Large Scale Holistic Video Understanding," arXiv:1904.11451v2 [cs.CV], 2019.

[19] DeepMind, [Online]. Available: https://deepmind.com/research/open-source/kinetics. [Accessed 03 Feb 2021].

[20] Will Kay, et al., “The Kinetics Human Action Video Dataset,” arXiv:1705.06950 [cs.CV], May 2017.

[21] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” 2014.

[22] S. A. S. Alfasly, Y. Hu, T. Liang, X. Jin, Q. Zhao, and B. Liu, “Variational Representation Learning for Vehicle Re-Identification,” 2019, doi: 10.1109/ICIP.2019.8803366.

[23] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, “MMTM: Multimodal transfer module for CNN fusion,” 2020, doi: 10.1109/CVPR42600.2020.01330.

[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[25] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” 2017, doi: 10.1109/CVPR.2017.106.

[26] Z. X. Li and F. Q. Zhou, “FSSD: Feature fusion single shot multibox detector,” *arXiv*. 2017.

[27] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional Two-Stream Network Fusion for Video Action Recognition,” 2016, doi: 10.1109/CVPR.2016.213.

[28] O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[29] K. M. Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, “Rethinking Spatiotemporal Feature Learning For Video Understanding,” *Eur. Conf. Comput. Vis.*, 2018.

[30] T. Chen *et al.*, “MXNet : A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems arXiv : 1512 . 01274v1 [cs . DC] 3 Dec 2015,” *Emerald Gr. Publ. Ltd.*, 2015.

[31] X. Liao , . L. He, Z. Yang and C. Zhang, "Video-Based Person Re-identification via 3D Convolutional Networks and Non-local Attention," in *ACCV*, Perth, 2018.

[32] H. Kaiming , Z. Xiangyu, R. Shaoqing and S. Jian , "Deep Residual Learning for Image Recognition," in *CVPR*, LAS VEGAS, 2016.

[33] F. Christoph , F. Haoqi , M. Jitendra and H. Kaiming, "SlowFast Networks for Video Recognition," in *ICCV*, Seoul, 2019.

[34] TRECVID, [Online]. Available: https://trecvid.nist.gov/ [Accessed 27 Feb 2021].

[35] A. Yanagawa, et al., " Columbia University’s Baseline Detectors for 374 LSCOM Semantic Visual Concepts," Available: https://www.ee.columbia.edu/ln/dvmm/publications/07/Yanagawa_Columbia374.pdf [Accessed 27 Feb 2021]

[36] TRECVID 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking, [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv16.papers/tv16overview.pdf

[Accessed 27 Feb 2021].

[37] M. Naphade et al., "Large-scale concept ontology for multimedia," in IEEE MultiMedia, vol. 13, no.

3, pp. 86-91, July-Sept. 2006, doi: 10.1109/MMUL.2006.63.

[38] DeepAi, “Holistic Large Scale Video Understanding,” [Online]. Available:

https://deepai.org/publication/holistic-large-scale-video-understanding [Accessed 27 Feb 2021].

Received May 21, 2021, accepted June 8, 2021, date of publication June 17, 2021, date of current version June 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3089924

An Incremental Approach to Corpus Design and Construction: Application to a Large Contemporary Saudi Corpus

HEBAH ELGIBREEN[1,2], MOHAMMED FAISAL[1,3], MANSOUR AL SULAIMAN[1,4], (Member, IEEE), SHERIF ABDOU[5], MOHAMED AMINE MEKHTICHE[1,4], ABDULLAH M. MOUSSA[5], (Member, IEEE), YOUSEF A. ALOHALI[1,6], WADOOD ABDUL[1,4], GHULAM MUHAMMAD[1,4], (Senior Member, IEEE), MOHSEN RASHWAN[7], AND MOHAMMED ALGABRI[1,6]

[1]Center of Smart Robotics Research, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia
[2]Department of Information Technology, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia
[3]College of Applied Computer Sciences, King Saud University, Riyadh 11451, Saudi Arabia
[4]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia
[5]Department of Information Technology, Faculty of Computers and Artificial Intelligence, Cairo University, Giza 12613, Egypt
[6]Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia
[7]Department of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, Giza 12613, Egypt

Corresponding author: Hebah Elgibreen (hjibreen@ksu.edu.sa)

This work was supported by the Deputyship for Research and Innovation, Ministry of Education, Saudi Arabia, under Project DRI-KSU-1292.

ABSTRACT Due to the rapid developments in technology and the sudden expansion of social media use, Dialect Arabic has become an important source of data that needs to be addressed when building Arabic corpora. In this paper, thirty-three Arabic corpora are surveyed to show that despite all of the developments in the literature, Saudi dialect (SD) corpora still need further expansion. This paper contributes to the literature on SD corpora by creating the largest Saudi corpus – the King Saud University Saudi Corpus (KSUSC) – with +1B total words, including +119M SD words. The KSUSC not only is the newest and largest SD corpus but is also diverse, covering 26 domains in text collected from five different sources. This paper also contributes to the literature by developing a new incremental preprocessing system that is used to create relevant lexicons that are then used to clean and normalize the collected data. This incremental system is scalable and can be adapted for different resources and dialects. Moreover, the collection process for building the KSUSC is discussed in detail, and the challenges in collecting SD text with respect to each platform are highlighted. By the end of this paper, different design criteria are proposed and used with the KSUSC to conclude that the resulting corpus can be of great benefit to researchers who are interested in integrating the corpus with their own work or using its resulting lexicons with Saudi-based NLP tasks.

INDEX TERMS Saudi dialect, corpus, natural language processing, data preprocessing.

I. INTRODUCTION

Language corpus is a term used to describe a collection of texts written in one or more languages [1]. It is considered one of the most important sources of data in different areas, including information retrieval (IR), natural language processing (NLP), and computational linguistics (CL). This is because it can represent the written language and, hence, be used to process opinions and implement related applications. Compared to English corpora, Arabic corpora are poorly resourced and lack sufficient research and data, which negatively affects Arabic-based NLP practitioners [2].

The Arabic language is the mother language of Arab countries and one of the six official languages of the United Nations (UN) [3]. There are three main versions of the Arabic language: Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialect Arabic (DA). CA is the official language used in the Quran and during the medieval period, MSA

The associate editor coordinating the review of this manuscript and approving it for publication was Long Wang.

is the official language used in the modern period and by news outlets, and DA is the spoken language that is used in daily life and differs from one country/city to another. Moreover, there are many subcategories under these Arabic languages, and thus, Arabic is considered to be highly inflectional and to have a very complex morphology [4]. Therefore, the development of large Arabic corpora has been the focus of many researchers in recent years. However, most of these developments have focused on CA and MSA rather than DA; thus, Arabic NLP solutions perform poorly with DA data [5]. In DA, also known as colloquial language, one lemma can have hundreds of surface forms, and thus, it is a morphologically rich language [6]. DA differs from MSA syntactically, morphologically, and phonologically and does not have standard orthographies [7]. As emphasized in the literature [8], it is nearly impossible to have one NLP solution that can process all variations of Arabic. It is also important to have a corpus that reflects the current use of language [9], and thus, new terms that reflect recent events around the world (such as the COVID-19 pandemic and the Saudi Vision 2030 framework) must be incorporated into corpora.

Although MSA used to be considered the dominant written language, due to the social media platforms emerging around the world, DA has become more frequently written than MSA [10]. Moreover, Saudis are considered some of the most active users of social media. It was reported by Alruily [4] that there are more than 11 million Arabic accounts in Twitter, with 27.4 million tweets being published a day; Saudi Arabia has the most active users, with 30% of these tweets. In [11], 61% of the 175M Arabic tweets considered were found to be from Saudi Arabia.

This paper focuses on Saudi dialect (SD) corpora and highlights the current challenges involved in such corpora. In particular, from the literature on SD corpora, it is found that MSA language dominates most of the large corpora, while SD text has not been introduced in many of them. The size of the current SD corpora is very small, and it ranges only from thousands of words to a couple million words. Although 2∼3M words might be considered a large number, it is not enough for NLP tasks that target the Arabic language [12]. Moreover, the lack of diversity is also a challenge, as SD corpora text is mostly collected from Twitter. Although Twitter can be a rich source of data, it reflect only one social aspect of the Arabic-speaking community, and the diversity of opinions and domains is important. Therefore, this paper contributes to the literature on SD corpora as follows:

First, a comprehensive survey of 33 Arabic corpora is conducted. This survey describes each corpus and its availability. The corpora are categorized by their language, source, purpose, size, text date, accessibility and limitations on reuse. This survey can help readers understand the progress of Arabic corpora and the current limitations of SD corpora.

Second, extensive data collection is conducted with documents containing over 1.2B words. The data are collected from different sources and diverse domains. In particular, pre-exiting corpora are considered in addition to new online sources. Facebook, YouTube, and Twitter as well as other websites are used to collect text discussing recent events. Moreover, MSA language is introduced with SD to enrich the data and cover different semantic-based tasks. The statistics of the collected resources are summarized to show the domain of each source, and challenges faced when collecting SD text from each platform are also highlighted in this paper.

Third, a new incremental preprocessing system is proposed in this paper to create SD lexicons and use them to clean and normalize the text. Due to the diversity of the collection sources, it was almost impossible to use one round of cleaning that would work on all kinds of data; even the preexisting corpora needed a certain level of preprocessing to be compatible with the collected text. One of the main advantages of having an incremental approach is to avoid the need for stemming which is very challenging for dialectal Arabic and, thus, develop a system that is not restrict to certain rules or predefine lexicon. Thus, the proposed system is designed to incrementally create two types of lexicons to identify common ASCII and unwanted symbols. These lexicons will be used to clean MSA and SD text from different resources and platforms and then validate the data to eliminate irrelevant characters or incomplete text. The incremental process introduced in this system allows it to be scaled to other languages in DA.

Fourth, a contemporary linguistic corpus for the Saudi language is designed and created in this paper. This corpus is named the KSUSC corpus, and to the authors' knowledge, it is the newest and largest SD corpora to date. The KSUSC is the largest Saudi corpus and is diverse and up to date, with clearly defined design criteria. Its content is classified based on source, domain, and time. It includes +1B words, +161M sentences, and +14M unique words, covering 26 different domains. It is designed to overcome the current limitations of SD corpora and utilize the preprocessing system proposed in this paper.

This paper is organized as follows. First, a literature review of relevant corpora is presented. Second, the collection process is described, and the details of the collected resources are also summarized. Third, the architecture of the proposed preprocessing system is presented, and examples are illustrated at each phase. Finally, the KSUSC corpus design is presented, and all its statistics are discussed.

II. LITERATURE REVIEW

In the literature on the Arabic corpus, several corpora have been built in the past few years. This section surveys 33 Arabic corpora to summarize and compare their key criteria and motivate the contribution of this paper. One of the oldest Arabic corpora in the literature is the Penn Arabic Treebank (PATB) corpus [13], which was introduced in 2001. It has had three full releases of morphologically and syntactically annotated data: (1) the Arabic Treebank: Part 1, which consists of 166K words of written MSA newswire from the Agence France Presse corpus; (2) the Arabic Treebank: Part 2, which consists of 144K words from Al-Hayat distributed by Ummah

Arabic News Text; and (3) the Arabic Treebank: Part 3, which consists of 350K words of morphologically annotated newswire text from An-Nahar.

Cross Lingual Arabic Blog Alerts (COLABA) [14] is a project that was proposed for creating AD resources and NLP tools. In this project, an Arabic dialect corpus was created and consisted of four dialects: Egyptian, Iraqi, Levantine, and Moroccan. Given that the language used on social media is typically DA, one of the main objectives of COLABA was to illustrate the significant impact of using the corpus for DA processing with NLP applications. Accordingly, information retrieval (IR) was chosen as the main testbed application to process DA.

The Open Source Arabic Corpora (OSAC) [15] is a freely accessible MSA corpus that contains approximately 18M words and 500K keywords after stopword removal. It was collected from websites from 22,429 text documents in 10 categories: economics, history, entertainment, education & family, religious & fatwas, sports, heath, astronomy, law, stories, and cooking recipes. This corpus was used to determine the impact of preprocessing on Arabic text classification.

The OPUS Multilingual Corpus [16] is a growing language resource of parallel corpora and related tools. The goal of OPUS is to provide freely available data sets in various formats together with basic annotation that is useful for applications in computational linguistics, translation studies and cross-linguistic corpus studies. The overall goal of the OPUS project was to make parallel resources freely available. OPUS covers a substantial amount of newspaper texts and some other smaller collections from various online sources. OPUS has been extended to several large collections, such as TED [16] which is a parallel corpus of TED talk subtitles provided by CASMACAT. The files were originally provided by the Web Inventory of Transcribed and Translated Talks, consisting of 15 languages and a total number of 3.81M sentence fragments. Another example is MultiUN [17], which is a collection of translated documents from the United Nations with 6 languages and a total of 81.41M sentence fragments. Another well-known extended dataset is OpenSubtitles [18], which is a collection of translated movie subtitles from http://www.opensubtitles.org. It is a cleaner version of the subtitles using improved sentence alignment and better language checking with 62 languages and a total of 3.35G sentence fragments.

Yet Another Dialectal Arabic Corpus (YADAC) [19] is another multigenre dialectal Arabic corpus collected from Twitter, blogs/forums and online knowledge market services in which both questions and answers are user-generated. For this study, 15M search queries were randomly selected and used to crawl the web over a period of 7 months – May 2011 to November 2011. After applying the threshold model of dialect identification, the total size of YADAC reached 6M wordform tokens and 457K wordform types. Forty-one percent of the text was collected from online knowledge market services, 32% from microblogs, and 27% from blogs and forums.

The Arabic corpus for Egyptian tweets [20] is an Egyptian dialect corpus that consists of several general topics from Twitter. The corpus contains 22,834 tweets collected from May 2011 to December 2011. It is a subset of the YADAC corpus and uses a function-based annotation scheme in which words are labeled based on their grammatical functions rather than their morpho-syntactic structures.

AWATIF [21] is a multigenre corpus with MSA text that is labeled for subjectivity and sentiment analysis (SSA) at the sentence level. This corpus consists of 5,382 sentences and is labeled using both regular and crowdsource methods. It was collected from three different sources: 2855 sentences from Part 1 V 3.0 (ATB1V3) of the PATB corpus, 1019 sentences from 30 Wikipedia Talk Pages (WTP), and 1508 sentences from web forums.

The Multi-Dialectal Corpus of Arabic [11] is another multidialectal Arabic corpus collected based on the geographical information of tweets. In this corpus, 175 million tweets were collected in March 2014, and 62M Arabic tweets were selected. Selected tweets were filtered based on dialectal words to extract 6.5M tweets (i.e., 3.7% of the original tweets), of which 3.99M (61%) were from SA, 880K (13%) were from EG, 707K (11%) were from KW, 302K (5%) were from AE, 65k (2%) were from QA, and the remaining (8%) were from other countries such as Morocco and Sudan.

SANA [22] is a corpus that is a large-scale, multigenre, multidialect, and multilingual lexicon used for subjectivity and sentiment analysis. It includes 224,564 sentences that were collected from online newswires, chat turns, Twitter tweets, and YouTube comments. It includes text from MSA, Egyptian DA and Levantine DA along with English glosses. On the other hand, the Tunisian Dialect Corpus (TunDiaWN) [23] uses a corpus-based approach to create WordNet resources for the Tunisian dialect. It consists of 32,848 words collected from social media (Twitter, Facebook, etc.), written theatrical pieces, dictionaries, transcriptions of spontaneous speech, etc.

The King Abdulaziz City for Science and Technology (KACST) Arabic corpus [2], is one of the largest MSA corpora and was designed to overcome the limitations of existing Arabic corpora. The corpus texts were collected from several sources and contain more than 731 million words from 869,800 texts. The KACST Arabic corpus was designed and constructed to accommodate a large-sized and sufficiently diverse Arabic corpus able to represent the many varieties of Arabic language across three main dimensions: time, region, and genre. Such a corpus can be used for different research interests, ranging from linguistic studies at various levels to the development of NLP applications.

The 1.5 Billion Arabic Corpus [24] is a linguistic corpus for MSA language that includes more than 5M newspaper articles collected from 10 news sources between December 2013 and June 2014. This corpus covers several categories, including politics, literature, arts, technology, sports, economy, culture, and many other subject matters. It consists of more than 1.5 billion words and 3.3 million unique words. The

main purpose of creating this corpus was to provide a free Arabic language tool to researchers.

The King Saud University Corpus of Classical Arabic (KSUCCA) [25] is another CA corpus, and it consists of approximately 50M words. The corpus includes six categories: religion, linguistics, literature, science, sociology, and biography. The main purpose of KSUCCA is to help in studying the distributional lexical semantics of words in the holy Quran. It is used in ongoing research that attempts to study the meanings of words used in the holy Quran through the analysis of their distributional semantics in contemporaneous texts.

The Gumar corpus [26] is a large-scale Gulf Arabic (GA) dialect corpus that includes a number of subdialects from six countries of the Gulf Cooperation Council: Saudi Arabia (SA), Bahrain (BH), Kuwait (KW), Oman (OM), UAE, and Qatar (QA). It consists of 112 million words from 1,200 forum novels. It is 60.52% SA, 13.35% EA, 5.91% KW, 1.13% OM, 0.65% QA, and 0.49% BH. Moreover, approximately 10% of the text is identified as GA (other), which happens when, for example, a novel contains a combination of several GA dialects due to the existence of different characters in the novel or to the novel being authored by multiple writers with different dialects.

Another version of Gumar corpus was introduced in 2018 and called Gumar Emirati [27]. This version is the first large-scale morphologically manually annotated corpus for the UEA language. This corpus includes approximately 200,000 words selected from eight novels in Emirati Arabic. The selected texts are annotated for tokenization, part-of-speech, lemmatization, English glosses and dialect identification. The corpus includes the native spoken variety in the Gulf Cooperation Council; however, it still lags behind the resource and tool creation of other Arabic dialects, given the considerable amount of dialectal content online. Another morphologically annotated corpus that is considered the first morphologically annotated corpus for the Palestinian dialect is the Curras corpus (Jarrar *et al.* 2017). The Palestinian language is a very commonly spoken version of DA. The Curras corpus consists of approximately 43,000 words and was collected from a variety of resources (Facebook, Twitter, blogs, forums, Palestinian stories, Palestinian terms, TV shows).

A Saudi Twitter Corpus [28] was introduced in 2016 and consists of 4700 SD tweets that are used for sentiment analysis. This corpus includes data from Twitter and covers several domains, such as sports, economy, and politics. The intention behind building this corpus was to create the first reliably annotated Twitter data for SD language. Additionally, another corpus that was introduced for sentiment analysis and machine language applications is the Book Reviews in Arabic Dataset (BRAD) [29]. It consists of approximately 2,781,805 sentences in MSA and dialects and was extracted from 510,598 Arabic reviews collected from 4993 books and 76530 reviewers/users. The balanced clean subset contains 156,506 reviews, and each review is annotated with a scale from 1 to 5.

One of the largest Arabic booking review corpora is the Hotel Arabic-Reviews Dataset (HARD) [30]. It was designed for subjective sentiment analysis and machine language applications. It consists of 490,587 MSA and dialectal hotel reviews collected from the Booking.com website, which specializes in online accommodations booking. The collected reviews are structured as follows: rating, title of the review, positive aspect(s) of the accommodation, negative aspect(s) of the accommodation, reviewer's username, and country of residence.

Arabic Sentiment Analysis of Saudi Tweets (AraSenTi-Tweet) [6] is an MSA and SA corpus collected from Twitter. It contains approximately 2.2M tweets, with 17,573 of them being Saudi tweets. The corpus is manually annotated for sentiment and labeled with four labels for sentiment: positive, negative, neutral and mixed. Another corpus created based on Twitter is the Multi-Dialect Arabic Sentiment Twitter Dataset (MD-ArSenTD) [31] which is a multidialect Arabic corpus collected from tweets from 12 Arab countries (KW, SA, QA, UAE, Jordan, Lebanon, Palestine, Syria, Algeria, Morocco, Tunisia, Egypt) and annotated for sentiment and dialect. The Twitter4J API [32] was used to collect 470K tweets posted from 3/1/2017 to 4/30/2017. Then, 14,400 total tweets were selected, and 1,200 tweets from each country were selected and annotated.

Tashkeela [33] is an MSA and CA corpus consisting of 75M words collected from 97 Islamic classical books using a semiautomatic web crawling process. About 867,913 words, representing 1.15% of the corpus text, was in MSA and is crawled from the Internet; while 74,762,008 words contained in 97 books, representing 98.85% of the corpus, was collected from Shamela Library. Additionally, the ANT Corpus [34] is another online MSA corpus of news articles collected from the Tunisian news website; it contains more than 865,500 words collected from 10,000 articles in 9 categories. This corpus can be used for the text classification process. In addition, Arabic Text Corpus [35] is an Arabic text corpus with more than 233k words built from three different sources: Quranic text, Classical Arabic text, and Modern Arabic text. The corpus was collected from the Quran, contemporary Arabic corpora, and the InAra Arabic corpus. According to the authors, the corpus will be freely available to researchers in the future.

Several corpora were also created solely for SD language. One such corpus is the Dialectal Saudi Twitter Corpus (Saudi Dialect) [4], which is an SD corpus containing 207,452 tweets generated by 101 Saudi Twitter users and collected in 2017. Moreover, the SaUdi corpus for NLP Applications and Resources (SUAR) [10] is another SD corpus; it consists of 104,079 words from different online resources (blogs, forums, Instagram, Twitter, WhatsApp, YouTube). The corpus was automatically annotated using the MADAMIRA tool and was considered a pilot study to explore possible directions for facilitating the morphological annotation of the Saudi corpus.

Multi Arabic Dialect Applications and Resources (MADAR) [36] is the first large parallel corpus of 25 Arabic city dialects; it consists of more than 12,000 sentences. The goal of developing MADAR was to create a corpus with large number of dialects and a unified framework with common annotation guidelines and decisions. It can be used in applications such as dialect identification (DID) and machine translation (MT). In addition, the Single-labeled Arabic News Articles Dataset (SANAD) [37] is a large Arabic corpus collected from three news portals, AlKhaleej, AlArabiya, and Akhbarona. It consists of approximately 200k articles in seven categories that are available to the research community for Arabic computational linguistics. SANAD was collected from three main news portals, including AlKhaleej, AlArabiya, and Akhbarona. It is freely available to the public online.

The Jordan Comprehensive Contemporary Arabic (JCCA) corpus [9] is a 100-million-word corpus consisting of MSA as written and spoken in Arab countries. Within the corpus, 87% of the texts are from written sources divided into 9 categories: applied sciences, arts, belief and thought, commerce and finance, imaginative works, leisure, natural and pure sciences, social sciences, and world affairs;. The remaining 13% comes from transcribed spoken language includes transcripts of spontaneous conversations (4.2%) and context-governed spoken language (6.2%) in the categories of educational/informative, business, public/institutional, and leisure.

One of the newest corpora in the literature is Habibi corpus [38], which is a multidialect multinational corpus of Arabic song lyrics from 18 different Arab countries. It consists of 500,000 sentences, 3.5M words from 30,000 Arabic songs from 6 Arabic dialects (Egyptian, Gulf, Levantine, Iraqi, Sudanese and Maghrebi) sung by individuals from 18 different Arabic countries (Egypt, SA, Lebanon, Iraq, Sudan, KW, Syria, UAE, Morocco, Tunisia, Yamen, Jordan, QA, BH, Algeria, OM, Palestine, and Libya). Another corpus, created in 2020, is the Arabic Sentiment Analysis Dataset [39], which is an SD corpus consisting of 15,149 words and is built from tweets discussing several social issues in Saudi Arabia related to the Saudi Vision 2030 framework. It is manually annotated according to the sentiment conveyed in the text and is mainly used for sentiment classification.

From all the literature discussed above, the details of each corpus can be summarized as seen in Table 1.

From Table 1, it can be seen that the SD corpus is a domain that needs further contributions and that the available Arabic corpora are not enough to cover the gaps. First, MSA language dominates most of the large datasets, while SD text is not introduced in many of them. This creates an issue because SD differs from MSA syntactically, morphologically, and phonologically [7]; especially since people are increasingly using Dialectal Arabic, while MSA is limited to formal resources. Regardless of the similarity between the two languages, more SD text needs to be collected to allow Arabic NLP models to process such text.

The second issue in the literature that can be highlighted is that even though there have been attempts to collect SD text, these attempts have been limited in size. As summarized in Table 1, SD corpora only range from thousands of words to a couple million words. Although 2∼3M words might sound a large number, it is not enough for NLP tasks that target dialectal language. This is because, for example, DS does not have standard orthographies, which makes processing it more challenging and thus requires a very large amount of data.

The third limitation of the current SD corpora is the lack of diversity. As shown in Table 1, corpora that include SD language are usually collected from one source (mostly Twitter); only SUAR introduced corpora that include SD language are usually collected from one source (mostly Twitter); only SUAR introduced different sources, yet it only includes 104,079 words. Although Twitter can be a rich source of data, it can show only one social aspect of the community. YouTube and websites are as frequently used as Twitter, and they can show different aspects, especially when targeting resources with diverse titles and topics.

In conclusion, regardless of the rich literature on Arabic corpora, SD corpora are still lacking and need further contributions. Thus, the following section will propose a new model to collect and build a new Saudi corpus that is large, up to date, and diverse.

III. DATA COLLECTION

To collect as much data as possible while using diverse sources, data were collected from the available open corpora in addition to new sources on the web. Due to the difficulty of acquiring valid SD text and how expensive it is to clean such text, it was also decided to introduce MSA text into the collected data. This is because there is some similarity between MSA and dialectical language [40] that can be useful for semantic-based tasks.

The overall statistics of the collected resources are summarized in Table 2. The total amount of collected text was 184,146,256 sentences, 1,238,539,863 words, and 26,674,484 unique words, of which 126,090,964 words were SD. The texts were collected from five different sources, including preexisting corpora that are available to the public, websites, Facebook, YouTube, and Twitter. In terms of the number of unique words from each source, YouTube is the richest source of vocabulary for SD text, while the preexisting corpora are the richest source for MSA text.

From Table 2, it can be seen that some collected data included both MSA and SD (i.e., the source was mixed). This usually happens when the text includes official announcements written in MSA language in addition to comments written by people in SD language. The only exception is in the preexisting corpora, which included either only MSA or only SD text. On YouTube and Twitter, it was difficult to find pure MSA text. The collected text was identified as either SD

TABLE 1. Existing arabic corpora.

#	Corpus	Language	Source	Size	Year	Accessibility	Reuse
1	PATB	MSA	Newswire	+ 1.3 M words	2004-2011	Commercial	Private
2	COLABA	Egyptian Iraqi Levantine Moroccan	Arabic social media	--	2010	Private	Private
3	OSAC	MSA	Websites	+18M words	2010	Public	Public
4	OPUS (MultiUN)	6 languages	Documents	81.41M sentence fragments	2010	Public	Public
5	YADAC	Egyptian	Twitter Blogs Forums	6M word form tokens 457K word form types	2012	Not Available	Not Available
6	Arabic corpus for Egyptian tweets	Egyptian	Twitter	22,834 tweets	2012	Private	Private
7	AWATIF	MSA	PATB Part1 Wikipedia Web forum	5,382 sentences	2012	Private	Private
8	OPUS (TED)	15 languages	TED Talk subtitles	3.81M sentence fragments	2013	Public	Public
9	Multi-Dialectal Corpus of Arabic	MSA Dialectal Arabic	Twitter	62M tweets	2014	Private	Private
10	SANA	English Egyptian Levantine MSA	Newswire Chat turns Twitter YouTube	224,564 entries (sentence)	2014	Private	Private
11	TunDiaWN	Tunisian	Twitter Facebook TripAdviser Theater Dictionaries Transcripts	32,848 words	2014	Private	Private
12	KACST	MSA	Website crawling	+731M words	2014	Restricted	Private
13	1.5 Billion Arabic Corpus	MSA	Newspapers	+1.5B words	2014	Private	Private
14	KSUCCA	Classical Arabic MSA	Almaktabah Alshamilah website	+50M words	2014	Public	Public
15	Gumar	GD MSA	Forum novels	112M words	2016	Restricted	Private
16	Gumar Emirati	UEA	Forum novels	200K words	2018	Restricted	Private
17	Curras	Palestinian	Facebook Twitter Blogs Forums Documents TV Shows	43K Words	2016	Public	Personal
18	Saudi Twitter Corpus	Saudi	Twitter	4,700 tweets	2016	Private	Private
19	BRAD	MSA Dialectal	Book reviews	2,781,805 sentences	2016	Public	Public
20	HARD	MSA Dialectal	Booking.com website	490,587 reviews	2016	Public	Public
21	AraSenTi-Tweet	MSA SD	Twitter	2.2M	2017	Private	Private
22	MD-ArSenTD	KW, SA, QA, UAE, Jordan, Lebanon, Palestine, Syria, Algeria, Morocco, Tunisia, Egypt, MSA	Twitter	14,400 tweets	2017	Private	Private
23	Tashkeela	Classical Arabic MSA	Ancient books Online Libraries	75M words	2017	Public	Public
24	ANT Corpus	MSA	Tunisian news websites	+86500 words	2017	Public	Public
25	Arabic Text Corpus	Classical Arabic MSA	Quran Contemporary Arabic corpus InAra Arabic corpus	233K words	2018	Private	Private

TABLE 1. *(Continued.)* **Existing arabic corpora.**

26	Dialectal Saudi Twitter Corpus	SD	Twitter	207,452 tweets	2017	Public	Public
27	SUAR	SD	Blogs Forums Instagram Twitter WhatsApp YouTube	104,079 words	2018	Private	Restricted
28	MADAR	AD English French MSA	Traveling Expression Corpus	+12K sentences	2018	Restricted	Restricted
29	OPUS (OpenSubtitles)	62 language	Movie subtitles	3.35G sentence fragments	2018-2020	Public	Public
30	SANAD	MSA	News	200K articles	2019	Public	Public
31	JCCA	MSA	Newspapers Books Online sources	100M word	2019	Private	Private
32	Habibi	Egyptian, Gulf, Levantine, Iraqi, Sudanese and Maghrebi	Song lyrics	+3.5M words	2020	Public	Public
33	Arabic Sentiment Analysis Dataset	SD	Twitter	15,149 words	2020	Public	Public

TABLE 2. **Total statistics of collected data.**

Source	Language	No. of Unique Words	No. of Sentences	No. of Words
Pre-existing Datasets	MSA	14,761,449	164,615,349	1,026,315,738
	SD	383,690	515,782	4,144,771
Web Crawling	MSA	1,938,577	722,340	32,529,087
	SD	112,025	547,433	2,524,225
Facebook	MSA	988,479	1,356,175	21,640,486
	Mixed	1,368,222	2,302,685	27,122,297
YouTube	SD	5,453,422	12,565,229	106,599,767
	Mixed	251,193	247,937	4,367,554
Twitter	SD	1,320,775	1,231,816	12,822,201
	Mixed	96,652	41,510	473,737
Total	MSA	17,688,505	166,693,864	1,080,485,311
	SD	7,269,912	14,860,260	126,090,964
	Mixed	1,716,067	2,592,132	31,963,588
	SUM	**26,674,484**	**184,146,256**	**1,238,539,863**

or mixed SD and MSA. This is because both YouTube and Twitter are informal platforms, as is also the case with news websites.

One last interesting observation from Table 2 is that Facebook was not a rich source of SD text. This is because Saudis rarely use this platform for communication and, instead, use YouTube and Twitter. For that reason, most of the resources collected from Facebook were text written with MSA followed by the small number of comments in SD language (mixed with MSA). Further information on the details of each source is described in the following sections.

A. PRE-EXISTING CORPORA COLLECTION

Corpora that are available for public use with MSA and SD language were collected in this stage. The statistical details of the existing corpora that were collected are summarized in Figure 1.[1] Note that the information shown in this figure describes the portion collected and not the whole corpora. This is because some corpora include different languages that are not related to the Saudi language, so they were eliminated in the collection phase.

Starting with MSA corpora, illustrated in Figure 1. (a), it is clear that the literature is rich with MSA text that can be reused. OPUS corpora are the richest source of MSA text and include data from 2010 to 2019. Other corpora cover the years in between. Thus, collected MSA corpora include text related to past events. On the other hand, when considering the SD corpora illustrated in Figure 1. (b), the text is more recent and dated between 2017 and 2020. However, the collected SD corpora are smaller than the MSA corpora, as the former range between 23,000 words and 3M words (5000∼249000 unique words). Thus, more text needs to be collected, and more recent events need to be covered to enrich the literature on SD text.

In addition to the size limitation, the collected corpora are still limited in terms of the number of domains represented. Figure 2 illustrates the 10 domains covered by the collected corpora. To build a diverse and scalable corpus, more domains need to be considered.

B. WEBSITE CRAWLING

Due to the need to collect relatively new text from the web, the web crawling technique was used to collect text from different websites and news outlets. Note that each website has its own selectors and also pages have different selectors. Thus, most of existing crawlers collected all sentences from

[1]Throughout this paper, figures with statistics are represented in logarithmic scale for better visualization.

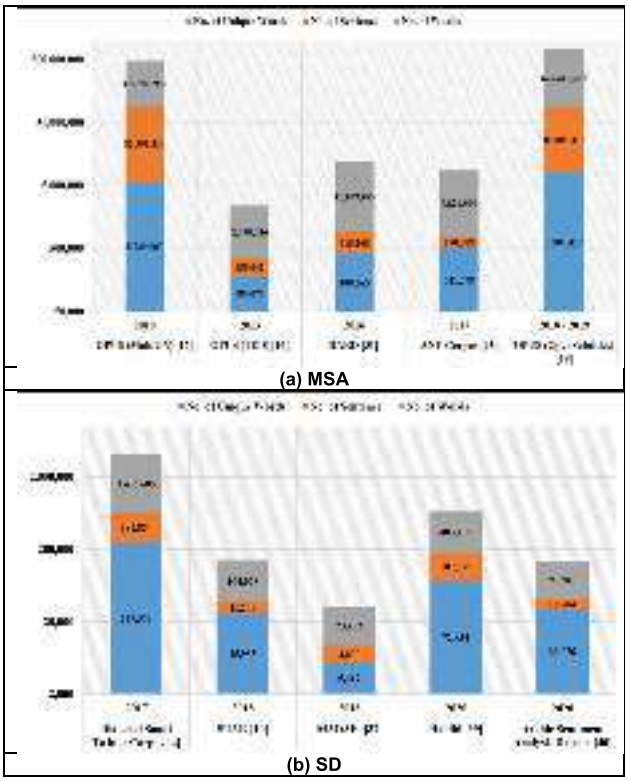


FIGURE 1. Collected existing corpora; (a) statistics on the MSA corpora, (b) statistics on the SD corpora.

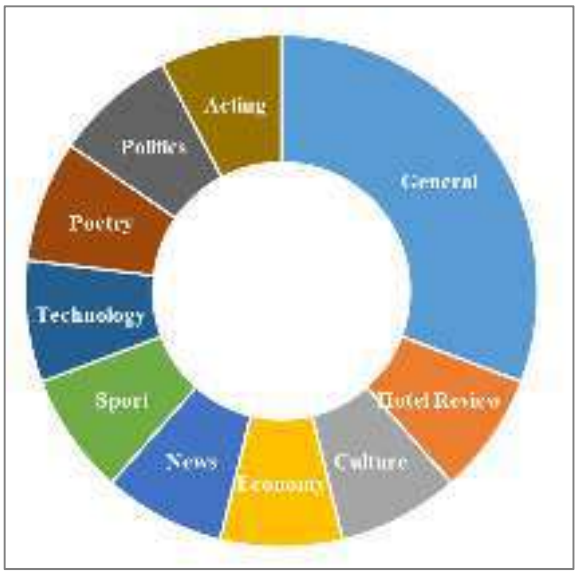


FIGURE 2. Domains of the collected pre-existing corpora.

a page without considering the repetition of the sentences with different selectors. For that reason, a special crawler was programmed to deal with websites' selectors and collect the required text (which is the body of a page). In particular, the implemented crawler works as follows:

1. Scan every link in a website.
2. Extract the content according to given selectors.
3. Save the body according to the language type. If the page has no body, as with main pages or pages that contain links to other pages, nobody will be stored.
4. After crawling all pages, XML tags are removed from the text and only the body of the pages are preserved.
5. Merge all bodies of a website in one text file.

Ten different websites were crawled, as listed in Table 3. These websites are rich in content and cover news or medical topics or sell products. Only Arabic pages were extracted; these were then manually scanned to identify whether they had MSA or SD language or both.

The statistical details of the crawled websites are summarized in Table 3. It can be seen that websites with medical information and news outlets included only MSA text, which is expected since these platforms are formal resources. However, on commercial platforms that sell products, the text extracted was SD because it included the sellers' description of their products in addition to consumers' comments, which were written in their spoken language.

By combining all domains of the crawled text, as illustrated in Figure 3, it can be seen that the total number of words collected from SD websites (+2M words) is much less than that collected from MSA sources (+32M words). It was found that websites can be a rich source of MSA text but not of SD text. This can be attributed to the fact that the young Saudi generation is not using website platforms as those before it have, and social media platforms are becoming more frequently used, as will be emphasized in the following section.

C. SOCIAL MEDIA COLLECTION

Social media platforms are becoming a major research source for rich information, especially for dialectal text [39], and the Saudi community has witnessed a massive increase in the use of social media in the last 10 years [6]. For research purposes, social media platforms have provided APIs for developers, allowing them to collect user comments and blog text fitting certain criteria. In this paper, it was decided to collect text from three social media platforms: YouTube, Twitter, and Facebook. The statistical details of the collected text are summarized in Table 4.

From Table 4, it can be seen that SD text was differentiated better on Twitter and YouTube, while on Facebook, almost 50% of the collected text was mixed and the rest was MSA. In general, the total number of words collected from social media exceeded +17M sentences, +173M words, and +9M unique words. Most of the collected text was SD rather than MSA. Mixed text, however, occurred in all social media platforms. When comparing the sources of social media, it became clear that YouTube was the richest source of data for SD text, and Twitter came in second. The details of each source and how the API was used to collect the data are shown in the following section.

In addition to its size, the diversity of the data is also important. Figure 4 shows the domains of the collected sources and the number of datasets collected for each domain. It is

TABLE 3. Websites crawled.

Website	Language	Domain	No. of Sentences	No. of Words	No. of Unique Words
https://www.webteb.com/	MSA	Medicine	476,298	8,929,536	131,933
https://ajel.sa/	MSA	News	1,163	34,691	10,614
http://aletq.com/	MSA	News	92,609	961,372	163,414
http://al-madina.com/	MSA	News	7,752	273,945	40,349
https://www.bbc.com/arabic	MSA	News	33,255	968,393	101,730
https://twasul.info/	MSA	News	11,150	316,322	56,733
https://ar.wikipedia.org/	MSA	General	100,113	21,044,828	1,433,804
https://sa.opensooq.com/	SD	Economy	363,095	1,698,155	53,483
https://haraj.com.sa/	SD	Economy	79,618	298,346	30,804
https://www.aswaqcity.com/	SD	Economy	104,720	527,724	27,738
			1,269,773	**35,053,312**	**2,050,602**

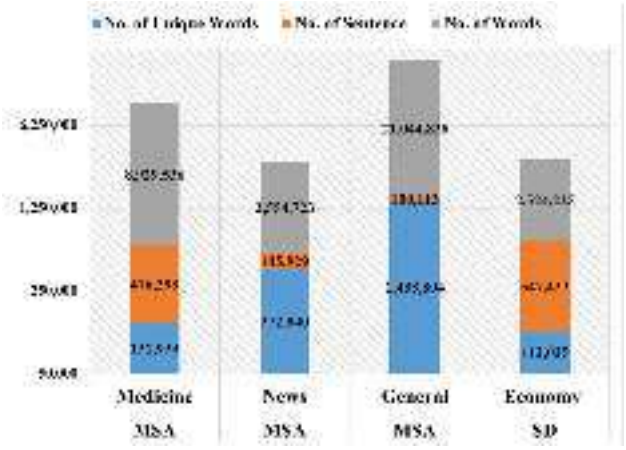


FIGURE 3. Website crawling statistics.

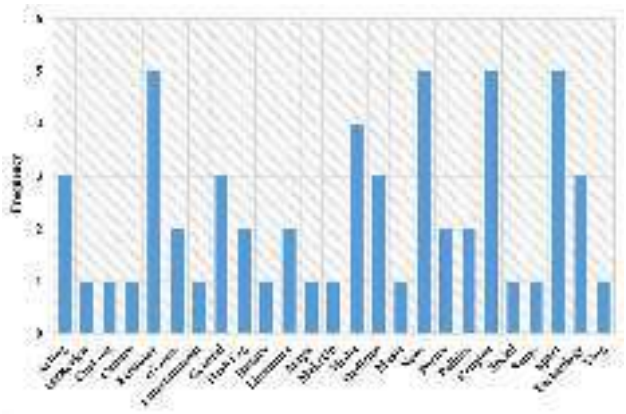


FIGURE 4. Domain distribution across the collected social media dataset.

clear from this figure that the collected data are diverse and cover different domains and topics. Data collected from social media have been categorized into 26 different domains. Each domain provided one to five different datasets. The details of the number of words in each domain are explained next and vary depending on the source from which the text was collected.

1) YOUTUBE RESOURCES

To collect data from YouTube, the commentThread [41] API was used. This tool allows developers to fetch user replies and comments from their channels and videos. In fetching

TABLE 4. Social media data statistics.

Source	Language	No. of Sentences	No. of Words	No. of Unique Words
Facebook	Mixed	2,302,685	27,122,297	1,368,222
	MSA	1,356,175	21,640,486	988,479
		3,658,860	**48,762,783**	**2,356,701**
Twitter	Mixed	41,510	473,737	96,652
	SD	1,231,816	12,822,201	1,320,775
		1,273,326	**13,295,938**	**1,417,427**
YouTube	Mixed	247,937	4,367,554	251,193
	SD	12,565,229	106,599,767	5,453,422
		12,813,166	**110,967,321**	**5,704,615**

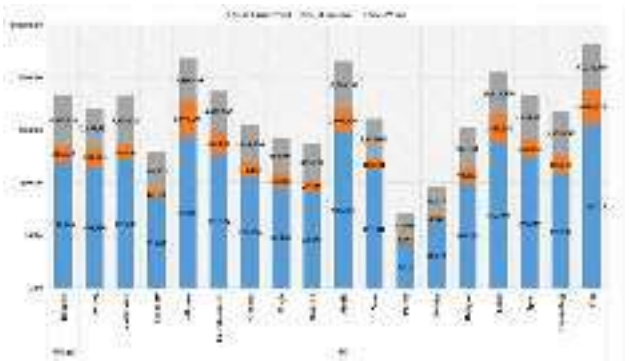


FIGURE 5. Collection statistics for youtube data.

the comments from a channel or a video, it was important to decide what channels and videos to collect text from. First, Arabic channels were examined, and if these channels included videos or playlists introduced by a Saudi announcer, then they were flagged. Then, the comment section of the flagged videos was examined, and if it had a large number of Saudi comments, then the video URL was sent to commentThread to collect the data. As a result of this process, it was possible to collect a total of 110,967,321 words, 12,813,166 sentences, and 5,704,615 unique words from YouTube. As illustrated in Figure 5, the only domain in which SD and MSA could not be differentiated was religious content. This is because people write in their spoken language and include MSA text from the Quran or the Prophet's sayings.

It is interesting to note in Figure 5 that one of the richest sources for SD text on YouTube is vlog-style videos. Users

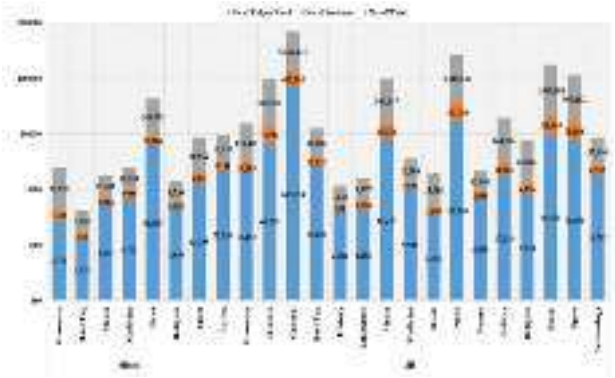


FIGURE 6. Collection statistics for twitter data.

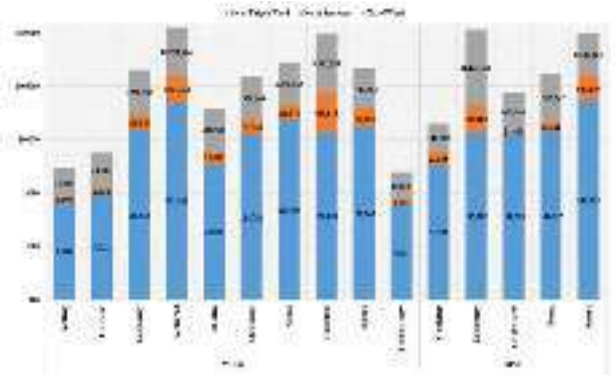


FIGURE 7. Collection statistics for facebook data.

very actively commented on these videos and engaged with the channel owner. On the other hand, poetry content was relatively scarce, and the amount of content in the rest of the domains depended on how active the channel was.

2) TWITTER RESOURCES

To collect relevant and important text from Twitter, different criteria were used during the collection process. The first criterion used was streaming Saudi tweets identified by geographical information. Of course, during this process, non-SD texts were found and labeled as mixed. The second criterion used was archived tweets based on important keywords (hashtags) related to recent events in Saudi Arabia, such as #المرور_السعودي or #شكرا_لكل_معلم_سعودي. A third criterion used to collect the archived data was relevant user accounts, such as well-known Saudi influencers that have millions of Saudi followers and tweets. Note that all tweets, retweets, and comments were also collected.

As illustrated in Figure 6, mixed text is not affected by the domain but rather by the tweet poster. If the posting account is from an official entity, the collected text will be a mix of MSA and SD. This is because the accounts of official entities usually post in MSA, but their tweets become mixed with comments written in SD. Moreover, it was also a challenge to identify the domain of +5M words of text, and thus, these words were labeled as general text. This is because many tweets were replies or discussions about different topics, so they did not belong to a certain domain. The news domain came in second place (following the general domain) in terms of the number of SD words (+2M words) and first place in terms of mixed text (+300,000 words).

3) FACEBOOK RESOURCES

Facebook provides posts and comments related to a given page through GraphAPI [42]. The first step in collecting text from Facebook was to collect pages in different domains, including news, politics, culture, sports, cooking, and advertisements. The FindMyFBID tool [43] was used to convert the page URL to an ID to retrieve the data with GraphAPI. A post and its corresponding comments cannot be retrieved all at once, so all posts from the last three years are retrieved first; then, for each post, a new request is sent to fetch its comments and replies.

From Figure 7, it is clear that SA is not popular on Facebook; thus, the data collected from this platform were either MSA or mixed. As with Twitter, MSA text was related to the owner rather than the domain, so text that was published by official entities was in MSA, while posts written by regular users were a mixture of MSA and SD. Domains such as acting, technology, and culture contained the least SD content, while the rest of the domains are interchangeable.

IV. PROPOSED KSUSC SYSTEM ARCHITECTURE

From the literature on NLP, different preprocessing tools have been developed [44]–[53]. However, due to the diversity of sources and the nature of SD language, it was impossible to use these tools to preprocess the collected text. Using one round of normalization and cleaning would not work on all kinds of data; even the preexisting corpora needed a certain level of preprocessing to be compatible with the proposed KSUSC corpus. Moreover, stemming is very challenging for dialectal Arabic [54] since it does not follow standard Arabic grammar rules for term conflation and attaching affixes to words. Therefore, an incremental process was adopted for the proposed system to continuously update the lexicons and build the KSUSC corpus. One of the advantages of having an incremental system is to avoid the need for stemming and develop a system that can scale and not restrict to certain rules or predefine lexicon. The proposed system was developed using Python regular expression and is divided into five main processes, as illustrated in Figure 8. First, data are processed to build lexicons for all accented characters and unwanted symbols. Second, the data are normalized to unify identical characters that are found with different representations. Third, irrelevant information and repeated sentences and words are removed. Fourth, numbers and English words are masked. Finally, in the fifth step, the corpus is encoded and validated to ensure that all characters are recognized and normalized. This is of particular importance because semantic corpora such as the KSUSC can be used with non-Arabic models such as BERT, which need to be strictly normalized and encoded.

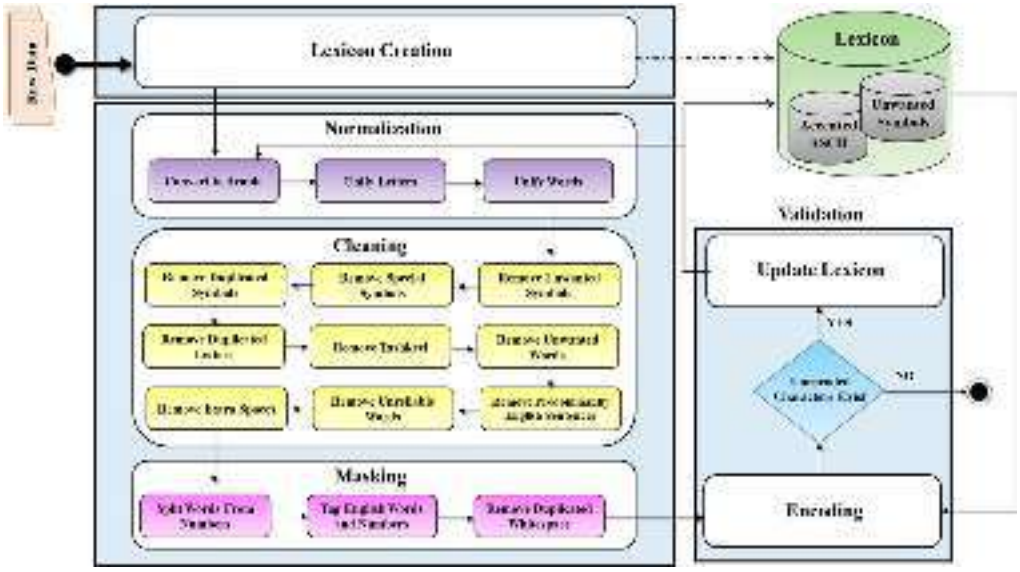


FIGURE 8. KSUSC system architecture.

Updating the lexicons incrementally allows the proposed KSUSC system to be scalable to incorporate new SD documents. The details on how the developed processes works will be explained in the following sections and examples of the proposed system will also be illustrated.

A. LEXICON CREATION

Before starting to normalize and clean the text, it is important to build a lexicon for all variations of letters and symbols and for special unwanted symbols. This is of particular importance because the data are collected from various places, and thus, it is possible that the letters were written in different formats or using different keyboard settings. Moreover, irrelevant symbols and tags could have been included during the collection process. Thus, the lexicon is initially empty, then the collected text is scanned, and all unrecognized symbols and letters are extracted incrementally in the validation phase. Then, the list is investigated to create two lexicons: Accented ASCII and Unwanted Symbols. The 'Accented ASCII' lexicon includes all accented characters and their equivalent ASCII codes; an example is shown in Table 5. The 'Unwanted Symbols' lexicon includes all characters that are not recognized. After creating these lexicons, it is possible to use them with any subsequent new data.

B. NORMALIZATION

While collecting the data, it was noticed that a character's Unicode code can be different from one source to another. For example, some people write numbers in Hindi ASCII, while others write them in Arabic ASCII; sometimes spaces or punctuation are written with Arabic keyboards, while other times, they are written with English keyboards. All these characters have the same meaning but might confuse the semantic model if they are represented differently. Thus, it is important to first normalize the collected data regardless of the source from which it was collected. To do that, Python regular expression was used to implement three normalization processes as follows.

TABLE 5. Sample of ASCII lexicons for 'ا' and 'ي'-accented characters.

Letter	ASCII	Variants	Variants ASCII	Letter	ASCII	Variants	Variants ASCII
ا	1575	أ	1649	ي	1610	ۑ	65268
		ٱ	65166			ۑ	65267
		إ	65165			ي	65266
		آ	64337			ي	65265
		ٱ	1493			ى	1744
		أ	64336			ۑ	64510
		ٱ	1503			ى	1745
						ۑ	64511
						ى	64484
						ۑ	64486

1) CONVERT TO ARABIC

All characters, spaces, and punctuation written in English are converted to Arabic, and the numbers written in Hindi ASCII are converted to English ASCII.

2) UNIFY LETTERS

In Arabic text, letters can be written differently depending on their position in the word or the region of the writer. For example, the letter 'ك' can be written as ک, ﻙ, ك, ﻛ, etc. However, it is impossible for Python regular expressions to recognize the similarity since each variation has a different ASCII code. Thus, accented characters are converted to ASCII characters using the previously built Accented ASCII lexicon.

مقالة جميلة توضح مقارنة رائعة بين Ionic و React_Native وأيهم أفضل، معلومة React Native عمل بهAirbnb، Instagram 1428//26هـ، المعدل بالمراسيم الملكية رقم (م/70) وتاريخ 1437/11/6هـ، ورقم (م/73) وتاريخ 1439/7/18هـ، ورقم (م/115) وتاريخ 1439/12/5هـ، وبعد الصادرة بالقرار الوزاري رقم (7019) وتاريخ 1429/7/3هـ، وبناءً على ما تقتضيه المصلحة العامة.. يُقرر ما يلي:	(a)
مقالة جميلة توضح مقارنة رائعة بين Ionic وReact _ Nativeوأيهم أفضل، معلومة React Native عمل بهInstagram، Airbnb 1428//26هـ، المعدل بالمراسيم الملكية رقم (م/70) وتاريخ 1437/11/6 هجري، ورقم (م/73) وتاريخ 1439/7/18 هجري، ورقم (م/115) وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم (7019) وتاريخ 1429/7/3 هجري، وبناءً على ما تقتضيه المصلحة العامة.. يُقرر ما يلي:	(b)

FIGURE 9. Sample of data in the normalization phase: (a) before normalization, (b) after normalization.

TABLE 6. Abbreviation vs. full words.

Abbreviation	Words
ﷺ	صلى الله عليه وسلم
ﷻ	جل جلاله
هـ	هجري

3) UNIFY WORDS

In Arabic text, some words can be abbreviated using one character and, thus, should be unified as either abbreviated or full words. For example, Table 6 shows some abbreviations and their corresponding words.

After applying the three processes of normalization, all letters should be recognizable and have the same meaning, making them ready to enter the cleaning phase. Figure 9 shows a sample of the text before and after the normalization phase is applied. Note that the changes are in blue.

C. CLEANING

The text cleaning phase is used to eliminate unwanted words and characters and improve the performance of a semantic model. This step removes all HTML tags, extra whitespaces, special characters, punctuation, and duplicate letters as well as any sentences that are not related to the meaning of the text or have been corrupted during collection. In particular, Python regular expression was used to implement the following nine processes that clean the data and remove any noise.

1) REMOVE UNWANTED SYMBOLS

Unwanted symbols (e.g., $;*()" #!@^&-+={}[],|'<>-□﴿﴾□» «؟?؟□□É□￼ 👍⍨) are replaced by whitespace in this task, while full stops and commas are left in the text because they help indicate the length of the sentences. Note that some symbols were not recognized at the beginning, and thus, the Unwanted Symbols lexicon was updated accordingly. For example, Table 7 shows some symbols that were not recognized at the beginning and indicates how they were removed from the text. Moreover, Figure 10 shows a sample of the text before removing unwanted symbols and after removing them (colored in red).

2) REMOVE SPECIAL SYMBOLS

Some special symbols (e.g., \/:%-$) should be removed from the text; however, they have a meaning when they appear with a number. For example, a colon appears with numbers when it is a time (12:30 pm), and a backslash or dash with numbers can represent dates (1/1/2020 or 1-1-2020). Thus, it is important to preserve these special symbols when they occur with numbers and remove them if they do not. Figure 11 shows a sample of the text before and after removing unwanted symbols. Note that the symbols with numbers (highlighted in green) were kept, while the other symbols (highlighted in red) were removed.

3) REMOVE DUPLICATED SYMBOLS

After removing unrelated symbols, it may be noticed that full stop and commas are sometimes unnecessarily repeated. Thus, the duplication should be removed, as shown in Figure 12 (colored in red).

4) REMOVE DUPLICATED LETTERS

Speakers sometimes repeat a letter in a word to emphasize their emotion. For example, in English, people might write ‘Good’ as ‘Goooooood’ to emphasize that they like something. Similarly, this type of redundancy occurs in Arabic and should be removed to standardize the data. For instance, the redundant letters ‘و’ in the word ‘كفوووو’ will be removed to obtain the correct word ‘كفو’. One exception is the word ‘هههه’, which is an expression of laughing out loud. In this expression, the letter ‘هـ’ is repeated several times depending on the user; however, if it is removed, the system might confuse it with the letter ‘هـ’, which means Hijri, and mistakenly transform it to the word ‘هجري’. Thus, the expression for laughing out loud was transformed into two letters instead of one. For example, Figure 13 shows how the word ‘هههههههههه’ is transformed to a word with two letters, ‘هه’, which also means laughing.

5) REMOVE TASHKEEL

In Arabic, there are special punctuation marks called ‘Tashkeel’ that are used to emphasize certain vowel sounds, such as ‘◌ٓ,,,◌ً,◌ٌ,◌ٍ,◌َ,◌ُ,◌ِ,◌ّ,◌ْ,◌ۖ,◌ٗ,ـ,◌ۗ,◌٘,◌ٙ,◌ٚ,◌ۛ,◌ۜ,◌۟,◌۠,◌ۡ,◌ۢ,◌ۣ,◌ۤ,◌ۧ,◌ۨ,◌۪,◌۫,◌۬,◌ۭ,

TABLE 7. Removal of unknown symbols.

Symbol	Text Before Cleaning	Text After Cleaning
ɵ	ضحكت ضحك هه يآبطني ɵ ِ ɵ بطناهة □	ضحكت ضحك هه يآبطني بطناهة
ِ	ضحكت ضحك هه يآبطني ɵ ِ ɵ بطناهة □	
□	ضحكت ضحك هه يآبطني ɵ ِ ɵ بطناهة □	
°	انا حاسه انك حاط ١٨ عشان تجذب المتابع ° ° ʖ	انا حاسه انك حاط ١٨ عشان تجذب المتابع
ʖ	انا حاسه انك حاط ١٨ عشان تجذب المتابع ° ° ʖ	

(a)	وتفيد مجلة «Heliyo» ، بأن علماء جامعة البلطيق الفيدرالية ، بالتعاون مع علماء جامعة مقاطعة كيميروفو ، درسوا تأثير الأشعة فوق. بالقرار الوزاري رقم(7019) "أر أبت إذا صليت الصلو ات المكتو بات و صمت ر مضان و أحللت الحلال و حر مت الحر ام و لم أز د على ذلك شيئا ، أأدخل...
(b)	وتفيد مجلةHeliyo ، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق. بالقرار الوزاري رقم 7019 أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل...

FIGURE 10. Sample of data with unwanted symbols: (a) before removal, (b) after removal.

(a)	26//1428هجري ، المعدل بالمراسيم الملكية رقم (م/70) وتاريخ 6/11/1437هجري ، ورقم (م/73) وتاريخ 1439/7/18هجري ، ورقم (م/115) وتاريخ 5/12/1439هجري ، وبعد الصادرة بالقرار الوزاري رقم (7019) وتاريخ 3/7/1429هجري ، وبناءً على ما تقتضيه المصلحة العامة.. يُقرر ما يلي: خطورة الغيبة محمد مختار الشنقيطير
(b)	26//1428هجري، المعدل بالمراسيم الملكية رقم (م/70) وتاريخ 6/11/1437هجري، ورقم (م/73) وتاريخ 1439/7/18هجري، ورقم (م/115) وتاريخ 5/12/1439هجري، وبعد الصادرة بالقرار الوزاري رقم (7019) وتاريخ 3/7/1429هجري، وبناءً على ما تقتضيه المصلحة العامة.. يُقرر ما يلي خطورة الغيبة محمد مختار الشنقيطير

FIGURE 11. Sample of data with special symbols: (a) before removal, (b) after removal.

(a)	أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا ، أأدخل...
(b)	أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل.

FIGURE 12. Sample of data with duplicate symbols: (a) before removal, (b) after removal.

(a)	ضحكت ضحك هههههههههه يآبطني ɵ ِ ɵبطناهة□
(b)	ضحكت ضحك هه يآبطني ɵ ِ ɵبطناهة□

FIGURE 13. Sample of data with duplicate letters: (a) before removal, (b) after removal.

ً,ٌ,ٍ,َ,ُ,ِ,ّ,ْ,ٓ,ٔ,ٕ,ٖ,ٗ,٘,ٙ,ٚ,ٛ,ٜ,ٝ,ٞ,ٟ,٠,١,٢,٣,٤. These marks are removed to standardize the text, as shown in Figure 14.

6) REMOVE UNWANTED WORDS

In the text, there are words and sentences with special uses that show emotions or are used as metadata, which has no relevance to the semantic meaning of the text itself and includes HTML tags, URLs, emails, hashtags, emojis, and other special metadata. These words and sentences are removed in this step, as shown in Figure 15. There is one exception during the removal process, which is related to hashtags. Since hashtags have meaning when they appear inside a sentence, it was decided that hashtags should be removed only when they appear separately as one word in a line. However, when a hashtag appears in a sentence, it is processed so that the hash and the

(a)	باركَ اللهُ فيكِ يادكتوره خطورة الغيبة كلُ الشرقية تنتظرك بشوق سددَ الباري خطاكِ
(b)	بارك الله فيك يادكتوره خطورة الغيبة كل الشرقية تنتظرك بشوق سدد الباري خطاك

FIGURE 14. Sample of data with arabic punctuation marks: (a) before removal, (b) after removal.

(a)	قم بزيارة الموقع https://website.com/main وارسل بياناتك إلى الإيميل mymai@site.com RT حيو حارتي 2020 أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل ... بارك الله فيك يادكتوره خطورة الغيبة -محمد مختار الشنقيطي تصاميم دعوية، كل الشرقية تنتظرك بشوق، سدد الباري خطاك وياليت قومي يعلمون
(b)	قم بزيارة الموقع وارسل بياناتك إلى الإيميل حيو حارتي 2020 أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا ، أأدخل ... بارك الله فيك يادكتوره خطورة الغيبة -محمد مختار الشنقيطي تصاميم دعوية ، كل الشرقية تنتظرك بشوق ، سدد الباري خطاك وياليت قومي يعلمون

FIGURE 15. Sample of data with unwanted words: (a) before removal, (b) after removal.

(a)	وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية ، بالتعاون مع علماء جامعة مقاطعة كيميروفو ، درسوا تأثير الأشعة فوق اكتبي باليوتويب how to make background to my iphone اتوقع كذا 1428//26هجري ، المعدل بالمراسيم الملكية رقم (م/70) وتاريخ 1437/11/6هجري ، ورقم (م/73) وتاريخ 1439/7/18هجري ، ورقم (م/115) وتاريخ 1439/12/5هجري ، وبعد الصادرة بالقرار الوزاري رقم (7019) وتاريخ 1429/7/3هجري ، وبناءً على ما تقتضيه المصلحة العامة.. يُقرر ما يلي:
(b)	وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي

FIGURE 16. Sample of data with sentences dominated by English words: (a) before removal, (b) after removal.

underscore are replaced with whitespace and the word itself is preserved.

7) REMOVE PREDOMINANTLY ENGLISH SENTENCES

During the cleaning process, sentences with more than 25% of their words in English are removed from the text. Figure 16 shows an example of two sentences that have English words, and one of these sentences will be removed (colored in red).

8) REMOVING UNRELIABLE WORDS

During the collection process, whitespace may be removed in some places so that words are combined. To reduce the need to manually revise these words, it was decided that sentences with words longer than 15 letters should be removed, as illustrated in Figure 17.

9) REMOVE EXTRA SPACES

Extra whitespace occurs between punctuation and text or numbers. Thus, these spaces are removed, as shown in Figure 18.

D. MASKING

After cleaning the text, it is important to mask English words and numbers for future use with semantic models. These words can also be manually translated into Arabic in the

وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعةكيميروفودرسواتاثيرالأشعةفوق اعمل لايكات 256655895566655222544325565665566566 مرة	(a)

FIGURE 17. Sample of data with unreliable words: (a) sentences that will be removed.

وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري ، ورقم م /73 وتاريخ 1439/7/18 هجري ، ورقم م /115 وتاريخ 1439/12/5 هجري ، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 3 / 1429/7 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي	(a)
وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي	(b)

FIGURE 18. Sample of data with duplicate spaces: (a) before removal, (b) after removal.

وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي	(a)
وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي	(b)

FIGURE 19. Sample of data requiring words to be split from numbers: (a) before splitting, (b) after splitting.

future. To automate this process, tags are used for masking, and the following processes are applied.

1) SPLIT WORDS FROM NUMBERS

Before masking the numbers, it is important to separate them from text since they are combined with other words in some documents, as shown in Figure 19.

2) TAG ENGLISH WORDS AND NUMBERS

All English words are masked with the tag [unknown], while numbers are tagged as [number], as shown in Figure 20.

3) REMOVE DUPLICATED WHITESPACE

After applying all the previous steps, whitespace may be duplicated due to replacing unwanted symbols or letters with spaces. Thus, these duplicates are removed, as shown in Figure 21.

E. DATA VALIDATION

After cleaning the data, it is important to validate it and correct any mistakes found. For example, in building this corpus, it was noticed that some symbols were still not recognized, and letters had different ASCII representations even if they had the same meaning. Thus, this phase introduces two main processes used to validate the data and correct unrecognized letters or symbols.

1) ENCODING

All Arabic letters, in addition to full stops and commas, were matched to a list of English letters. Then, the collected data were encoded to English, and any letter or symbol that was not recognized and was not found in the lexicon was added to a list called a non-encoded list.

2) UPDATING LEXICONS

In order to incrementally update the two lexicon proposed in this system, after encoding the documents, the non-encoded list was investigated. If the list was not empty, Arabic letters were added to the matching entry in the Accented ASCII lexicon, and other symbols were added to the Unwanted Symbols lexicon.

The normalization and cleaning phases were repeated after updating the lexicons, and all files were encoded incrementally until the non-encoded list was empty. Figure 22 shows the encoding results after validation. All letters were

حيو حارتي 2020 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي شوف نسبة البطارية بين الساعة 10:03 و الساعة 10:54، الف مبروك الله يوفقكم ويسعدكم وبما انك في سان فرانسيكو جرب المطاعم خاصة كفطور. mamas nopa zazie كل التوفيق	(a)
حيو حارتي number]] انا حاسه انك حاط [numberعشان تجذب المتابع [numberهجري، المعدل بالمراسيم الملكية رقم م [numberوتاريخ [numberهجري، ورقم م [numberوتاريخ number] [numberهجري، ورقم م [numberوتاريخ [numberهجري، وبعد الصادرة بالقرار الوزاري رقم [number] وتاريخ [numberهجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي شوف نسبة البطارية بين الساعة [numberو الساعة number]]، الف مبروك الله يوفقكم ويسعدكم وبما انك في سان فرانسيكو جرب المطاعم خاصة كفطور. [unknown] [unknown] كل التوفيق و النجاح [unknown]	(b)

FIGURE 20. Sample of data when tagging english words and numbers: (a) before tagging, (b) after tagging.

وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي	(a)
وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق 1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ 1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي	(b)

FIGURE 21. Sample of data with duplicate whitespace: (a) before removal, (b) after removal.

translated to a matching English letter, and thus, all words were recognized.

V. USE CASE

For the output of the proposed system, this section presents a full use case and shows the result of each process in the KSUSC system. Starting from the original sample, presented in Figure 23, the text was normalized as shown in Figure 24, cleaned as shown in Figure 25, masked as shown in Figure 26, and finally validated based on the encoding shown in Figure 27. The resulting lexicons can be incrementally updated whenever a new document is introduced.

VI. PROPOSED KSUSC CORPUS DESIGN

After collecting the data from various resources and cleaning them using the KSUSC system, it was decided to design the KSUSC corpus to be the largest Saudi corpus to date. To design this new corpus, the following criteria were validated:

- Corpus Size: The designed corpus is a large corpus with more than 1B words.
- Corpus Languages: The languages of the final corpus are MSA and SD. Introducing MSA text into the collected data would enrich the text due to the similarity between MSA and dialectical language.
- Material Mode: The material of the KSUSC is written text because it can be easily collected and validated. However, in the future, spoken materials will also be considered for inclusion.
- Corpus Dates: The KSUSC corpus covers past materials taken from preexisting corpuses (up to 2010) in addition to recent new content written by the end of 2020.
- Corpus Source: The text was collected from five resources, including preexisting corpora, websites, and different social media platforms.
- Corpus Domains: The KSUSC is a diverse corpus that covers more than 26 domains.

After validating the proposed criteria, it was possible to design and build the KSUSC corpus, which includes a total of 161,795,667 sentences, 1,183,156,600 words, and 14,240,747 unique words. Metadata were introduced to archive the text, and the outcome was copyrighted, as will be explained next. For more details about the corpus statistics, see the Appendix at the end of this paper.

A. TEXT DISTRIBUTION

To uncover the design criteria of the KSUSC corpus, the general statistics are outlined in Table 8. From this table, it is clear that the KSUSC is a large corpus and includes both

حيو حارتي [number] ضحكت ضحك هه يآبطني بطناهة انا حاسه انك حاط [number] عشان تجذب المتابع الفديوهات لاتثبت انه المدرس ولا واضح الشكل ويمكن احد مثلها اساسا مو اثبات صريح وتفيد مجلة [unknown]، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق [number] هجري، المعدل بالمراسيم الملكية رقم [number] وتاريخ [number] هجري، ورقم [number] وتاريخ [number]هجري، ورقم [number] وتاريخ [number] هجري، وبعد الصادرة بالقرار الوزاري رقم [number] وتاريخ [number]هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي شوف نسبة البطارية بين الساعة [number] والساعة [number]، الف مبروك الله يوفقكم ويسعدكم وبما انك في سان فرانسيكو جرب المطاعم خاصة كفطور. [unknown] [unknown] [unknown] كل التوفيق والنجاح الرحمة المهداة صلى الله عليه وسلم أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل بارك الله فيك يادكتوره خطورة الغيبة محمد مختار الشنقيطي، تصاميم دعوية، كل الشرقية تنتظرك بشوق، سدد الباري خطاك ويالبيت قومي يعلمون	**(a)**
Mjh MGQJj number VMcJ VMc gg HWfGgI GfG MGSg Gfc MGW number YTGf JLPH GdeJGHY GdaOjhgGJ dGJKHJ Gfg GdeOQS hdG hGVM GdTcd hjecf GMO eKdgG GSGSG eh GKHGJ UQjM hJajO eLdI unknown , HCf YdeGA LGeYI GdHdWjb GdajOQGdjI, HGdJYGhf eY YdeGA LGeYI ebGWYI cjejQhah, OQShG JCKjQ GdCTYI ahb number gLQj , GdeYOd HGdeQGSje GdedcjI Qbe e number hJGQjN number gLQj , hQbe e number hJGQjN number gLQj , hQbe e number hJGQjN number gLQj , hHYO GdUGOQI HGdbQGQ GdhRGQj Qbe number hJGQjN number gLQj , hHfGA Ydi eG JbJVjg GdeUdMI GdYGeI. jbQQ eG jdj Tha fSHI GdHWGQjI Hjf GdSGYI number h GdSGYI number , Gda eHQhc Gddg jhabce hjSYOce hHeG Gfc aj SGf aQGfSjch LQH GdeWGYe NGUI caWhQ. unknown unknown unknown cd GdJhajb hGdfLGM GdQMeI GdegOGI Udi Gddg Ydjg hSde CQCjJ EPG UdjJ GdUdhGJ GdecJhHGJ hUeJ QeVGf hCMddJ GdMdGd hMQeJ GdMQGe hde CRO Ydi Pdc TjFG, CCONd HGQc Gddg ajc jGOcJhQg NWhQI GdZjHI eMeO eNJGQ GdTfbjWj, JUGeje OYhjI, cd GdTQbjI JfJXQc HThb, SOO GdHGQj NWGc hjGdjJ bhej jYdehf	**(b)**

FIGURE 22. **Sample of data during the validation phase: (a) before encoding, (b) after encoding.**

TABLE 8. **KSUSC distribution across languages.**

Language	No. of Sentences	No. of Words	No. of Unique Words
MSA	146,969,746	1,032,814,633	8,029,018
SD	12,441,955	119,877,091	5,014,593
Mixed	2,383,966	30,464,876	1,197,136
	161,795,667	**1,183,156,600**	**14,240,747**

MSA and SD language. The size of the MSA text is +8M unique words, +146M sentences, and ∼1B words, while the size of the SD and mixed texts is +6M unique words, +14M sentences, and +150M words. Although the total number of words in the KSUSC that come from SD text is much smaller than the total number of words that come from MSA text, the former is still considerably larger than the amount of SD text in all other available corpora. Moreover, when comparing the unique number of words between MSA and SD, it can be seen that MSA constitutes 56% of the total number of unique words, while SD and mixed text constitute the rest. Thus, the KSUSC is still rich in SD vocabularies and morphologies.

With respect to the date criterion, as illustrated in Figure 28, it is clear that the KSUSC date is concentrated towards the end of the date range (between 2018 and 2020). This is of particular importance for ensuring that new vocabularies

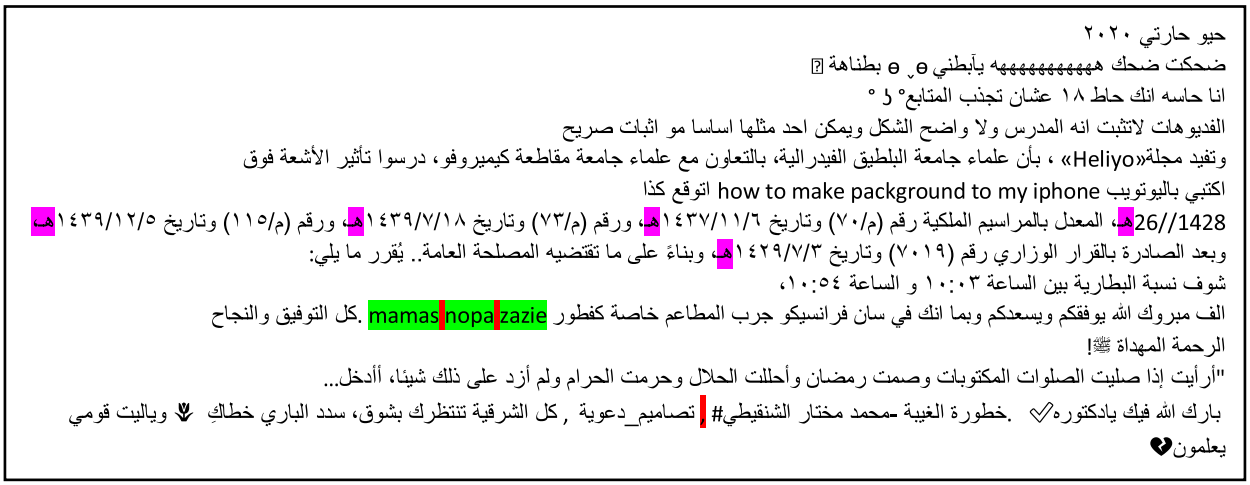

حيو حارتي ٢٠٢٠
ضحكت ضحك هههههههههههه يابطني ɵ ِ ɵ بطناهة ؟
انا حاسه انك حاط ١٨ عشان تجذب المتابع° ل °
الفديوهات لاتثبت انه المدرس ولا واضح الشكل ويمكن احد مثلها اساسا مو اثبات صريح
وتفيد مجلة«Heliyo» ، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق
اكتبي باليوتويب how to make packground to my iphone اتوقع كذا
26//1428هـ، المعدل بالمراسيم الملكية رقم (م/٧٠) وتاريخ ١٤٣٧/١١/٦هـ، ورقم (م/٧٣) وتاريخ ١٤٣٩/٧/١٨هـ، ورقم (م/١١٥) وتاريخ ١٤٣٩/١٢/٥هـ،
وبعد الصادرة بالقرار الوزاري رقم (٧٠١٩) وتاريخ ١٤٢٩/٧/٣هـ، وبناءً على ما تقتضيه المصلحة العامة.. يُقرر ما يلي:
شوف نسبة البطارية بين الساعة ١٠:٠٣ و الساعة ١٠:٥٤،
الف مبروك الله يوفقكم ويسعدكم وبما انك في سان فرانسيكو جرب المطاعم خاصة كفطور mamas nopa zazie .كل التوفيق والنجاح
الرحمة المهداة ﷺ!
"أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل...
بارك الله فيك يادكتوره✓ .خطورة الغيبة -محمد مختار الشنقيطي# تصاميم_دعوية , كل الشرقية تنتظرك بشوق، سدد الباري خطاكِ ويااليت قومي يعلمون

FIGURE 23. Original sample of the use case.

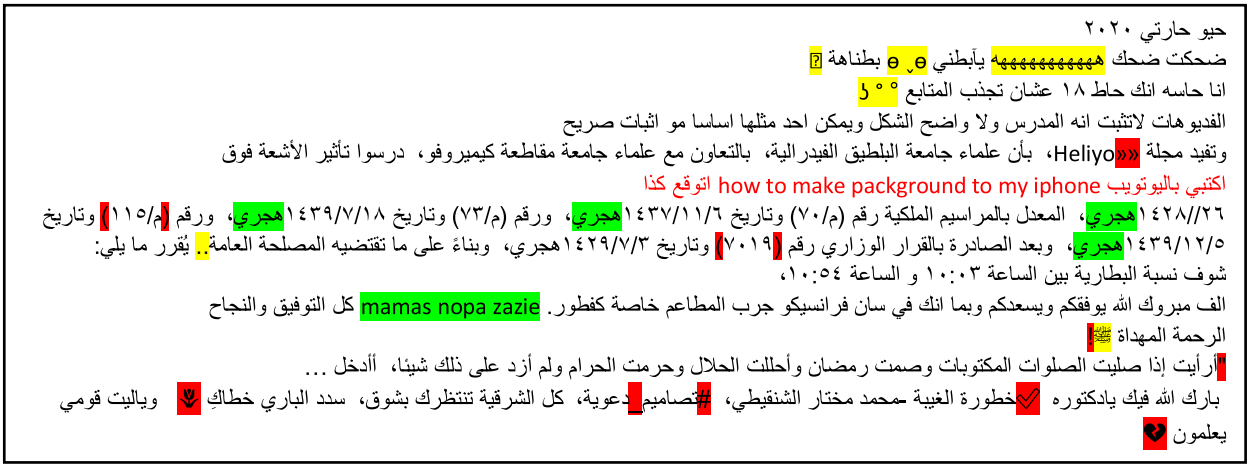

حيو حارتي ٢٠٢٠
ضحكت ضحك هههههههههههه يابطني ɵ ِ ɵ بطناهة ؟
انا حاسه انك حاط ١٨ عشان تجذب المتابع ° ° ل
الفديوهات لاتثبت انه المدرس ولا واضح الشكل ويمكن احد مثلها اساسا مو اثبات صريح
وتفيد مجلة «»Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق
اكتبي باليوتويب how to make packground to my iphone اتوقع كذا
٢٦//١٤٢٨هجري، المعدل بالمراسيم الملكية رقم (م/٧٠) وتاريخ ١٤٣٧/١١/٦هجري، ورقم (م/٧٣) وتاريخ ١٤٣٩/٧/١٨هجري، ورقم (م/١١٥) وتاريخ
١٤٣٩/١٢/٥هجري، وبعد الصادرة بالقرار الوزاري رقم (٧٠١٩) وتاريخ ١٤٢٩/٧/٣هجري، وبناءً على ما تقتضيه المصلحة العامة.. يُقرر ما يلي:
شوف نسبة البطارية بين الساعة ١٠:٠٣ و الساعة ١٠:٥٤،
الف مبروك الله يوفقكم ويسعدكم وبما انك في سان فرانسيكو جرب المطاعم خاصة كفطور. mamas nopa zazie كل التوفيق والنجاح
الرحمة المهداة ﷺ!
"أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل ...
بارك الله فيك يادكتوره ✓خطورة الغيبة -محمد مختار الشنقيطي، #تصاميم_دعوية، كل الشرقية تنتظرك بشوق، سدد الباري خطاكِ ويااليت قومي يعلمون

FIGURE 24. Use case sample after normalization.

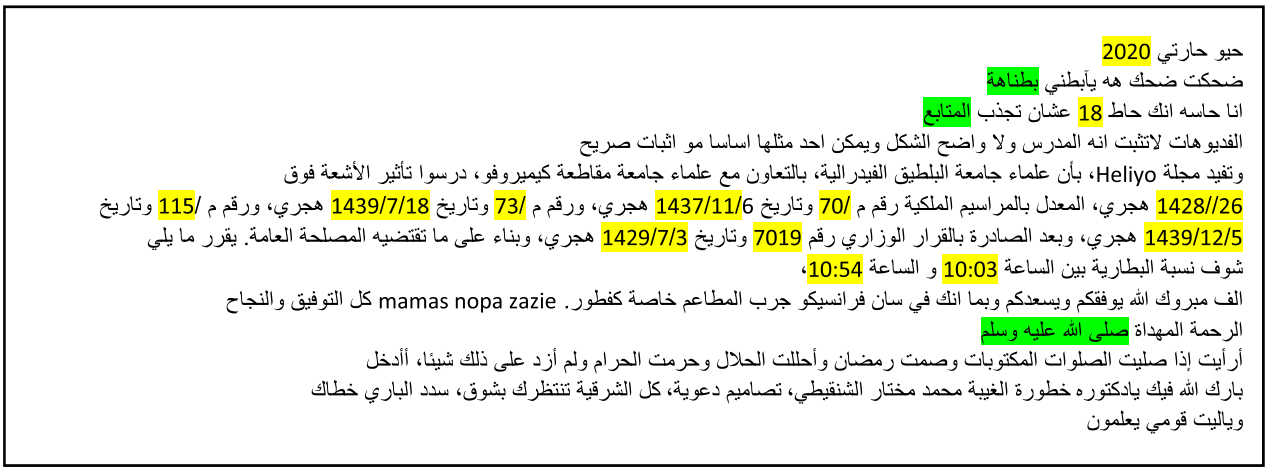

حيو حارتي 2020
ضحكت ضحك هه يابطني بطناهة
انا حاسه انك حاط 18 عشان تجذب المتابع
الفديوهات لاتثبت انه المدرس ولا واضح الشكل ويمكن احد مثلها اساسا مو اثبات صريح
وتفيد مجلة Heliyo، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق
1428//26 هجري، المعدل بالمراسيم الملكية رقم م /70 وتاريخ 1437/11/6 هجري، ورقم م /73 وتاريخ 1439/7/18 هجري، ورقم م /115 وتاريخ
1439/12/5 هجري، وبعد الصادرة بالقرار الوزاري رقم 7019 وتاريخ 1429/7/3 هجري، وبناء على ما تقتضيه المصلحة العامة. يقرر ما يلي
شوف نسبة البطارية بين الساعة 10:03 و الساعة 10:54،
الف مبروك الله يوفقكم ويسعدكم وبما انك في سان فرانسيكو جرب المطاعم خاصة كفطور. mamas nopa zazie كل التوفيق والنجاح
الرحمة المهداة صلى الله عليه وسلم
أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل
بارك الله فيك يادكتوره خطورة الغيبة محمد مختار الشنقيطي، تصاميم دعوية، كل الشرقية تنتظرك بشوق، سدد الباري خطاك
ويااليت قومي يعلمون

FIGURE 25. Use case sample after cleaning phase.

reflecting recent events are covered by the KSUSC. Note that this figure reflects the number of unique words with respect to time in an effort to avoid noisy or redundant text. There are some missing years from the source data, but this does not affect the quality of the corpus since data from the last three years that can help with new tasks and challenges, such as

حيو حارتي [number]
ضحكت ضحك هه يآبطني بطناهة
انا حاسه انك حاط [number] عشان تجذب المتابع
الفديوهات لاتثبت انه المدرس ولا واضح الشكل ويمكن احد مثلها اساسا مو اثبات صريح
وتفيد مجلة [unknown]، بأن علماء جامعة البلطيق الفيدرالية، بالتعاون مع علماء جامعة مقاطعة كيميروفو، درسوا تأثير الأشعة فوق
[number] هجري، المعدل بالمراسيم الملكية رقم [number] وتاريخ [number] هجري، ورقم [number] وتاريخ [number]هجري، ورقم
[number] وتاريخ [number] هجري، وبعد الصادرة بالقرار الوزاري رقم [number] وتاريخ [number]هجري، وبناء على ما تقتضيه
المصلحة العامة. يقرر ما يلي
شوف نسبة البطارية بين الساعة [number] والساعة [number]،
الف مبروك الله يوفقكم ويسعدكم وبما انك في سان فرانسيكو جرب المطاعم خاصة كفطور. [unknown] [unknown] [unknown] كل التوفيق
والنجاح
الرحمة المهداة صلى الله عليه وسلم
أرأيت إذا صليت الصلوات المكتوبات وصمت رمضان وأحللت الحلال وحرمت الحرام ولم أزد على ذلك شيئا، أأدخل
بارك الله فيك يادكتوره خطورة الغيبة محمد مختار الشنقيطي، تصاميم دعوية، كل الشرقية تنتظرك بشوق، سدد الباري خطاك
ويباليت قومي يعلمون

FIGURE 26. Use case sample after masking.

Mjh MGQJj number
VMcJ VMc gg HWfGgI
GfG MGSg Gfc MGW number YTGf JLPH GdeJGHY
GdaOjhgGJ dGJKHJ Gfg GdeOQS hdG hGVM GdTcd hjecf GMO eKdgG GSGSG eh GKHGJ UQjM
hJajO eLdI unknown , HCf YdeGA LGeYI GdHdWjb GdajOQGdjI, HGdJYGhf eY YdeGA LGeYI ebGWYI cjejQhah,
OQShG JCKjQ GdCTYI ahb
number gLQj , GdeYOd HGdeQGSje GdedcjI Qbe e number hJGQjN number gLQj , hQbe e number hJGQjN
number gLQj , hQbe e number hJGQjN number gLQj , hHYO GdUGOQI HGdbQGQ GdhRGQj Qbe number hJGQjN
number gLQj , hHfGA Ydi eG JbJVjg GdeUdMI GdYGeI. jbQQ eG jdj
Tha fSHI GdHWGQjI Hjf GdSGYI number h GdSGYI number ,
Gda eHQhc Gddg jhabce hjSYOce hHeG Gfc aj SGf aQGfSjch LQH GdeWGYe NGUI caWhQ. unknown unknown
unknown cd GdJhajb hGdfLGM
GdQMeI GdegOGI Udi Gddg Ydjg hSde
CQCjJ EPG UdjJ GdUdhGJ GdecJhHGJ hUeJ QeVGf hCMddJ GdMdGd hMQeJ GdMQGe hde CRO Ydi Pdc TjFG,
CCONd
HGQc Gddg ajc jGOcJhQg
NWhQI GdZjHI eMeO eNJGQ GdTfbjWj,
JUGeje OYhjI,
cd GdTQbjI JfJXQc HThb, SOO GdHGQj NWGc
hjGdjJ bhej jYdehf

FIGURE 27. Use case sample after encoding.

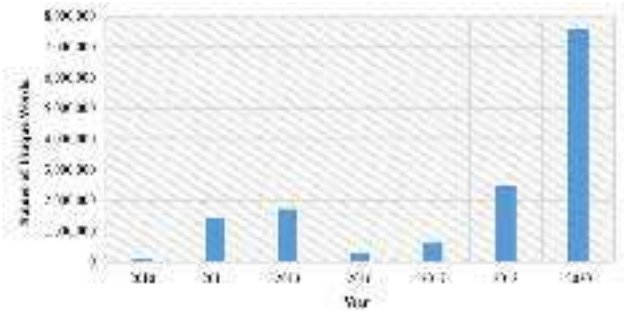


FIGURE 28. KSUSC designed distribution across timelines.

COVID-19 or the Saudi Vision 2030 framework, are of more interest here.

In terms of the text sources, as illustrated in Table 9, more than 102M words in SD language came from YouTube, +10M words came from Twitter, and +2M words came from other web crawled sites. However, +900M words from MSA text resulted from preprocessing the preexisting datasets, indicating that the text was indeed further cleaned and standardized. Moreover, +32M words in MSA resulted from web crawling, and +20M words came from Facebook.

Finally, the domain distribution is another design criterion that must be highlighted with respect to the number of unique words. From Figure 29, it can be seen that the domain that includes the highest number of unique words is the general domain (with 34%). This is because it includes text that is not focused on a certain topic. After that, text in the acting domain

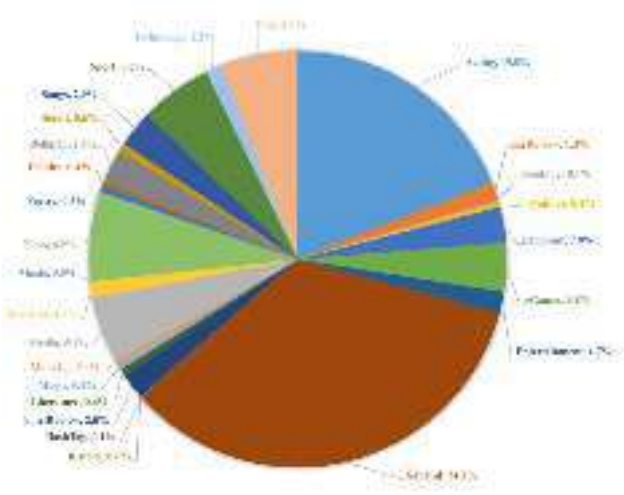


FIGURE 29. KSUSC designed distribution across domains.

TABLE 9. KSUSC designed distribution across sources.

	Language	No. of Sentences	No. of Words	No. of Unique Words
Pre-existing Dataset	MSA	145,193,442	980,222,925	5,642,982
	SD	502,562	4,142,961	335,718
		145,696,004	**984,365,886**	**5,978,700**
Facebook	Mixed	2,112,808	26,001,220	922,787
	MSA	1,127,112	20,409,001	622,403
		3,239,920	**46,410,221**	**1,545,190**
Twitter	Mixed	39,027	447,612	75,539
	SD	1,009,425	10,894,402	769,388
		1,048,452	**11,342,014**	**844,927**
Website	MSA	649,192	32,182,707	1,763,633
	SD	255,565	2,476,153	97,791
		904,757	**34,658,860**	**1,861,424**
YouTube	Mixed	232,131	4,016,044	198,810
	SD	10,674,403	102,363,575	3,811,696
		10,906,534	**106,379,619**	**4,010,506**

comes in second place, constituting 19% of all unique words, and the news domain comes in third place with approximately 7% of all unique words. Next, the corpus is distributed similarly between the rest of the 23 domains, with the music (0.02%) and history (0.03%) domains having the smallest number of unique words.

Although these statistics show the percentage of unique words in each domain, it must be clarified that these domains are based on the source text and not the vocabulary itself. In other words, if a word appears in sports-related text, this does not mean that the word does not appear in news-related text; rather, this means that it will not appear again in the same domain. Therefore, it was important to characterize the corpus with different metadata and archive it for further use.

B. TEXT METADATA

Any corpus in the literature has to be categorized and described with metadata. To facilitate further improvement in the future, it was decided to record as much information as possible about the corpus collected. In particular, KSUSC metadata included the source name, domain, file name, URL of the sources (or query/region), year of the text, and the total number of words, the number of sentences, and the number of unique words. These criteria will allow researchers to restrict their data to a specific source, period of time, domain, or even a query or region. These clear design criteria and metadata will be helpful in the corpus compilation process and when validating the accuracy of a certain task that is applied across the corpus.

C. PLAUSIBLE APPLICATIONS

This corpus is a general-purpose Saudi corpus that can be used when developing NLP applications for Saudi language or for linguistic studies at different domains. In particular, semantic-based applications are the targeted applications that can use the corpus as it is; in which methods can be applied on the collected text to measure the similarity between words in the same sentences and understand what it means accordingly. Examples of such applications are semantic analysis and search query. Note that any pre-trained models, such as BERT, can directly use the corpus to fine-tune over Saudi Dialectical text. Moreover, if the data is further annotated, classification methods can also be applied on the annotated corpus and used for more specific applications such as sentiment analysis [55]–[59] or sarcasm identification [60], [61].

D. CHALLENGES AND DIFFICULTIES

The process of collecting SD text revealed different challenges, and the construction of the KSUSC presented many difficulties. These challenges can be summarized as follows:

- It was difficult to distinguish SD text from other Gulf DA languages, and Saudi dialect experts had to be consulted.
- The social media text was very dirty and needed intensive incremental cleaning.
- No common SD lexicons were found; thus, they had to be created from scratch.
- Social media platforms had many restrictions when collecting the data, which extended the collection process to obtain enough text for the corpus.
- Some domains did not exist in current sources.
- Not all preexisting corpora were available or free to access.

Because of these challenges, the source of each text was included in the design criteria, but some domains and years were not fully covered.

E. COPYRIGHTS

The designed corpus was built from sources available online, and some have an active copyright, such as newspapers, magazines, books, tweets, and websites. Thus, the following actions have been taken: (1) bibliographic information about the corpus content is provided; (2) previews of the full text are restricted and are not available to the public; and (3) the collected text is not distributed and is locally used for research purposes. The corpus will be used according to the previously stated restrictions, and because it is intended for research

TABLE 10. KSUSC Detailed Statistics.

Source	Language	Category	Time	No. of Unique Words	No. of Sentences	No. of Words
Pre-existing Datasets	SD	General	2017	220,643	381,709	3,444,749
		General	2018	5,250	3,953	21,634
		Poetry	2020	62,595	99,125	477,770
		Politics	2020	22,810	7,417	95,481
		Social	2018	24,420	10,358	103,327
	MSA	Culture	2017	52,736	11,350	345,937
		Technology	2017	57,504	24,532	680,140
		Sport	2017	75,923	44,173	1,460,770
		News	2017	144,306	158,439	4,560,388
		Economy	2017	49,535	26,684	827,246
		hotelReview	2016	282,899	512,512	11,029,537
		Acting	2020	2,543,126	64,164,808	373,045,451
		General	2010	131,984	148,668	2,049,168
		General	2013	1,734,837	69,699,319	351,853,997
		General	2018	570,132	10,402,957	234,370,291
Website	SD	Economy	2020	97,791	255,565	2,476,153
	MSA	Medicine	2018	103,682	437,011	8,886,837
		News	2018	226,147	112,068	2,251,042
		General	2011	1,433,804	100,113	21,044,828
YouTube	SD	Acting	2020	141,894	214,421	1,926,650
		carReview	2020	190,485	251,437	3,858,461
		Economy	2020	45,452	18,655	296,064
		eGames	2020	495,436	2,401,397	18,667,161
		Entertainment	2020	238,918	490,526	4,457,690
		Magic	2020	58,450	65,064	555,379
		MakeUp	2020	55,393	37,453	441,649
		Media	2020	644,047	1,383,250	17,070,306
		General	2020	110,800	106,465	995,822
		News	2020	158,247	108,220	1,400,449
		Poetry	2020	6,798	2,067	17,859
		Politics	2020	15,325	4,748	62,413
		Religion	2020	72,616	82,043	891,185
		Songs	2020	405,908	1,158,668	10,412,711
		Sport	2020	220,248	349,165	3,975,687
		Technology	2020	94,785	150,694	1,840,975
		Vlog	2020	856,894	3,850,130	35,493,114
	Mixed	Religion	2020	198,810	232,131	4,016,044
Twitter	SD	Acting	2020	19,374	6,491	57,696
		Economy	2020	14,473	8,528	94,534
		eGames	2020	45,464	58,468	626,587
		HashTag	2020	20,124	9,630	82,683
		History	2020	3,606	606	6,730
		Literature	2020	3,527	1,477	8,864
		Media	2020	57,556	58,728	721,909
		Medicine	2020	8,604	2,156	22,760
		General	2020	301,142	501,076	4,853,438
		Music	2020	2,887	1,089	13,012
		News	2020	134,916	178,791	2,124,959
		Poetry	2020	4,748	1,216	9,266
		Politics	2020	12,669	9,560	131,335
		Religion	2020	6,111	4,256	49,634
		Social	2020	63,760	93,446	1,208,296
		Sport	2020	56,051	67,863	829,431
		Technology	2020	14,376	6,044	53,268
	Mixed	Economy	2020	2,369	1,342	20,056
		Hashtag	2020	904	307	2,132
		Media	2020	4,171	1,340	10,408
		Medicine	2020	5,570	1,905	13,972
		News	2020	48,891	27,553	336,968
		Religion	2020	3,919	852	8,514
		Sport	2020	9,715	5,728	55,562
Facebook	MSA	Cooking	2018	20,852	13,412	112,411
		Economy	2018	98,465	248,226	10,666,947
		Literature	2018	76,101	18,997	576,846
		News	2018	111,391	63,026	1,513,749
		Sport	2018	315,594	783,451	7,539,048
	Mixed	Acting	2018	6,038	1,289	20,498
		Culture	2018	7,313	2,394	41,547
		Economy	2018	87,463	98,739	1,422,899
		Media	2018	23,322	22,993	310,189
		Medicine	2018	64,813	71,616	1,008,600
		General	2018	362,386	944,522	10,656,985
		News	2018	159,900	188,972	2,274,879
		Religion	2018	89,348	606,274	8,582,885
		Sport	2018	117,848	174,813	1,667,435
		Technology	2018	4,356	1,196	15,303
Total				**14,240,747**	**161,795,667**	**1,183,156,600**

purposes, it is consistent with the current Saudi copyright law [62].

VII. CONCLUSION

A language corpus is one of the most important sources of data for researchers and can represent written language around the world. However, Arabic corpora lack sufficient research interest and data due to the time-consuming challenges faced when designing and building such a corpus. DA text is particularly challenging because it does not have standard orthographies. This paper surveyed 33 Arabic corpora to confirm that SD corpora are still in need of further expansion. Thus, +1.2B words of text were collected from five different sources. These sources included existing corpora and new text collected from websites and social media platforms in order to include past and recent vocabularies.

Thus, the paper introduced a new preprocessing system that is incremental and scalable to incorporate new data sources. The system validated the collected data and eliminated irrelevant characters and incomplete text. The incremental creation of SD lexicons will allow the system to scale to other languages in DA. As a result of this system, it was possible to design and build a new KSUSC corpus that is large, diverse, and up to date. The KSUSC corpus includes MSA and SD language covering 26 different domains. It is a large corpus with +1B words, +161M sentences, and +14M unique words. More than 35% of all unique words in the KSUSC are in SD language.

The scalability of the preprocessing system and the diversity and large size of the designed corpus will allow researchers to use the KSUSC for a wide number of tasks and to integrate their own corpora. Important steps that the authors are considering in the future include providing the preprocessing system as a tool for public use with different DA text and introducing text from speech resources into the corpus. This will allow the system to enrich the SD lexicons even further and allow other researchers to contribute to and reuse the system. Moreover, in the future, the corpus will be used with semantic-based applications such as semantic analysis and query search. The effect of mixing the SD text with MSA will be further investigated.

APPENDIX

See Table 10.

REFERENCES

[1] I. A. El-Khair, "Abu El-Khair corpus: A modern standard Arabic corpus," *Int. J. Recent Trends Eng. Res.*, vol. 2, no. 11, pp. 1–9, 2016.

[2] A. O. Al-Thubaity, "A 700M+ Arabic corpus: KACST Arabic corpus design and construction," *Lang. Resour. Eval.*, vol. 49, no. 3, pp. 721–751, Sep. 2015, doi: 10.1007/s10579-014-9284-1.

[3] United Nations. *Official Languages*. [Online]. Available: https://www.un.org/en/sections/about-un/official-languages/index.html#:~:text=There%20are%20six%20official%20languages,%2C%20French%2C%20Russian%20and%20Spanish

[4] M. Alruily, "Issues of dialectal Saudi Twitter corpus," *Int. Arab J. Inf. Technol.*, vol. 17, no. 3, pp. 367–374, May 2020, doi: 10.34028/iajit/17/3/10.

[5] N. Habash, R. Eskander, and A. Hawwari, "A morphological analyzer for Egyptian Arabic," in *Proc. 12th Meeting Special Interest Group Comput. Morphol. Phonol.*, 2012, pp. 1–9.

[6] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, "AraSenTi-tweet: A corpus for Arabic sentiment analysis of Saudi tweets," *Procedia Comput. Sci.*, vol. 117, pp. 63–72, Jan. 2017, doi: 10.1016/j.procs.2017.10.094.

[7] N. Y. Habash, "Introduction to Arabic natural language processing," *Synth. Lect. Hum. Lang. Technol.*, vol. 3, no. 1, pp. 1–187, Jan. 2010.

[8] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Trans. Asian Lang. Inf. Process.*, vol. 8, no. 4, pp. 1–22, Dec. 2009, doi: 10.1145/1644879.1644881.

[9] M. Sawalha, F. Alshargi, A. AlShdaifat, S. Yagi, and M. A. Qudah, "Construction and annotation of the Jordan comprehensive contemporary Arabic corpus (JCCA)," in *Proc. 4th Arabic Natural Lang. Process. Workshop*, 2019, pp. 148–157.

[10] N. Al-Twairesh, R. Al-Matham, N. Madi, N. Almugren, A.-H. Al-Aljmi, S. Alshalan, R. Alshalan, N. Alrumayyan, S. Al-Manea, S. Bawazeer, N. Al-Mutlaq, N. Almanea, W. B. Huwaymil, D. Alqusair, R. Alotaibi, S. Al-Senaydi, and A. Alfutamani, "SUAR: Towards building a corpus for the Saudi dialect," *Procedia Comput. Sci.*, vol. 142, pp. 72–82, Jan. 2018, doi: 10.1016/j.procs.2018.10.462.

[11] H. Mubarak and K. Darwish, "Using Twitter to collect a multi-dialectal corpus of Arabic," in *Proc. EMNLP Workshop Arabic Natural Lang. Process. (ANLP)*, Stroudsburg, PA, USA, 2014, pp. 1–7, doi: 10.3115/v1/W14-3601.

[12] G. Lebboss, G. Bernard, N. Aliane, A. Abdallah, and M. Hajjar, "Evaluating methods for building Arabic semantic resources with big corpora," in *Proc. 9th Int. Joint Conf. Comput. Intell. (IJCCI)*, Funchal, Portugal, C. Sabourin, J. J. Merelo, K. Madani, and K. Warwick, Eds. Cham, Switzerland: Springer, 2019, pp. 179–197.

[13] M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki, "The Penn Arabic treebank: Building a large-scale annotated Arabic corpus," in *Proc. NEMLAR Conf. Arabic Lang. Resour. Tools*, Sep. 2016, pp. 102–109.

[14] M. Diab, N. Habash, O. Rambow, M. Altantawy, and Y. Benajiba, "COLABA: Arabic dialect annotation and processing," in *Proc. LREC Workshop Semitic Lang. Process.*, Jan. 2016, pp. 66–74, 2010.

[15] M. K. Saad and W. M. Ashour, "OSAC: Open source Arabic corpora," Osac Open Source Arabic Corpora, Tech. Rep., 2010, vol. 10.

[16] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," stanbul, Turkey, Tech. Rep., 2012, pp. 2214–2218.

[17] A. Eisele and Y. Chen, "MultiUN: A multilingual corpus from united nation documents," presented at the 7th Int. Conf. Lang. Resour. Eval. (LREC), 2010.

[18] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*, 2016, pp. 923–929.

[19] R. Al-Sabbagh and R. Girju, "YADAC: Yet another dialectal Arabic corpus," in *Proc. 8th Int. Conf. Lang. Resour. Eval. (LREC)*, 2012, pp. 2882–2889.

[20] R. Al-Sabbagh and R. Girju, "A supervised POS tagger for written Arabic social networking corpora," in *Proc. 11th Conf. Natural Lang. Process. KONVENS, Empirical Methods Natural Lang. Process.*, vol. 5, Sep. 2012, pp. 39–52.

[21] M. Abdul-Mageed and M. Diab, "AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis," in *Proc. 8th Int. Conf. Lang. Resour. Eval. (LREC)*, 2012, pp. 3907–3914.

[22] M. Abdul-Mageed and M. Diab, "SANA: A large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, 2014, pp. 1162–1169.

[23] R. Bouchlaghem, A. Elkhlifi, and R. Faiz, "Tunisian dialect wordnet creation and enrichment using Web resources and other wordnets," in *Proc. EMNLP Workshop Arabic Natural Lang. Process. (ANLP)*, Stroudsburg, PA, USA, 2014, pp. 104–113, doi: 10.3115/v1/W14-3613.

[24] I. A. El-khair, "1.5 Billion words Arabic corpus," 2016, *arXiv:1611.04033*. [Online]. Available: http://arxiv.org/abs/1611.04033

[25] M. Alrabiah, A. Al-Salman, and E. Atwell, "The design and construction of the 50 million words KSUCCA," in *Proc. 2nd Workshop Arabic Corpus Linguistics*, 2013, pp. 5–8.

[26] S. Khalifa, N. Habash, D. Abdulrahim, and S. Hassan, "A large scale corpus of Gulf Arabic," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*, 2016, pp. 4282–4289.

[27] S. Khalifa, N. Habash, F. Eryani, O. Obeid, D. Abdulrahim, and M. A. Kaabi, "A morphologically annotated corpus of Emirati Arabic," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, 2019, pp. 3839–3846.

[28] A. Assiri, A. Emam, and H. Al-Dossari, "Saudi Twitter corpus for sentiment analysis," *Int. J. Comput. Inf. Eng.*, vol. 10, no. 2, pp. 272–275, 2016.

[29] A. Elnagar and O. Einea, "BRAD 1.0: Book reviews in Arabic dataset," in *Proc. IEEE/ACS 13th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2016, pp. 1–8, doi: 10.1109/AICCSA.2016.7945800.

[30] A. Elnagar, Y. S. Khalifa, and A. Einea, "Hotel Arabic-reviews dataset construction for sentiment analysis applications," in *Intelligent Natural Language Processing: Trends and Applications*, K. Shaalan, A. E. Hassanien, and F. Tolba, Eds. Cham, Switzerland: Springer, 2018, pp. 35–52.

[31] R. Baly, G. El-Khoury, R. Moukalled, R. Aoun, H. Hajj, K. B. Shaban, and W. El-Hajj, "Comparative evaluation of sentiment analysis methods across Arabic dialects," *Procedia Comput. Sci.*, vol. 117, pp. 266–273, Jan. 2017, doi: 10.1016/j.procs.2017.10.118.

[32] Y. Yamamoto, "Twitter4j—A java library for the Twitter API," Tech. Rep., 2014.

[33] T. Zerrouki and A. Balla, "Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems," *Data Brief*, vol. 11, p. 147, Apr. 2017.

[34] A. Chouigui, O. B. Khiroun, and B. Elayeb, "ANT corpus: An Arabic news text collection for textual classification," in *Proc. IEEE/ACS 14th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Oct. 2017, pp. 135–142, doi: 10.1109/AICCSA.2017.22.

[35] K. Meskaldji, S. Chikhi, and I. Bensalem, "A new multi varied Arabic corpus," in *Proc. 3rd Int. Conf. Pattern Anal. Intell. Syst. (PAIS)*, Oct. 2018, pp. 1–5, doi: 10.1109/PAIS.2018.8598524.

[36] H. Bouamor, N. Habash, M. Salameh, W. Zaghouani, O. Rambow, D. Abdulrahim, O. Obeid, S. Khalifa, F. Eryani, A. Erdmann, and K. Oflazer, "The MADAR Arabic dialect corpus and lexicon," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, 2019, pp. 3387–3396.

[37] O. Einea, A. Elnagar, and R. Al Debsi, "SANAD: Single-label Arabic news articles dataset for automatic text categorization," *Data Brief*, vol. 25, Aug. 2019, Art. no. 104076, doi: 10.1016/j.dib.2019.104076.

[38] M. El-Haj, "Habibi—A multi dialect multi national Arabic song lyrics corpus," in *Proc. 12th Lang. Resour. Eval. Conf.*, Marseille, France, 2020, pp. 1318–1326.

[39] S. N. Alyami and S. O. Olatunji, "Application of support vector machine for Arabic sentiment classification using Twitter-based dataset," *J. Inf. Knowl. Manage.*, vol. 19, no. 1, Mar. 2020, Art. no. 2040018.

[40] N. Habash and O. Rambow, "MAGEAD: A morphological analyzer and generator for the Arabic dialects," in *Proc. 21st Int. Conf. Comput. Linguistics, 44th Annu. Meeting Assoc. Comput. Linguistics*, 2006, pp. 681–688.

[41] YouTube. (2020). *CommentThreads: YouTube Data API*. [Online]. Available: https://developers.google.com/youtube/v3/docs/commentThreads

[42] FaceBook. (2020). *Graph API: FaceBook for Developers*. [Online]. Available: https://developers.facebook.com/docs/graph-api/

[43] FindMyFBID. *Find Your Facebook ID*. [Online]. Available: https://findmyfbid.com/

[44] A. Pasha, M. Al-Badrashiny, M. T. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. Roth, "MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, May 2014, pp. 1094–1101.

[45] N. Habash and O. Rambow, "Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop," in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2005, pp. 573–580.

[46] N. Habash, O. Rambow, and R. Roth, "MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization," in *Proc. 2nd Int. Conf. Arabic Lang. Resour. Tools (MEDAR)*, vol. 41, 2009, p. 62.

[47] M. Diab, K. Hacioglu, and D. Jurafsky, "Automated methods for processing Arabic text: From tokenization to base phrase chunking," in *Arabic Computational Morphology: Knowledge-Based and Empirical Methods*. Norwell, MA, USA: Kluwer, 2007.

[48] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A fast and furious segmenter for Arabic," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Demonstrations*, Stroudsburg, PA, USA, Jun. 2016, pp. 11–16, doi: 10.18653/v1/N16-3003.

[49] O. Obeid, N. Zalmout, S. Khalifa, D. Taji, M. Oudah, B. Alhafni, G. Inoue, F. Eryani, A. Erdmann, and N. Habash, "CAMeL tools: An open source Python toolkit for Arabic natural language processing," in *Proc. 12th Lang. Resour. Eval. Conf.*, May 2020, pp. 7022–7032. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.868

[50] M. Althobaiti, U. Kruschwitz, and M. Poesio, "AraNLP: A java-based library for the processing of Arabic text," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, 2014, pp. 4134–4138.

[51] D. A. Said, N. M. Wanas, N. M. Darwish, and N. H. Hegazy, "A study of text preprocessing tools for Arabic text categorization," in *Proc. 2nd Int. Conf. Arabic Lang. Resour. Tools*, Jan. 2009, pp. 230–236.

[52] M. Attia, "A large-scale computational processor of the Arabic morphology, and applications," Fac. Eng., Cairo Univ., Egypt, Giza, Egypt, Tech. Rep., 2000.

[53] K. M. Darwish, "Probabilistic methods for searching OCR-degraded Arabic text," Tech. Rep., 2004.

[54] G. Alwakid, T. Osman, and T. Hughes-Roberts, "Towards improved saudi dialectal Arabic stemming," in *Proc. Int. Conf. Comput. Inf. Sci. (ICCIS)*, Apr. 2019, pp. 1–5, doi: 10.1109/ICCISci.2019.8716408.

[55] A. Onan, S. Korukoglu, and H. Bulut, "LDA-based topic modelling in text sentiment classification: An empirical analysis," *Int. J. Comput. Linguistics Appl.*, vol. 7, no. 1, pp. 101–119, 2016.

[56] A. Onan, "Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach," *Comput. Appl. Eng. Educ.*, vol. 29, no. 3, pp. 572–589, May 2021.

[57] A. Onan, "Sentiment analysis on Twitter based on ensemble of psychological and linguistic feature sets," *Balkan J. Electr. Comput. Eng.*, vol. 6, no. 2, Apr. 2018, Art. no. 2, doi: 10.17694/bajece.419538.

[58] A. Onan, "Deep learning based sentiment analysis on product reviews on Twitter," in *Big Data Innovations and Applications*. Cham, Switzerland: Springer, 2019, pp. 80–91.

[59] A. Onan, "Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks," *Concurrency Comput., Pract. Exper.*, p. e5909, Jun. 2020.

[60] A. Onan, "Topic-enriched word embeddings for sarcasm identification," in *Software Engineering Methods in Intelligent Algorithms*. Cham, Switzerland: Springer, 2019, pp. 293–304.

[61] A. Onan and M. A. Toçoğlu, "A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification," *IEEE Access*, vol. 9, pp. 7701–7722, 2021, doi: 10.1109/ACCESS.2021.3049734.

[62] Saudi Authority for Intellectual Property. *Copyright Law Issued by Royal Decree No. M/41 Dated 2/7/1424 AH Amended by the Council of Ministers Resolution No. (536) Dated 19/10/1439 AH*. [Online]. Available: https://www.saip.gov.sa/wp-content/uploads/2019/10/Copyright-Law.pdf

HEBAH ELGIBREEN is currently an Assistant Professor with the Department of Information Technology, College of Computer and Information Sciences, King Saud University (KSU). She is also a Research Affiliate with the Department of Mechanical Engineering (MRL Lab), Massachusetts Institute of Technology (MIT); the Director of the AI Center of Advance Study, KSU; and a member of the Center of Smart Robotics Research, KSU. She is specialized in artificial intelligence and machine learning. Her research currently focuses on using ML approaches to improve collaborative robotics motions in shared environment.

MOHAMMED FAISAL received the master's and Ph.D. degrees from King Saud University, in 2012 and 2016, respectively. He currently works as an Assistant Professor and supervise the Unit of Innovation and Entrepreneurship, College of Applied Computer Science, King Saud University, where he also a Robotics Consultant with the Center of Smart Robotics Research. In recent years, he has published scores of scientific research in refereed and classified scientific journals and conferences. In 2019, he granted a patent for a tree harvesting tool from the USA, and authored the book titled *Developing Protocols for Monitoring Using Wireless Sensor Network and Unmanned Aerial Vehicle*, in 2013. His research is currently focusing on the use of robots and artificial intelligence to improve the quality of life. He won the President of the Republic of Yemen Youth Award for the Applied Sciences Branch, in 2013, and many scientific and research awards and medals.

MANSOUR AL SULAIMAN (Member, IEEE) received the Ph.D. degree from Iowa State University, USA, in 1987. Since 1988, he has been with the Department of Computer Engineering, King Saud University (KSU), Riyadh, Saudi Arabia, where he is currently a Professor with the Department of Computer Engineering and the Director of the Center of Smart Robotics Research. His research areas include automatic speech/speaker recognition, automatic voice pathology assessment systems, computer-aided pronunciation training systems, and robotics. He was the Editor-in-Chief of the *Journal of King Saud University Computer and Information Sciences*.

SHERIF ABDOU received the B.Sc. and M.Sc. degrees in computer science and automatic control from the University of Alexandria, Egypt, in 1993 and 1997, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Miami, USA, in 2003. In 2003, he joined BBN Technologies as a Senior Staff Scientist with the Arabic Language Team of the Ears Project to provide affordable reusable speech-to-text decoding for the Defense Advanced Research Projects Agency, DARPA. In 2005, he was appointed as the

Research and Development Manager of Research and Development International (RDI) Company, where he is leading a team to develop several products for Human Language Technologies, such as NLP, ASR, TTS, OCR, and language teaching with a special focus on the technologies of the Arabic language. In 2005, he joined Cairo University, where he is currently a Professor and the Chair of the Department of Information Technology, Faculty of Computers and Information. He is also a member of the review committee and has more than 60 papers in distinguished conferences and articles in journals in the HLT field. He is also the principal investigator and the co-principal investigator of several research projects in the areas of language learning, virtual tutors, Web monitoring, and intelligent contact centers.

MOHAMED AMINE MEKHTICHE was born in Medea, Algeria, in 1987. He received the B.S. and M.S. degrees in electronic engineering from the University of Blida, in 2010 and 2012, respectively. From 2014 to 2021, he was a Researcher with the Center of Smart Robotic Research, King Saud University, Saudi Arabia. His current research interest includes image processing stereo vision.

ABDULLAH M. MOUSSA (Member, IEEE) received the B.Sc. degree from the Department of Electrical Engineering, Suez Canal University, Egypt, and the M.Sc. degree from the Department of Electrical Engineering, Port Said University, Egypt. He is currently a Research Assistant with the Faculty of Computers and Artificial Intelligence, Cairo University. He has several publications in different domains, including natural language processing and image processing. His current research interests include natural language processing, computer vision, and machine learning. He is also an Active Reviewer for several journals, including IEEE Access, IEEE Transactions on Industrial Electronics, and IEEE Transactions on Intelligent Transportation Systems.

YOUSEF A. ALOHALI received the Ph.D. degree in AI from the School of Engineering, Concordia University, Montréal, Canada. He is currently an Associate Professor of artificial intelligence with the Department of Computer Science, King Saud University, Riyadh, Saudi Arabia. He led digital transformation in a number of national organizations at various time spans, including King Saud University (2007–2012), Ministry of Higher Education (2013–2014), and Ministry of Education (2015–2016). He headed the executive office at Tatweer for Educational Technologies (TETCO), from 2016 to 2019. He has a number of publications in the fields of AI and its applications.

WADOOD ABDUL received the Ph.D. degree in signal and image processing from the University of Poitiers, France, in 2011. He is currently working as an Associate Professor with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University. He has published over 70 papers in well-reputed conferences and articles in journals. He developed the Communications Laboratory by Lucus Nulle and the Biometrics Laboratory funded by ZKTeco, King Saud University. His research interests are focused on multimedia security, biometrics, agriculture applications, privacy, medical image processing, and video understanding, where he is working on several externally funded research projects. He received the Best Faculty Award from the College of Computer and Information Sciences, King Saud University, in 2017.

GHULAM MUHAMMAD (Senior Member, IEEE) received the B.S. degree in computer science and engineering from the Bangladesh University of Engineering and Technology, in 1997, the M.S. and Ph.D. degrees in electronic and information engineering from Toyohashi University and Technology, Japan, in 2003 and 2006, respectively. He is currently a Professor with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University (KSU), Riyadh, Saudi Arabia. He is involved in many research projects as a principal investigator and a co-principal investigator. He has supervised more than 15 Ph.D. and Master Thesis. He has authored or coauthored more than 250 publications, including IEEE/ACM/Springer/Elsevier journals, and flagship conference papers. He owns two U.S. patents. His research interests include signal processing, machine learning, the IoTs, medical signal and image analysis, AI, and biometrics. He received the Best Faculty Award of the Department of Computer Engineering, KSU, from 2014 to 2015. He was a recipient of the Japan Society for Promotion and Science (JSPS) Fellowship from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

MOHSEN RASHWAN is currently a Professor of communications with the Department of Electronics and Communications, Faculty of Engineering, Cairo University, Egypt. He had over 100 articles published in international proceedings and conferences. He had over 70 thesis under his supervision (finished and current: 26 Ph.D.'s and 47 M.Sc.'s). Many of his ex-postgraduate (M.Sc. and Ph.D.) students are currently recruited in the core of top world hi-tech companies and research centers, such as IBM-WRC, Lucent Technologies, and Microsoft. Over 350 graduation projects are realized under his supervision with the Department of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, in different applications, such as the digital processing of speech, audio, image, and video, pattern recognition or classification, OCR, document analysis, biometry, software, and hardware design. He has been awarded many prizes to some of those projects. He has a grant of research and development of many projects from the Information Technology Academia Collaboration (ITAC) Program initiated by the Information Technology Industry Development Agency (ITIDA) for RDI exclusively to produce its Arabic off-line OCR. From April 2010 to June 2011, he was the principal investigator of this project. He had Mediterranean Arabic Language and Speech Technologies (MEDAR). This project runs under the EU's FP7 Research and Development Grant Program. This project is shared by 15 partners from 11 M.E. and European countries, from February 2008 to August 2010.

MOHAMMED ALGABRI received the master's degree from King Saud University, where he is a currently pursuing the Ph.D. degree with the Department of Computer Science, College of Computer and Information Sciences. His research interests include speech processing, pronunciation error detection, and deep learning.

PLOS ONE

RESEARCH ARTICLE

Semantic textual similarity for modern standard and dialectal Arabic using transfer learning

Mansour Al Sulaiman[1,5☯], Abdullah M. Moussa[2☯*], Sherif Abdou[2], Hebah Elgibreen[3,5,6], Mohammed Faisal[4,5,6], Mohsen Rashwan[7]

1 Department of Computer Engineering, College of Computer and Information Sciences (CCIS), King Saud University, Riyadh, Saudi Arabia, **2** Faculty of Computers and Artificial Intelligence, Cairo University, Giza, Egypt, **3** Information Technology Department, College of Computer and Information Science, King Saud University, Riyadh, Saudi Arabia, **4** Center of AI & Robotics, Kuwait College of Science and Technology (KCST), Kuwait City, Kuwait, **5** Center of Smart Robotics Research, College of Computer and Information Science, King Saud University, Riyadh, Saudi Arabia, **6** Artificial Intelligence Center of Advance Studies (Thakaa), King Saud University, Riyadh, Saudi Arabia, **7** Faculty of Engineering, Cairo University, Giza, Egypt

☯ These authors contributed equally to this work.
* a.m.moussa@ieee.org

Citation: Al Sulaiman M, Moussa AM, Abdou S, Elgibreen H, Faisal M, Rashwan M (2022) Semantic textual similarity for modern standard and dialectal Arabic using transfer learning. PLoS ONE 17(8): e0272991. https://doi.org/10.1371/journal.pone.0272991

Editor: Omar A. Alzubi, Al-Balqa Applied University Prince Abdullah bin Ghazi Faculty of Information Technology, JORDAN

Received: January 5, 2022

Accepted: July 31, 2022

Published: August 11, 2022

Data Availability Statement: All relevant data are within the paper and its Supporting information files.

Funding: The authors extend their appreciations to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number DRI-KSU-1292. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Semantic Textual Similarity (STS) is the task of identifying the semantic correlation between two sentences of the same or different languages. STS is an important task in natural language processing because it has many applications in different domains such as information retrieval, machine translation, plagiarism detection, document categorization, semantic search, and conversational systems. The availability of STS training and evaluation data resources for some languages such as English has led to good performance systems that achieve above 80% correlation with human judgment. Unfortunately, such required STS data resources are not available for many languages like Arabic. To overcome this challenge, this paper proposes three different approaches to generate effective STS Arabic models. The first one is based on evaluating the use of automatic machine translation for English STS data to Arabic to be used in fine-tuning. The second approach is based on the interleaving of Arabic models with English data resources. The third approach is based on fine-tuning the knowledge distillation-based models to boost their performance in Arabic using a proposed translated dataset. With very limited resources consisting of just a few hundred Arabic STS sentence pairs, we managed to achieve a score of 81% correlation, evaluated using the standard STS 2017 Arabic evaluation set. Also, we managed to extend the Arabic models to process two local dialects, Egyptian (EG) and Saudi Arabian (SA), with a correlation score of 77.5% for EG dialect and 76% for the SA dialect evaluated using dialectal conversion from the same standard STS 2017 Arabic set.

Introduction

Recognizing the similarity between two sentences is a vital process in many applications since the text is one of the most important media for communication [1]. This makes Semantic

Competing interests: The authors have declared that no competing interests exist.

Textual Similarity (STS) a critical pre-step in several domains such as information retrieval, document classification, machine translation, textual summarization, question answering, short answer grading, semantic search, and conversational systems [2]. For example, In the information retrieval problem, the most common criterion used to retrieve information is key sentences. Given a set of available media such as documents or videos, millions of them for practical applications, the user can query the system by entering a sentence to describe the content of the medium to be viewed. The same medium can be retrieved using several sentences. i.e., the user can use a different query other than the key sentences that are associated with the medium to describe it. For any efficient retrieval process, the system should be able to recognize the correlation between similar, but different, queries [3].

STS and sentence embeddings

While there are several ways to tackle the problem of STS, the most promising ones are based on word/sentence embeddings. Sentence embeddings are vector representations of sentences in which each vector is mathematically close in the space to other vectors that represent semantically close meaning. Embeddings can be calculated using different algorithms such as Word2Vec [4], GloVe [5], and BERT [6]. BERT and BERT-Like models are generally based on self-supervised machine learning techniques that make use of the huge amounts of unlabeled text data available on the internet. While BERT is not intentionally created to generate embeddings, it can be adjusted to generate sentence embeddings of good quality. BERT models set new state-of-the-art performance on various sentence classification and sentence-pair regression tasks. To generate a sentence-pair similarity score, BERT uses a cross-encoder: Two sentences are passed to the transformer network and the target value is predicted using a simple regression method for the output. However, this setup is unsuitable for various applications due to the high number of possible combinations to be checked. In [7], the authors proposed a method to generate effective sentence embeddings from BERT models, and several other models have been suggested for such a line of adaptations.

Motivation

While measuring semantic similarity of texts is applied widely for some languages, for example, English, The Arabic version of the problem has three main limitations. The first one is that the methods proposed to handle the problem for the Arabic language are not of good performance. The second issue is that the development of STS models always requires the availability of semantic similarity annotated corpus with considerable size [8]. Unfortunately, this type of resource is not available for low resources languages such as Arabic. The third problem is that the written form of dialectal Arabic doesn't have lexical standards. So, there is always a need for approaches that can minimize the gap between the performance of Arabic STS models and the level of STS models of widely investigated languages like English. The motivation of this work is to overcome these challenges. and to provide a methodology for handling these issues. The general advantages and contributions of this work are provided in the next section.

Contributions

The main contributions proposed in the paper are the following:

- Proposing three approaches to tackle the problem of Arabic STS. The first is to use automatic machine translation to translate English STS data to Arabic and to use the translated data for converting Arabic BERT models into STS Arabic models. The second approach is to interleave English STS data with Arabic BERT models to generate enhanced Arabic STS models.

The third approach is based on knowledge distillation models that are optimized using proposed translated Arabic STS datasets.

- The development of a new data resource of professional translation for 1.3K pairs of sentences from their original form in English to MSA, Egyptian Arabic, and Saudi Arabic versions.
- Proposing different models that advance the state-of-the-art performance in the STS task in MSA with limited resources.
- The development, to the best of our knowledge, of first STS models for Egyptian Arabic and Saudi Arabic.

The rest of the paper is organized as follows: Section 2 illustrates the related work and literature review; Section 3 provides the details of the proposed approaches, the developed datasets, and the developed models. Section 4 includes the experimental results and Section 5 includes comparisons with the state-of-the-art results. Finally, section 6 includes the conclusions and some prospects for our planned future work.

Related work

Lexical-based similarity

Because semantic textual similarity has many applications in natural language processing, the general form of the problem has attracted a lot of attention from the community [9, 10]. However, it has gained a less but considerable interest regarding the Arabic language. While there are several methods tried to tackle the problem, these methods can be categorized into two main tracks: lexical-based similarity and semantic-based similarity [11]. Lexical-based similarity relies on calculating the correlation between the character streams of two sentences to be compared. This process can be applied to the level of characters or the level of words. While applying this process to the level of characters is relatively simple, it is not robust enough to extract the real similarity between two sentences. Computing the correlation between two texts based on words is better than character level [12]. Methods for measuring similarity between words are using several distance measures to compute the relevance between two terms [13]. Some examples of these measures are Jaccard distance and Levenshtein distance [14, 15].

Semantic-based similarity

Semantic-based sentence similarity methods can be divided into three classes: word-based sentence similarity, structure-based sentence similarity, and vector-based sentence similarity methods [13]. In word-based sentence similarity, the sentence is handled as a list of words, and the correlation between the words in the two sentences is compared [16]. In structure-based sentence similarity, several methods have been suggested that use language grammar [17], Part-Of-Speech (POS) [18] and words order [19]. Vector-based sentence similarity methods rely on calculating sentence embeddings that describe each sentence as a mathematical vector. These methods are based on corpus analysis. The vector representing each sentence can be calculated by training a model using a sufficiently large corpus. Many techniques have been presented to provide sentence embeddings. For example, Kiros et al. in [20] proposed a method named Skip-Thought that trains an encoder-decoder framework to try to predict the surrounding sentences. In [21], the authors proposed a method that uses siamese transformers and siamese DAN networks to generate sentence embeddings. Cer et al. [22] proposed Universal Sentence Encoder which used unsupervised learning with a transformer network. Conneau

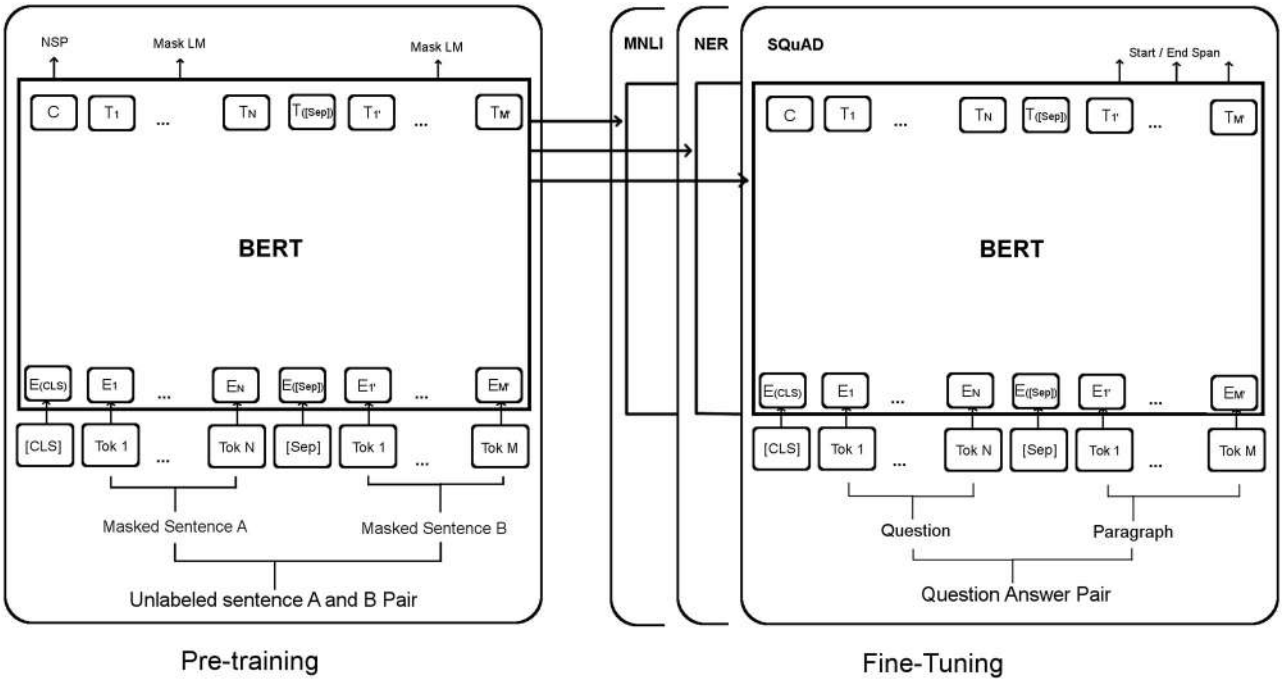


Fig 1. The main framework of BERT.

https://doi.org/10.1371/journal.pone.0272991.g001

et al. [23] proposed InferSent, a siamese BiLSTM network with max-pooling over the output. This method used labeled data of Stanford Natural Language Inference dataset (SNLI) [24] and the MultiGenre NLI dataset (MultiNLI) [25].

BERT embeddings

The main recent approaches to calculate sentence embeddings are based on utilizing robust language models such as BERT. BERT (Fig 1), which stands for Bidirectional Encoder Representations from Transformers, is designed to train masked language models from an unlabeled text by conditioning on both left and right contexts in all layers of a transformer network. Such a language model randomly masks a specific percentage of input tokens and the objective of the training is to predict the original masked tokens using only their context. BERT-based models can be used to generate sentence embeddings. There are several ways to utilize BERT for generating sentence embeddings. For example, by averaging the BERT output layer which is known as BERT embeddings, or by using the embedding of a special token the BERT uses as the first token for each input sentence (Known as the [CLS] token). Also, The BERT can be used in a sentence-pair regression mode to generate a similarity score. However, the embeddings generated by these methods are either not of good quality or not practical for most applications [7].

Several techniques have been proposed to enhance the accuracy of BERT-based sentence embeddings. For example, in [7], the authors present Sentence-BERT (SBERT). The SBERT model [7] takes as input a pair of sentences into siamese architecture which consists of two instances of a base model. Each instance produces an embedding using a pooling procedure. The two embeddings are compared and the manual estimated correlation scores are used to train the model for being oriented to the semantic similarity. In the testing phase, the testing

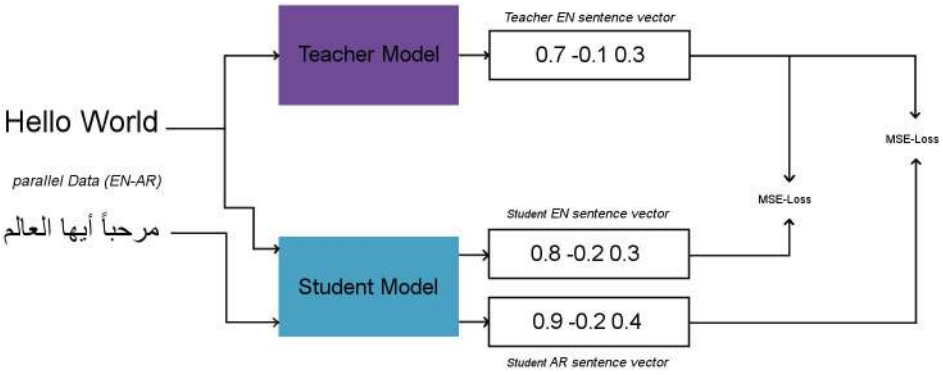


Fig 2. Given parallel data from two languages, a student model can be trained such that the generated vectors for the two languages sentences are close to the teacher language sentence vector.

https://doi.org/10.1371/journal.pone.0272991.g002

pairs of sentences are given as input to the same architecture and produce a cosine similarity value for each pair of sentences that can be compared with the manual given reference correlation score. The SBERT has been shown to achieve state-of-art performance for the English language STS tasks. To transfer such good performance to other languages, especially those with limited resources, a knowledge distillation approach was proposed [26]. In [26], the authors proposed an efficient method to extend existing sentence embedding models to new languages. Network learning is based on the concept that the original sentence and translated sentences should be mapped in the same location in the vector space. Given, for example, a teacher model of English, they presented an approach to train a student model of another language. They use the original teacher model to produce sentence embeddings for the source language and train a new system using translated sentences to simulate the original model. Fig 2 illustrates an overview of the method. However, using such a technique needs considerable amounts of parallel data from multiple languages to be effective.

The state-of-the-art

Transfer learning-based solutions for STS have been used in several recent studies. For example, in [27], the authors presented an STS system based on transfer learning. They used an approach that is utilizing RoBERTa [28] models and applied their work to a biomedical dataset. Their proposed methodology obtained an accuracy of 0.9. However, this accuracy was based on domain-specific data. Also in [29], Mutinda et. al. proposed Japanese BERT-based models for textual similarity. They also created two datasets that targeted the clinical medical domain to test their presented systems. They achieved a score of 0.904 on the clinical domain dataset. Furthermore, Yang et. al, in [30] explored 3 transformer-based models for clinical STS, BERT, XLNet [31] and, RoBERTa. They examined transformer models pre-trained using both clinical text and general English text. Their best-performing system was based on a RoBERTa model and obtained a Pearson correlation of 0.9065. However, such good results were due to applying the system to a domain-specific dataset.

Some techniques have been presented to handle the Arabic STS problem. In [26], the authors applied their knowledge distillation-based model on a standard Arabic dataset for testing proposed by [8] and got 79.1 based on Spearman rank correlation. Also, in [32] Nagoudi and Schwab proposed a combination of word embedding and word alignment techniques and then calculated sentence embedding as a sum of its content of word vectors to tackle the Arabic STS problem [9]. Also in [33], Nagoudi et al. proposed a sentence vectors-based method for

the cross-lingual similarity between Arabic and English sentences. and they found that using weighting based on POS can enhance their output results.

Proposed datasets and approaches

Data

In [8], the authors presented the evaluation of their organized task for Multilingual STS. They have proposed datasets for being used to train and test STS proposed models. The datasets are formatted into pairs of sentences. For each pair, there is a given manual score that indicates the correlation between the two sentences. This score is ranging from 0 (no correlation) to 5 (exact meaning). Table 1 provides some examples of various degrees of correlation between each pair of sentences in the STS datasets.

While Arabic STS was one of their organized tracks, the authors of [8] have provided an MSA Arabic dataset for training, This work adds to them a translation of another 1379 pairs of sentences from the English STS data. The translation has been completed by professional experts. A translation for the same dataset to Egyptian Arabic and Saudi Arabic variants has also been provided by this work. A dataset for testing has been presented in [8]. It consists of 250 pairs of sentences of MSA Arabic. The structure of the testing dataset and training dataset is similar. We proposed a translation of this testing dataset to Egyptian Arabic and Saudi Arabic to be used in evaluation. It is worth mentioning that the testing dataset presented by [8] is a standard measure that is used by state-of-the-art papers (for example [26]). Table 2 illustrates some examples of the proposed translations along with their original English texts.

Methodology

To develop our Arabic STS models, three approaches have been used. The first one is to train an SBERT-based model. Such a model is based on an Arabic BERT model that is converted to SBERT structure and fine-tuned using automatic translation to Arabic of the SNLI [24] and

Table 1. Examples of different levels of correlation between the sentences in STS dataset.

Correlation	Example
5	**The two sentences have the exact same meaning**
	I don't see why there should be any problem with this whatsoever.
	I don't see why that should be a problem.
4	**Some unimportant details are different but the two sentences are almost the same**
	A black and white photo of a man driving a car and someone with a motorcycle.
	A black and white photo of a man in a classic car and a man with a classic motorcycle.
3	**The two sentences are roughly equivalent, but there are some important different details**.
	A woman is talking on a cell phone.
	A man and woman are talking on the phone.
2	**The two sentences are not the same but they share some of the details**
	A man is playing the piano.
	A man played the guitar.
1	**The two sentences share the same topic but they are not equivalent**.
	A person is slicing some onions.
	A woman is chopping herbs.
0	**The two sentences are completely different**
	The train heads down the tracks and along the hedge.
	A dog on the floor of a patio looks at a cat on the fence.

https://doi.org/10.1371/journal.pone.0272991.t001

Table 2. Some examples of the proposed translations along with original English sentences.

Correlation	Language/Dialect	Sentence1	Sentence2
5	English	There is a cook preparing food.	A cook is making food.
	MSA	هناك طباخ يعد الطعام	الطباخ يصنع الطعام.
	Egyptian	فيه طباخ بيجهز الأكل	الطباخ بيجهز الاكل.
	Saudi	فيه طباخ قاعد يسوي الأكل.	فيه طباخ قاعد يسوي الأكل.
4	English	You have to decide what you want to get out of this.	You have to find out what works for you.
	MSA	عليك أن تقرر ما تريد للتمكن من الخروج من هذا	عليك أن تعرف ما يناسبك.
	Egyptian	المفروض تعرف انت عايز ايه عشان تقدر تخرج من الموضوع ده	لازم تعرف اللي يناسبك
	Saudi	لازم تعرف وش تبي عشان تقدر تطلع من هالشي.	لازم تعرف اللي يناسبك.
3	English	A harnessed dog leaps over a log as another dog follows behind.	A brown and white dog is leaping over a log in a field while another dog follows behind it.
	MSA	يقفز كلب ملجم فوق جذع خشبي بينما يتبعه كلب آخر.	يقفز كلب بني وأبيض فوق جذع خشبي في حقل بينما يتبعه كلب آخر.
	Egyptian	كلب متكم فوق جذع شجرة وفي كلب تاني بيعمل زيه	كلب بني في ابيض بينط فوق جذع خشبي في جنينة وراه كلب تاني
	Saudi	كلب مكمم يطمر على جذع خشبي ويلحقه كلب ثاني.	كلب لونه بني وأبيض يطمر على جذع خشبي في مزرعة ويلحقه كلب ثاني.
2	English	Two men wearing traditional clothing is standing outside.	Three women wearing black vests and gray shirts are talking outside of a building.
	MSA	يقف رجلان يرتديان ملابس تقليدية في الخارج.	ثلاث نساء يرتدين سترات سوداء وقمصان رمادية يتحدثن خارج مبنى.
	Egyptian	رجلين واقفين بره لابسين هدوم عادية	تلت ستات لابس جواكت سوده وقمصان رمادي بيتكلموا برا المبني
	Saudi	رجالين لابسين ملابس رحية وواقفين برا.	ثلاث حريم لابسين جكيتات سوداء وبلايز رصاصية يسولفون برا المبنى.
1	English	My answer to your question is "Probably Not".	This is a part answer to your question
	MSA	جوابي عن سؤالك هو ربما لا.	هذا جزء من إجابة سؤالك
	Egyptian	جواب عالسؤال بتاعك هو ممكن لا	دة جزء من إجابة سؤالك
	Saudi	جوابي على سؤالك هو ممكن لا.	هذا جزء من جواب سؤالك.
0	English	There is no test that can tell you if it is sealed or not.	There is no code telling you that you cannot do this.
	MSA	لا يوجد اختبار يمكن أن يخبرك ما إذا كان محكم الإغلاق أم لا.	لا يوجد رمز يخبرك أنه لا يمكنك القيام بذلك.
	Egyptian	مفيش حاجة ممكن تقولك اذا كان مقفول كويس ولا لا	مافيش رمز هيقولك انك مافيش حاجة تعملها
	Saudi	مافيه إختبار يعلمك إن كان مقفل زين أو لا.	مافيه دليل يقول لك إنك ماتقدر تسويه.

https://doi.org/10.1371/journal.pone.0272991.t002

MultiNLI [25] English data sets. The M2M100 Many-to-Many multilingual model proposed by [34] has been used for automatic translation of SNLI and MultiNLI datasets to MSA. To build the SBERT-based model, the translated data have been used to convert the ArabicBERT model into an SBERT version. The second approach is based on interleaving English STS data with Arabic BERT models using transfer learning. In this approach, we started with an Arabic BERT-based model. This model has been fine-tuned to be converted to an SBERT model. This was done using English data from SNLI and MultiNLI English datasets and from original STS dataset. As will be seen in the Experimental Results section, this approach considerably impoved the accuracy of the model. The third approach is to utilize knowledge distillation-based STS models as a base and fine-tune the models using the proposed translated dataset to increase the accuracy of the models when used for Arabic STS. First, the pairs of sentences in the translated dataset have been inputted into siamese architecture which consists of two instances of a base model. Each instance produces an embedding using a pooling procedure. The two embeddings are compared and the manual estimated correlation scores are used to guide the network to fine-tune the model for being oriented to the dialect of the input data. Second, in the testing phase, each generated model has been verified using a similar architecture that takes the testing pairs of sentences as input and produces a cosine similarity value for each pair of sentences that can be compared with a manual given reference correlation score. Fig 3 summarizes the framework used in the third approach. The details of implemented experiments are explained in the following section.

Experimental results

The proposed models have been tested on a standard dataset for testing proposed by [8]. As mentioned before, this dataset has been translated to Egyptian and Saudi Arabic by native speakers of both dialects. Three groups of tests have been applied. In the first group, the accuracy of MSA models has been checked. While in the second and third groups, the generated models of Egyptian and Saudi Arabic have been tested. The accuracy measure that has been

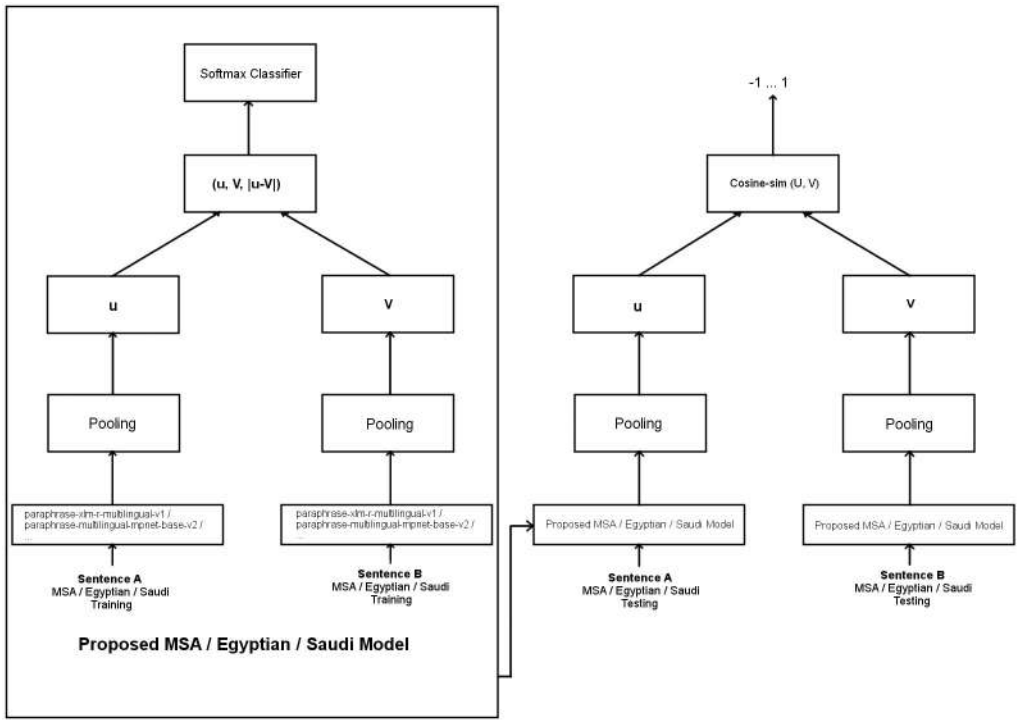


Fig 3. A framework of models generation using the third approach.

https://doi.org/10.1371/journal.pone.0272991.g003

used is the standard Spearman rank correlation between the cosine similarity of sentence representations and reference labels of testing datasets. The following is a brief description of the current state-of-the-art STS models and the base models that have been utilized in our experiments:

- **ArabicBERT**: ArabicBERT was the first pre-trained BERT model for Arabic. It is proposed by Safaya et al. in [35].
- **ARBERT**: proposed by Abdul-Mageed et al. in [36]. It is an Arabic large scale masked language model that targets modern standard Arabic.
- **stsb-xlm-r-multilingual**: It is a natural language processing model implemented in Transformer library. It was trained on SNLI + MultiNLI and on STS benchmark dataset. The model is a multilingual version, trained on parallel data for 50+ languages [26].
- **distiluse-base-multilingual-cased-v1**: A multilingual knowledge distilled version of multilingual Universal Sentence Encoder. Supports 15 languages including Arabic and English [26].
- **distiluse-base-multilingual-cased-v2**: It is a multilingual knowledge distilled version of multilingual Universal Sentence Encoder. While v1 model supports 15 languages, this version supports 50+ languages. However, performance on the 15 languages mentioned above are reported to be a bit lower [26].
- **quora-distilbert-multilingual**: It is the multilingual version of quora-distilbert-base, fine-tuned with parallel data for 50+ languages [26].
- **paraphrase-xlm-r-multilingual-v1**: A multilingual version of paraphrase-distilroberta-base-v1, trained on parallel data for 50+ languages [26].

Table 3. Accuracy of machine translation based and interleaved MSA models tested based on Spearman rank correlation between the cosine similarity of sentence representations and the reference labels of the testing dataset in [8].

Base Model	Training data	Score
ArabicBERT bert-base	SNLI and MultiNLI datasets translated using M2M100 model into MSA	0.4798
ArabicBERT bert-base	SNLI and MultiNLI English datasets	0.6525
ARBERT	SNLI and MultiNLI English datasets	0.708
ARBERT	SNLI and MultiNLI English datasets then STS for 1 epoch	**0.7364**

https://doi.org/10.1371/journal.pone.0272991.t003

- **paraphrase-multilingual-mpnet-base-v2**: It is the multilingual version of paraphrase-mpnet-base-v2, trained on parallel data for 50+ languages [26].

The following tables show the results of MSA, Egyptian Arabic, and Saudi Arabic experiments respectively. For each table, the base model, the training/fine-tuning data, and the accuracy measured in Spearman/cosine similarity are shown respectively.

Approaches evaluation

As can be seen in Table 3, in the first experiment, the first approach has been checked. The M2M100 model has been used to automatically translate SNLI and MultiNLI datasets to MSA. M2M100 is a Many-to-Many multilingual translation model proposed by Facebook that can translate directly between any pair of 100 languages. The translated data have been used to convert the ArabicBERT model into an SBERT model. As illustrated in Table 3, when the translated version of SNLI and MultiNLI has been used, the spearman score was around 0.48. But when the original English versions of SNLI and MultiNLI have been used to build the SBERT model, the spearman score was over 0.65. This means that the accuracy achieved using the original SNLI and MultiNLI English version is better than the accuracy we got using the translated version. This may be due to the inaccuracies in the translated version. So, It is not recommended to use automatic language translation-based solutions to tackle the STS problem; at least with the current maturity level of automatic translation.

To check the second approach, another experiment has been conducted. We have started with the ARBERT model, which is an Arabic BERT-based model, and fine-tuned it using English data to convert it into an SBERT model, In this direction, two trials have been tested, in the first trial, only SNLI and MultiNLI English datasets have been used for model conversion. while in the other trial, SNLI and MultiNLI datasets have been utilized and then a fine-tuning process has been applied using original STS data [8] for one epoch. The first trial provided a spearman score of around 0.70 while with the second trial, we got an accuracy of over 0.73. From these two trials, It can be seen that interleaving English data with Arabic-based models is more promising than the translation-based solution.

In the third approach, It has been checked how efficiently to use knowledge distillation-based solutions. For this purpose, several experiments have been conducted. As shown in Table 4, our translated 1.3k pairs of sentences have been used to fine-tune several state-of-the-art STS models. The best Spearman score achieved using this approach was over 0.81 using paraphrase-multilingual-mpnet-base-v2 proposed by [26] as a base model and our proposed translated dataset along with original data presented by [8]. Using a similar procedure, the proposed translated versions of STS data to Egyptian Arabic and Saudi Arabic have been used to fine-tune the state-of-the-art models. As illustrated in Table 5, In the case of Egyptian Arabic, the proposed translated data have been successfully used to fine-tune the base model paraphrase-xlm-r-multilingual-v1 proposed by [26] with a Spearman score of 0.775. In the case of

Table 4. Accuracy of knowledge distillation-based MSA models tested based on Spearman rank correlation between the cosine similarity of sentence representations and the reference labels of the testing dataset in [8].

Base Model	Fine-tuning data	Score
quora-distilbert-multilingual	translated 1.3K MSA pairs of sentences	0.7665
distiluse-base-multilingual-cased-v2	translated 1.3K MSA pairs of sentences	0.7752
distiluse-base-multilingual-cased-v1	translated 1.3K MSA pairs of sentences	0.7778
stsb-xlm-r-multilingual	translated 1.3K MSA pairs of sentences	0.7785
paraphrase-xlm-r-multilingual-v1	translated 1.3K MSA pairs of sentences	0.7918
paraphrase-xlm-r-multilingual-v1	translated 1.3K MSA pairs of sentences + original Arabic STS	0.7999
paraphrase-multilingual-mpnet-base-v2	translated 1.3K MSA pairs of sentences	0.8012
paraphrase-multilingual-mpnet-base-v2	translated 1.3K MSA pairs of sentences + original Arabic STS	**0.8103**

https://doi.org/10.1371/journal.pone.0272991.t004

Saudi Arabic. the proposed translated Saudi data along with the original data proposed by [8] have been utilized to fine-tune state-of-the-art base models. Table 6 provides the details of the experiments done in this direction. As shown in Table 6, the best Spearman score achieved was over 0.76 by fine-tuning the base model distiluse-base-multilingual-cased-v2.

Comparisons with state-of-the-art

To test the quality of the proposed models, they have been compared to state-of-the-art counterparts. While different methods have been assessed on various datasets at testing, our results can be compared to methods that used the MSA testing dataset suggested in [8]. Table 7 illustrates the comparisons with the best current MSA models.

As shown in Table 7, the proposed model for MSA enhanced the state-of-the-art result by around an absolute 2%. It is worth mentioning that transfer learning-based solutions depend on the similarity between the domain of the base model and the domain of the new model. While the base model (paraphrase-multilingual-mpnet-base-v2) of the proposed MSA model

Table 5. Accuracy of main Egyptian models tested based on Spearman rank correlation between the cosine similarity of sentence representations and the reference labels of the testing dataset in [8] after translation to Egyptian Arabic.

Base Model	Fine-tuning data	Score
paraphrase-multilg-mpnet-base-v2	translated 1.3K Egyptian pairs of sentences	0.7345
paraphrase-multilg-mpnet-base-v2	original Arabic STS then the translated 1.3K Egyptian pairs of sentences	0.763
paraphrase-xlm-r-multilingual-v1	original Arabic STS then the translated 1.3K Egyptian pairs of sentences	0.7647
paraphrase-xlm-r-multilingual-v1	translated 1.3K Egyptian pairs of sentences	**0.7751**

https://doi.org/10.1371/journal.pone.0272991.t005

Table 6. Accuracy of main Saudi Arabian models based on Spearman rank correlation between the cosine similarity of sentence representations and the reference labels of the testing dataset in [8] after translation to Saudi Arabic.

Base Model	Fine-tuning data	Score
paraphrase-xlm-r-multilingual-v1	translated 1.3K Saudi pairs of sentences	0.7441
paraphrase-xlm-r-multilingual-v1	original Arabic STS then the translated 1.3K Saudi pairs of sentences	0.752
distiluse-base-multilingual-cased-v2	translated 1.3K Saudi pairs of sentences	0.7608
distiluse-base-multilingual-cased-v2	original Arabic STS then the translated 1.3K Saudi pairs of sentences	**0.7622**

https://doi.org/10.1371/journal.pone.0272991.t006

Table 7. Comparisons between the proposed models and current state-of-the-art Arabic STS models based on Spearman rank correlation between the cosine similarity of sentence representations and the reference labels of the testing dataset in [8].

Variant	Model	Spearman/Cosine similarity
MSA	quora-distilbert-multilingual	0.7075
	distiluse-base-multilingual-cased-v1	0.7586
	distiluse-base-multilingual-cased-v2	0.7734
	stsb-xlm-r-multilingual	0.7867
	paraphrase-xlm-r-multilingual-v1	0.791
	paraphrase-multilingual-mpnet-base-v2	0.791
	proposed MSA model	**0.8103**
Egyptian	**Model**	**Spearman/Cosine similarity**
	quora-distilbert-multilingual	0.5811
	paraphrase-multilingual-mpnet-base-v2	0.6847
	distiluse-base-multilingual-cased-v2	0.6950
	stsb-xlm-r-multilingual	0.7200
	distiluse-base-multilingual-cased-v1	0.7237
	paraphrase-xlm-r-multilingual-v1	0.7516
	proposed Egyptian model	**0.7751**
Saudi	**Model**	**Spearman/Cosine similarity**
	quora-distilbert-multilingual	0.5706
	paraphrase-multilingual-mpnet-base-v2	0.6784
	stsb-xlm-r-multilingual	0.6879
	paraphrase-xlm-r-multilingual-v1	0.7145
	distiluse-base-multilingual-cased-v1	0.7310
	distiluse-base-multilingual-cased-v2	0.7410
	proposed Saudi model	**0.7622**

https://doi.org/10.1371/journal.pone.0272991.t007

has been trained on large scale amounts of data [26], the proposed new model has been fine-tuned using small dataset of only a few thousands of sentence pairs. This is promising because it indicates that the results can be even improved more without a need for new large scale datasets.

While there are no models in the literature that intentionally target Egyptian and Saudi Arabic, the state-of-the-art multilingual model that supports MSA Arabic provides a good result for the Egyptian Arabic variant. But the contributed model that targets the Egyptian Arabic boosts the result by 2.4% absolute enhancement. And the proposed Saudi-focused model also provided around 2% absolute gain. However, the gap between the accuracy achieved in MSA versus the Egyptian and Saudi dialects is still considerable. This is largely because the base model used has been trained on MSA data, while the Egyptian and Saudi variants didn't appear in the training data of their base models. To tackle this problem in the future, is it planned to automatically extract parallel data of high quality between MSA and Egyptian Arabic and between MSA and Saudi Arabic. And then using these data to boost the performance of Egyptian and Saudi models to match the level of MSA.

Discussions and conclusions

In this paper, the semantic textual similarity problem has been addressed with a focus on the Arabic language and two of the major Arabic dialectical variants: Egyptian and Saudi Arabic. The Arabic language is one of the low-resourced languages. This produces a considerable lag

of accuracy between semantic textual similarity models of Arabic and their counterparts in rich-resourced languages such as English. The suggested work has been presented to tackle this problem. The main contributions proposed in the paper can be summarized in the following: First, the problem of limited resources for Arabic STS has been addressed by three approaches. The first approach is to utilize automatic machine translation to translate English STS data to Arabic and to use the translated data for converting Arabic BERT models into STS Arabic models. The second approach is to interleave English STS data with Arabic BERT models to produce improved Arabic STS models. The third approach is based on utilizing knowledge distillation-based models as a base and fine-tuning them using a proposed translated dataset to improve the performance for Arabic STS. Also, we contributed a manual translation of a large subset from the STS competition dataset [8]. It has been translated to modern standard Arabic, Egyptian Arabic, and Saudi Arabic by professional translators. Moreover, the developed models that enhanced the accuracy for modern standard Arabic STS by around absolute 2% gain over the state-of-the-art level have been presented. The models have been tested on the standard dataset used by the community. Furthermore, the work presented the details and experiments of the developed STS models for Egyptian Arabic and Saudi Arabic, which achieved gains of around absolute 2.4% and 2% respectively.

Based on these results, the main conclusions to be considered are the following: Delivering high-quality data to the community is of special importance to improve the accuracy of STS models of low-resourced languages such as Arabic. Also, knowledge distillation based solutions are competitive to tackle the STS problem. Furthermore, the accuracy of Egyptian Arabic and Saudi Arabic STS models can be boosted considerably even with using relatively small proposed datasets.

Limitations and future work

Although the suggested work presents significant improvement for Arabic MSA STS, there are still some limitations to be considered. First, there is a large gap between the accuracy of MSA STS models when compared with the state-of-the-art of English STS. It is important to minimize this gap to support models integration in practical applications. Also, evaluation measures should consider the embedded semantic information included in sentences such as named entities. Our future work plan is to check the robustness of the developed Arabic STS models by evaluating them using different downstream tasks such as question answering and Quora Question Pairs problem [37]. Moreover, it is planned to expand our work by targeting new important Arabic dialects such as Maghribi and Levantine variants.

Supporting information

S1 Data.
(ZIP)

Author Contributions

Conceptualization: Abdullah M. Moussa, Sherif Abdou, Hebah Elgibreen.

Data curation: Abdullah M. Moussa, Sherif Abdou, Hebah Elgibreen, Mohammed Faisal.

Formal analysis: Abdullah M. Moussa, Sherif Abdou, Hebah Elgibreen, Mohammed Faisal.

Funding acquisition: Abdullah M. Moussa, Mohammed Faisal, Mohsen Rashwan.

Investigation: Abdullah M. Moussa, Hebah Elgibreen, Mohammed Faisal.

Methodology: Abdullah M. Moussa, Sherif Abdou.

Project administration: Mansour Al Sulaiman, Sherif Abdou, Mohsen Rashwan.

Resources: Mansour Al Sulaiman.

Software: Abdullah M. Moussa.

Supervision: Mansour Al Sulaiman, Sherif Abdou, Mohsen Rashwan.

Validation: Abdullah M. Moussa, Sherif Abdou, Hebah Elgibreen, Mohammed Faisal.

Visualization: Abdullah M. Moussa, Sherif Abdou.

Writing – original draft: Abdullah M. Moussa.

Writing – review & editing: Abdullah M. Moussa, Sherif Abdou, Hebah Elgibreen.

References

1. Chandrasekaran D.; Mago V. Evolution of Semantic Similarity—A Survey. ACM Computing Surveys (CSUR) 54.2. 2021, 1–37. https://doi.org/10.1145/3440755
2. Wang J.; Dong Y. Measurement of text similarity: a survey. Information. 2020 Sep; 11(9):421. https://doi.org/10.3390/info11090421
3. Boyce BR.; Boyce BR.; Meadow CT.; Kraft DH.; Kraft DH.; Meadow CT. Text information retrieval systems. Elsevier; 2017 Aug 16.
4. Mikolov T.; Chen K.; Corrado G.; Dean, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013 Jan 16.
5. Pennington J.; Socher R.; Manning, CD. Glove: Global vectors for word representation. InProceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 2014 Oct (pp. 1532-1543).
6. Devlin J.; Chang MW.; Lee K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.
7. Reimers N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084. 2019 Aug 27.
8. Cer D.; Diab M.; Agirre E.; Lopez-Gazpio I.; Specia, L. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. arXiv preprint arXiv:1708.00055. 2017 Jul 31.
9. Alian M.; Awajan A. Semantic similarity for english and arabic texts: a review. Journal of Information & Knowledge Management. 2020 Dec 2; 19(04):2050033. https://doi.org/10.1142/S0219649220500331
10. Al-Bataineh H.; Farhan W.; Mustafa A.; Seelawi H.; Al-Natsheh, HT. Deep contextualized pairwise semantic similarity for arabic language questions. In2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) 2019 Nov 4 (pp. 1586-1591). IEEE.
11. Abo-Elghit AH.; Al-Zoghby AM.; Hamza TT. Textual Similarity Measurement Approaches: A Survey (1). The Egyptian Journal of Language Engineering. 2020 Sep 15; 7(2):41–62.
12. Aljameel SS.; O'Shea JD.; Crockett KA.; Latham, A. Survey of string similarity approaches and the challenging faced by the Arabic language. In2016 11th International Conference on Computer Engineering & Systems (ICCES) 2016 Dec 20 (pp. 241-247). IEEE.
13. Farouk, M. Measuring sentences similarity: a survey. arXiv preprint arXiv:1910.03940. 2019 Oct 6.
14. Niwattanakul S.; Singthongchai J.; Naenudorn E.; Wanapu, S. Using of Jaccard coefficient for keywords similarity. InProceedings of the international multiconference of engineers and computer scientists 2013 Mar 13 (Vol. 1, No. 6, pp. 380-384).
15. Levenshtein, VI. Binary codes capable of correcting deletions, insertions, and reversals. InSoviet physics doklady 1966 Feb 1 (Vol. 10, No. 8, pp. 707-710).
16. Wang Z.; Mi H.; Ittycheriah, A. Sentence similarity learning by lexical decomposition and composition. arXiv preprint arXiv:1602.07019. 2016 Feb 23.
17. Lee MC.; Chang JW.; Hsieh TC. A grammar-based semantic similarity algorithm for natural language sentences. The Scientific World Journal. 2014 Apr 10; 2014. https://doi.org/10.1155/2014/437162 PMID: 24982952

18. Batanović V.; Bojić D. Using part-of-speech tags as deep-syntax indicators in determining short-text semantic similarity. Computer Science and Information Systems. 2015; 12(1):1–31. https://doi.org/10.2298/CSIS131127082B

19. Li Y.; Bandar Z.; McLean D.; O'shea, J. A Method for Measuring Sentence Similarity and its Application to Conversational Agents. InFLAIRS Conference 2004 May (pp. 820-825).

20. Kiros R.; Zhu Y.; Salakhutdinov RR.; Zemel R.; Urtasun R.; Torralba, A.; et al. Skip-thought vectors. In Advances in neural information processing systems 2015 (pp. 3294-3302).

21. Yang Y.; Yuan S.; Cer D.; Kong SY.; Constant N.; Pilar, P.; et al. Learning semantic textual similarity from conversations. arXiv preprint arXiv:1804.07754. 2018 Apr 20.

22. Cer D.; Yang Y.; Kong SY.; Hua N.; Limtiaco N.; John, RS.; et al. Universal sentence encoder. arXiv preprint arXiv:1803.11175. 2018 Mar 29.

23. Conneau A.; Kiela D.; Schwenk H.; Barrault L.; Bordes, A. Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364. 2017 May 5.

24. Bowman SR.; Angeli G.; Potts C.; Manning, CD. A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326. 2015 Aug 21.

25. Williams A.; Nangia N.; Bowman, SR. A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426. 2017 Apr 18.

26. Reimers N.; Gurevych, I. Making monolingual sentence embeddings multilingual using knowledge distillation. arXiv preprint arXiv:2004.09813. 2020 Apr 21.

27. Awatramani V.; Gupta, P. Natural Language Transfer Learning for Physiological Textual Similarity. In2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence) 2020 Jan 29 (pp. 733-737). IEEE.

28. Liu Y.; Ott M.; Goyal N.; Du J.; Joshi M.; Chen, D.; et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. 2019 Jul 26.

29. Mutinda FW.; Yada S.; Wakamiya S.; Aramaki E. Semantic Textual Similarity in Japanese Clinical Domain Texts Using BERT. Methods of Information in Medicine. 2021 Jun; 60(S 01):e56–64. https://doi.org/10.1055/s-0041-1731390 PMID: 34237783

30. Yang X.; He X.; Zhang H.; Ma Y.; Bian J.; Wu Y. Measurement of Semantic Textual Similarity in Clinical Texts: Comparison of Transformer-Based Models. JMIR medical informatics. 2020; 8(11):e19735. https://doi.org/10.2196/19735 PMID: 33226350

31. Yang Z.; Dai Z.; Yang Y.; Carbonell J.; Salakhutdinov RR.; Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems. 2019; 32.

32. Schwab, D. Semantic similarity of arabic sentences with word embeddings. In Third arabic natural language processing workshop 2017 Apr 3 (pp. 18-24).

33. Nagoudi EM.; Ferrero J.; Schwab D.; Cherroun, H. Word embedding-based approaches for measuring semantic similarity of arabic-english sentences. In International Conference on Arabic Language Processing 2017 Oct 11 (pp. 19-33). Springer, Cham.

34. Fan A.; Bhosale S.; Schwenk H.; Ma Z.; El-Kishky A.; Goyal S.; et al. Beyond english-centric multilingual machine translation. Journal of Machine Learning Research. 2021; 22(107):1–48.

35. Safaya A.; Abdullatif M.; Yuret, D. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. In Proceedings of the Fourteenth Workshop on Semantic Evaluation 2020 Dec (pp. 2054-2059).

36. Abdul-Mageed M.; Elmadany A.; Nagoudi, EM. ARBERT & MARBERT: deep bidirectional transformers for Arabic. arXiv preprint arXiv:2101.01785. 2020 Dec 27.

37. Chen Z, Zhang H, Zhang X, Zhao L. Quora question pairs. URL https://www.kaggle.com/c/quora-question-pairs. 2018.