دليل توسيم البيانات اللغوية

2021

إعداد

أستاذ دكتور عبدالمحسن بن عبيد الثبيتي

المركز الوطني لتحليل البيانات والذكاء الاصطناعي

رؤية VISION 2030
المملكة العربية السعودية
KINGDOM OF SAUDI ARABIA

مدينة الملك عبدالعزيز للعلوم والتقنية KACST

دليل توسيم البيانات اللغوية

2021

المحتويات

مستخلص

قد يكون من الصعب تصور وجود نظام كفوء لحوسبة اللغة أو معالجتها بدون وجود بيانات مرجعية موسّمة يدويا يتم من خلالها التأكد من دقة مخرجات النظام، وكذلك تدريبه لو كان النظام معتمدًا على خوارزميات تعلم الآلة. وتعاني اللغة العربية من شح واضح في البيانات المرجعية الموسمّة لأغراض دراسات اللسانيات الحاسوبية ومعالجة اللغة العربية وتطبيقاتهما، ولعل السبب الرئيسي وراء هذا الشح هو ما يتطلبه بناء مثل هذه البيانات المرجعية من جهد ووقت ومال. وبالرغم من أهمية وجود هذه البيانات فإن الأهم هو أن تكون هذه البيانات كافية ومتسقة التوسيم وذات جودة تحقق الهدف منها. ويحاول هذا الدليل أن يسلط الضوء على أهمية وجود هذه البيانات المرجعية الموسمة مع أمثلة متنوعة على أنواع التوسيم مع التركيز على ما يختص بالعربية. ولأن كل عملية توسيم آلية يمكن ردها لتكون عملية تصنيف آلي، يشرع الدليل في شرح مختصر لخطوات بناء نموذج حاسوبي للتوسيم/التصنيف يوضح كيف تتعامل كل خطوة من خطواته مع البيانات المرجعية الموسمة وكيف نستخدم هذا النظام في نهاية المطاف للتحقق من كفاية البيانات وجودتها للمشكلة المطروحة. ثم يوضح الدليل أهم الأسئلة الواجب الإجابة عنها قبل البدء بمشروع بناء بيانات مرجعية موسمة جديدة وماهي أهم العوامل المؤثرة في عملية بناء البيانات المرجعية الموسمة يدويا وكيفية التعامل معها لتكون عملية التوسيم اليدوي للبيانات المرجعية أقل جهدا ووقتا وكلفة. وأخيراً يوضح الدليل كيفية معالجة حقوق الملكية الفكرية حتى لا يقع فريق العمل تحت طائلة القانون لو حصل خرق لقانون حقوق الملكية الفكرية وحتى لا يخسر الفريق أيضا ما بذله من جهد أو وقت ومال.

1. تمهيد

تُعتبر المدونات اللغوية في صورتها الخام مصدرا مفيدًا وغنيًّا للدراسات اللغوية، ولبناء النماذج الحاسوبية للغة. وتزداد فائدة المدونات اللغوية وتكون النتائج المستخلصة منها أكثر دقة وموثوقية عندما تكون مزودة بمعلومات لغوية إضافية توضح -على سبيل المثال- الوظائف النحوية للكلمات المُكوِّنة للجمل في نصوص هذه المدونات.

لنفترض أننا نبحث في المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية [1] (Al-Thubaity 2015) - والتي سوف نسميها لاحقا بالمدونة العربية - عن كلمة «عمر». تشير بيانات المدونة العربية بأن كلمة «عمر» تكررت في المدونة العربية 561,698 مرة في 125,576 نصا، ولكن أي كلمة من هذه الكلمات التي تجاوزت النصف مليون هي اسم العَلَم «عُمَر»، أو المصدر «عُمْر»، أو الفعلين الماضيين المبنيين للمعلوم «عَمَّرَ» و«عَمُرَ»، أو الفعل الماضي المبني للمجهول «عُمِّرَ»، أو أيّ صورة أخرى يمكن أن تكون مستعملة في العربية لصورة (العين والميم والراء) «عمر» ولم يُذكر فيما سبق. هل يستطيع الباحث فحص جميع السياقات التي وردت فيها كلمة «عمر» لتحديد نوع الكلمة التي يريد دراستها من بين الاحتمالات الموضحة أعلاه؟ نعم يستطيع؛ ولكنه جهد مُضْنٍ وشاقٌّ جدا.

وتزداد حيرة الباحث وارتباكه، وتقل جودة نتائجه لو بحث عن التلازم اللفظي للفعل الماضي المبني للمجهول «عُمِّرَ»، فالنتائج ستكون منحازة بشكل جلي لصالح اسم العلم «عُمَر». ويمكنك تصور كلمات أخرى مثل «عمرها» و «طالب» و«كتب» و «نافع» وغيرها من الأمثلة.

يبدو أن الحل الأوضح هو إضافة التشكيل لجميع نصوص المدونة لدفع مثل هذا اللبس. فلو نظرنا للمثال السابق نجد أن التشكيل حل ناجع بالنسبة للأفعال، ولكن ماذا عن الأسماء التي يتغير تشكيل الحرف الأخير فيها بحسب موقعها من الجملة فهي إما مرفوعة أو منصوبة أو مجرورة. يبدو أن مثل هذا الحل قد يفاقم المشكلة.

ولكن هناك خيار آخر يمكن استخدامه مع التشكيل أو بدونه ألا وهو إضافة قسم الكلام الذي تنتمي له كل كلمة في المدونة مثل: كون الكلمة حرفا أو اسم علم أو صفة أو فعل ماض أو غير ذلك من أقسام الكلام. نسمي إضافة معلومات لغوية لكلمات المدونة مثل التشكيل أو أقسام الكلام أو المعلومات الصرفية وغيرها من معلومات (توسيما) أو (تحشية) أو (تصنيفا). وكل هذه المصطلحات الثلاث مستخدمة في الأدبيات بالمعنى نفسه. وستلاحظ استخدامي لها بالتناوب في هذا الدليل.

---

[1] https://corpus.kacst.edu.sa/

1. تمهيد

وعند رغبتنا بتوسيم بيانات لغوية قد يكون من الأفضل أن نقوم بتوسيمها يدويا عندما تكون البيانات المطلوبة قليلة – عشرون ألف كلمة مثلا - ولا يتوفر نظام آلي لتوسيمها. ولكن لو نظرنا للمدونة العربية على سبيل المثال المكونة من مئات الملايين من الكلمات، هل يمكن توسيمها يدويا بأقسام الكلام مثلا؟ الحل الأمثل لمثل هذه الحالة هو استخدام التوسيم الآلي لأقسام الكلام. وهذا التوسيم الآلي بدوره بحاجة إلى بيانات مُوسَّمة يدويا، وكافية لتدريبه على القيام بالتوسيم بأقسام الكلام بكفاءة عالية.

للأسف، تعاني أبحاث اللسانيات الحاسوبية العربية ومعالجتها - بشكل عام - من شح واضح في توفر البيانات اللغوية الموسَّمة يدويا سواء ما كان منها مجانيا أو بمقابل مادي. وما هو متوفر منها إمَّا أن به مشاكل تعيق الاستفادة منه بشكل واسع، أو أنَّ حجم هذه البيانات غير كافٍ للحصول على نماذج حاسوبية قليلة الأخطاء. وقد بحثت -أثناء عملي على هذا الدليل - عن بيانات عربية موسمة في ثلاثة مواقع معروفة ومهمة هي: محرك البحث المتخصص بالبيانات من شركة جوجل[2]، وموقع Kaggle[3] الشهير المختص بتحليل البيانات والذي يوفر أكثر من خمسين ألف مجموعة من البيانات المختلفة، وموقع مكتبات جامعة كارنيغي ميلون[4] الأمريكية الذي يوفر روابط متعددة لمصادر البيانات الخاصة بتعلم الآلة. وما وجدته هو إما بيانات غير موسمة، أو مدونات صغيرة موسمة آليا، أو نصوص من موقع التواصل الاجتماعي تويتر مصنفة كتغريدات إيجابية أو سلبية أو محايدة، أو مدونة صغيرة جدا موسمة بالكينونات الأسمية إضافة إلى البنك الشجري العربي (Maamouri et al. 2004) المعروف لدى المتخصصين في مجال اللسانيات الحاسوبية ومعالجة اللغة العربية على نطاق واسع الصادر من اتحاد البيانات اللغوية بجامعة بنسلفانيا [5].

فالمدونات الموسمة آليا – في المصادر التي ذكرتها أعلاه - تم توسيمها باستخدام نماذج حاسوبية مبنية أساسا على توسيمات البنك الشجري العربي لأقسام الكلام الذي وجهت له انتقادات متعددة لعدم موائمته لأقسام الكلام العربي (التميمي 2019)، أو لبعض بنى اللغة العربية (المجدوب وآخرون 2019)، أو لعدم اتساق تشكيل الكلمات فيه (Al-Thubaity et al. 2020). فالبنك الشجري العربي مبني على أساس نظرية مفادها أن جميع اللغات لها نفس أقسام الكلام والتركيب النحوي فلذلك كانت أقسام الكلام فيه متطابقة مع أقسام الكلام الخاصة بالإنجليزية. وعلى الرغم من فائدة هذا التطابق في الترجمة الآلية أو في الدراسات التقابلية إلا إنه ليس مناسبا لدراسة اللغة العربية بشكل دقيق.

وعلى الرغم من كل الانتقادات التي توجه للبنك الشجري العربي فلا زال وسيضل مصدرا مهما ومؤثرا لبناء النماذج الحاسوبية الخاصة بالعربية لعدة أسباب، أهمها: أنه صادر من جهة علمية مرموقة وهي اتحاد البيانات اللغوية من جامعة بنسلفانيا التي عملت جاهدة على صدوره بأفضل ما يمكن، ولأنه أيضا استُخدم لفترات طويلة، وفي أبحاث كثيرة متعلقة بالعربية ممَّا يستدعي استخدامه باستمرار عند مقارنة أداء النماذج الحاسوبية الجديدة بتلك القديمة. وما لم تتوفر بيانات جديدة وموثوقة تنافسه في مجاله فإنه سيضل مستخدما.

---

(2) https://datasetsearch.research.google.com/
(3) https://corpus.kacst.edu.sa/
(4) https://guides.library.cmu.edu/machine-learning/datasets
(5) https://www.ldc.upenn.edu/

ويضاف إلى ما سبق من ملاحظات، اقتصار البنك الشجري العربي على نوع محدد من النصوص وهو النصوص الصحفية، بل اقتصاره أيضا على ثلاثة مصادر صحفية فقط، وهي: وكالة الأنباء الفرنسية وصحيفة الحياة وصحيفة النهار. ولا شك أن محدودية تنوع موضوعات البنك الشجري تحد من استخدامه في توسيم نصوص ذات طابع مختلف عن النصوص الصحفية مثل النصوص الأدبية أو العلمية أو الفنية أو التراثية بشكل عام. وهذه مشكلة معروفة في مجالات تعلم الآلة حيث يقل أداء النموذج الحاسوبي بشكل كبير عند تطبيقه على نصوص ذات طبيعة تختلف عن النصوص التي تدرب عليها (Manning 2011).

إنَّ مثل هذه الملاحظات على البنك الشجري تستدعي التفكير في حاجتنا إلى بناء مدونة جديدة توسم يدويا وفقا لقواعد اللغة العربية المتعارف عليها، أو باعتماد تقسيم تمام حسان لأقسام الكلام (حسان 2004) على سبيل المثال، أو حتى بوسوم البنك الشجري العربي - لو اقتنعنا به- بحيث نَسِم بيانات جديدة من النصوص التراثية أو من النصوص العلمية مثلا لنستطيع دراسة هذه الاختلافات في اللغة بشكل دقيق.

وعلى الرغم من كل الانتقادات التي توجه للبنك الشجري العربي فلا زال وسيضل مصدرا مهما ومؤثرا لبناء النماذج الحاسوبية الخاصة بالعربية لعدة أسباب، أهمها: أنه صادر من جهة علمية مرموقة وهي اتحاد البيانات اللغوية من جامعة بنسلفانيا التي عملت جاهدة على صدوره بأفضل ما يمكن، ولأنه أيضا استُخدم لفترات طويلة، وفي أبحاث كثيرة متعلقة بالعربية ممَّا يستدعي استخدامه باستمرار عند مقارنة أداء النماذج الحاسوبية الجديدة بتلك القديمة. وما لم تتوفر بيانات جديدة وموثوقة تنافسه في مجاله فإنه سيضل مستخدما.

ويضاف إلى ما سبق من ملاحظات، اقتصار البنك الشجري العربي على نوع محدد من النصوص وهو النصوص الصحفية، بل اقتصاره أيضا على ثلاثة مصادر صحفية فقط، وهي: وكالة الأنباء الفرنسية وصحيفة الحياة وصحيفة النهار. ولا شك أن محدودية تنوع موضوعات البنك الشجري تحد من استخدامه في توسيم نصوص ذات طابع مختلف عن النصوص الصحفية مثل النصوص الأدبية أو العلمية أو الفنية أو التراثية بشكل عام. وهذه مشكلة معروفة في مجالات تعلم الآلة حيث يقل أداء النموذج الحاسوبي بشكل كبير عند تطبيقه على نصوص ذات طبيعة تختلف عن النصوص التي تدرب عليها (Manning 2011).

إنَّ مثل هذه الملاحظات على البنك الشجري تستدعي التفكير في حاجتنا إلى بناء مدونة جديدة توسم يدويا وفقا لقواعد اللغة العربية المتعارف عليها، أو باعتماد تقسيم تمام حسان لأقسام الكلام (حسان 2004) على سبيل المثال، أو حتى بوسوم البنك الشجري العربي - لو اقتنعنا به- بحيث نَسِم بيانات جديدة من النصوص التراثية أو من النصوص العلمية مثلا لنستطيع دراسة هذه الاختلافات في اللغة بشكل دقيق.

1. تمهيد

ولو نظرنا لبيانات تويتر الموسمة والمتوفرة حاليا نجد أنها تعاني من مشكلة واضحة وهي: أن اهتمامات مستخدمي تويتر تختلف باختلاف البلدان وباختلافات الأوقات. فالتغريدات التي جمعت من مستخدمي تويتر في المملكة العربية السعودية قبل سنتين مثلا لن تصلح - في الأغلب - للتوسيم الآلي لمجموعة تغريدات جمعت وقت كتابة هذا التقرير، ولا تصلح أيضا لتغريدات جمعت في نفس الوقت السابق من بلد آخر مثل الإمارات العربية المتحدة أو جمهورية مصر العربية لاختلاف اللهجات والاهتمامات بين المغردين حسب بلدانهم. والبيانات قليلة الحجم على الرغم من أهميتها في البحث العلمي - عندما لا يتوفر غيرها- لفحص النماذج الحاسوبية الخاصة بالتصنيف، إلا أنها في الأغلب لا تفي بإنتاج نموذج حاسوبي، ذو كفاءة عالية، نستطيع استخدامه في تطبيقات واقعية، خارج المجال البحثي.

وكملخص لما ذكرته آنفا، فتوسيم البيانات مهم جدا لدراسة اللغة وفهمها سواء من خلال دراسة المدونات اللغوية الموسمة، أو من خلال النماذج الحاسوبية للنظريات المختلفة التي يسعى مجال اللسانيات الحاسوبية لبنائها، أو حتى في حل مشاكل أخرى حياتية يسعى مجال معالجة اللغات الطبيعية لحلها أو لتسهيل حياتنا. كما أن اختلاف النظريات في ذات الموضوع، أو اختلاف طبيعة النصوص أو الزمن الذي صدرت فيه، أو اختلاف لهجات ومناطق المتكلمين باللغة في هذه النصوص، أو اختلاف الاحتياجات العملية باعث حثيث لتوفير بيانات موسومة يدويا باستمرار للمساهمة في بناء نماذج حاسوبية ذات كفاءة عالية تساعد في حل المشاكل التي تواجهنا في المجالات المختلفة.

ولكن توفير بيانات كافية وموسومة يدويا وذات جودة عالية عمل مضن يستهلك الكثير من الوقت والجهد والمال، وعلاوة على ذلك فهو عمل مُمِّل للأسف! ويضاف إلى ما سبق عزوف اللغويين والحاسوبيين عن التعاون مع بعضهم البعض. ولعل هذه الأسباب هي التي حَدَّت بشكل كبير من توفر بيانات عربية موسَّمة يدويا. وما يسعى إليه هذا الدليل هو إيضاح الخطوات اللازمة لبناء مدونة/بيانات موسومة يدويا ذات جودة عالية تقليلا للتكاليف والوقت والجهد، وجعل عملية التوسيم اليدوي ممتعة للموسِّمين أثناء العمل على توسيم البيانات- ما أمكن - من خلال تسهيل وتسريع عملية التوسيم بالنسبة لهم.

وفيما تبقى من هذا الدليل سوف أقوم أولا: بعرض نماذج من عمليات التوسيم ليكون لدى القارئ الكريم فكرة كافية عن أهمية التوسيم، والحاجة له من خلال هذه النماذج. ثم أشرح بشكل مختصر الخطوات اللازمة لبناء نموذج حاسوبي للتوسيم لأهمية توفر مثل هذا النموذج في التحقق من جودة البيانات الموسمة وكفايتها للحاجة التي بنيت من أجلها. ثم أشرح بعد ذلك العوامل المؤثرة في عملية التوسيم اليدوي للبيانات. ثم أناقش بشكل مختصر موضوع مهم يختص بالبيانات وهو موضوع حقوق الملكية الفكرية ثم أختم هذه المقدمة بخلاصة وأمنية تتعلق بإنشاء بيانات عربية موسمة بكل- أو أغلب - ما تحتاجه أبحاث اللسانيات الحاسوبية العربية ومعالجتها آليا.

2. نماذج من عمليات التوسيم

إنَّ توسيم المدونات اللغوية المخصصة لدراسة اللغة من أوجهها المختلفة مثال واضح لفائدة التوسيم، وأثره النافع على جودة الدراسات القائمة على المدونات، ويمكن للتوسيم - أيضا- أن يشمل أي بيانات لغوية تستخدم في اللسانيات الحاسوبية أو في معالجة اللغة الطبيعية. بل إن النماذج الحاسوبية التي تستخدم في توسيم المدونات اللغوية هي نتاج أساسي للسانيات الحاسوبية المَعنيَّة ببناء نماذج حاسوبية للنظريات اللغوية المختلفة.

إنَّ جميع عمليات التوسيم تؤول بشكل أو بآخر إلى أن تكون عملية تصنيف. والتصنيف مبحث أساسي في اللسانيات الحاسوبية ومعالجة اللغات الطبيعية وفي مجالات حاسوبية مختلفة مثل معالجة الصور باستخدام الحاسب. والهدف الأساسي من التصنيف هو أن يتم وضع عنصر ما ضمن مجموعة محددة سلفا من التصنيفات. ففي البيانات النصية قد يكون هذا الكائن نصا أو جملة أو كلمة أو حتى حرفا ضمن مجموعة من الأحرف المتصلة. وقد تكون التصنيفات التي نريدها لهذه الكائنات أن تنتمي إليها مُنطلِقة ومُؤسَّسة على نظرية لغوية مثل نظرية تمام حسان لأقسام الكلام، أو نظرية في علم النفس مثل نظرية إكمان (Ekman 1992) في أقسام المشاعر البشرية الشائعة في تحليل مواقع التواصل الاجتماعي، وغير ذلك من النظريات العلمية. أو قد تكون هذه التصنيفات مُؤسَّسة على حاجات عملية مثل تحديد نوع الطعام الذي يطلبه شخص ما من خلال محادث آلي لأحد شركات توصيل الطعام المنتشرة هذه الأيام. وفي هذه الحالة ما على الشخص سوى أن يكتب طلبه بالصيغة والطريقة التي يرغبها ويرسلها من خلال التطبيق ويقوم برنامج آلي بتحديد أنواع الطعام الذي طلبه وكميته ثم يقوم المطعم بإرسالها له.

وفيما يلي مجموعة من الأمثلة الأساسية لعمليات التصنيف أو التوسيم التي يمكن أن تُجرى على البيانات اللغوية مع بعض الأمثلة التوضيحية مع الإشارة إلى مرجع لهذه الأمثلة ما أمكن:

1.2 تصنيف النصوص

أن يكون لدينا نص مكون من مجموعة من الجمل، ثم نصنف هذا النص بحسب الوسوم أو التصنيفات المحددة سلفًا، مثل: تصنيف النصوص الصحفية إلى سياسية، أو اقتصادية أو اجتماعية أو ثقافية وخلاف ذلك من التصنيفات (Khorsheed & Al-Thubaity 2013). أو قد تكون المشكلة هي تحديد أو تصنيف نوع البريد الإلكتروني الذي نستقبله إلى بريد إلكتروني مزعج أو غير مزعج (Kumar & Sonowal 2020). أو مثل تصنيف التغريدات في تويتر إلى تغريدات إيجابية، أو سلبية أو محايدة أو دعائية (Al-Thubaity et al. 2018) أو تصنيف الأسئلة التي يتلقاها نظام الإجابة على الاستفسارات إلى أسئلة تتعلق بالوقت أو المواقع أو الأشخاص أو الأفكار المجردة (Soares & Parreiras 2020). وينتمي تحديد جنس الكاتب هل هو ذكر أو أنثى أو تحديد من هو كاتب مقال معين إلى هذا النوع من التصنيف (Hriez & Awajan 2020).

2. نماذج من عمليات التوسيم

2.2 التقطيع

ويكون على مستوى النص أو الجملة أو الكلمة.

أ. **التقطيع على مستوى النص:** عندما يكون التقطيع على مستوى النص فهو مَعنِي بشكل أساسي بتقطيع النص إلى جمل أو عبارات. في حالة التقطيع إلى جمل، قد تكون هذه المهمة سهلة ولا تحتاج لعملية تصنيف آلية عندما يلتزم كاتب النص بعلامات الترقيم. فالنقطة التي تلي الكلمة «.» وعلامة الاستفهام «؟» وعلامة التعجب «!» كلها علامات جلية لنهاية الجملة. ولكن الوضع يكون صعبا جدا ومؤثرا على دقة المحللات الصرفية أو المحللات التركيبية عند تعاملها مع نصوص غير ملتزمة بعلامات الترقيم.

وتؤول عملية التقطيع هذه إلى عملية تصنيف عندما نحدد الكلمات أو المواضع الدالة على نهاية الجملة. فعلى سبيل المثال نَسِمُ الكلمات التي تقع في بداية الجملة داخل النص بالوسم «ب» وفي نهاية الجملة بالوسم «ن» وما عداهما بالوسم «ل» فتصبح مهمة المقطع الآلي هنا تصنيف كلمات النص إلى هذه التصنيفات الثلاث. وينطبق هذا المبدأ على جميع عمليات التقطيع التي سوف نذكرها لاحقا.

ويندرج تحت هذا النوع من تقطيع النص تصنيف أجزاء النص إلى أصناف محددة سلفا مثل المقدمة والاستدلال والنتيجة. وفي هذا النوع من التصنيف يتم تصنيف الجمل الموجودة في النص - بعد تقطيعه إلى جمل- إلى واحد من الأقسام المحددة سلفا. ثم تضم هذه الجمل في مجموعات بحسب تسلسلها في النص (Eshkol et al. 2020).

ب. **التقطيع على مستوى الجملة**: يشمل التقطيع على مستوى الجملة تحديد العبارات الأسمية أو الفعلية أو شبه الجملة على سبيل المثال. ولهذا النوع من التوسيم استخدامات متعددة مثل التحليل التركيبي (Anderson & Vilares 2018) والترجمة الآلية (Fügen & Kolss 2007). ويعتمد مثل هذا النوع من التصنيف على تصنيف آخر يختص بالكلمات وهو التوسيم بأقسام الكلام وسنذكره لاحقا في تصنيف الكلمات.

ج. **التقطيع على مستوى الكلمة:** يمكن أن يتم هذا النوع من التقطيع على مراحل متتالية تكون مخرجات كل مرحلة منها أساسا للمرحلة التي تليها. فأول مراحله هو فصل المتصلات (Almuhareb et al. 2019). والمتصلات إما أن تكون حروفا تلصق في أول الكلمة عند كتابتها مثل الواو اللام والفاء والباء والكاف أو تلصق في نهاية الكلمة وهي الضمائر المتصلة مثل الهاء والكاف وألف الإثنين.

فعلى سبيل المثال، فإن لفظة «وكتابك» هي في الحقيقة ثلاث كلمات - لكل منها وسم يختص بها من أقسام الكلام كما أنها لا تغير من بنية الكلمة عند التصاقها بها - هي حرف العطف «و» والاسم «كتاب» والضمير المتصل «ك». وفصل المتصلات هنا مثال واضح على أن عملية التقطيع تؤول إلى عملية تصنيف حيث يتم وضع الحرف «ق» مثلا أمام كل حرف ينبغي فصله قبل الكلمة وأمام بداية كل ضمير متصل يأتي في آخر الكلمة (الضمير المتصل قد يكون حرفا واحدا مثل «ك» أو حرفين مثل «هم» أو ثلاثة أحرف مثل «كما»). وهذه المرحلة سابقة مهمة بالنسبة لعمليات الوسم بأقسام الكلام.

أما المرحلة الثانية التي يمكن أن تتم بعد فصل المتصلات فهي فصل الزوائد الصرفية التي إن دخلت على أول الكلمة أو آخرها غيرت بنيتها الصرفية مثل حروف كلمة «نأيت» التي تدخل على أول الفعل الماضي (كتب) فتحوله لفعل مضارع (يكتب) أو تاء التأنيث التي تلحق بالاسم المفرد المذكر (طالب) فتحوله إلى اسم مفرد مؤنث (طالبة) أو الواو والنون التي تلحق بالاسم المفرد المذكر (معلم) فتحوله إلى جمع مذكر سالم (معلمون).

ولكن هذه الزوائد الصرفية ليس لها قسم خاص من أقسام الكلام. وما ينتج من هذا النوع من التقطيع مفيد في عمليات البحث في المدونات بحيث تمكننا بشكل أفضل من دراسة معاني الكلمات واستخدامها بحيث تضم جميع الكلمات باختلاف بنائها الصرفي تحت مظلة أصل واحد (Freihat et al. 2018) أو في محركات البحث لتحسين نتائجها (Alnaied et al. 2020).

3.2 تصنيف الكلمات

يشمل هذا التوسيم أنواعا متعددة من أشهرها التوسيم بأقسام الكلام. وقد تكون هذه الوسوم مختصرة مثل الاسم والحرف والفعل وعلامات الترقيم والكلمات الأجنبية مثلا. أو قد يكون أكثر تفصيلا من ذلك بحيث يحتوي على أكثر من 30 وسما (الثبيتي وآخرون 2012، Zalmout & Habash 2020). والتوسيم بأقسام الكلام يجب أن تسبقه عملية أخرى هي فصل المتصلات المذكورة سابقا خلال الحديث عن التقطيع على مستوى الكلمة. والتوسيم بأقسام الكلام له فائدته في الدراسات اللغوية القائمة على المدونات، أو كسابقة لعمليات تصنيف أخرى مثل التحليل التركيبي (Leech 2013) وفي الترجمة الآلية (Mishra & Raj 2020) وغيره من المجالات.

ومن الأنواع الأخرى لتصنيف الكلمات توسيم الكينونات الأسمية والمعني بتحديد أنواع معينة من الأسماء مثل أسماء الأشخاص والأماكن والمنظمات على سبيل المثال بحيث يكون لكل نوع من هذه الكينونات الأسمية المرغوب تحديده تصنيف خاص بها وما عداها من كلمات داخل الجملة تأخذ تصنيفا واحدا مشتركا (Liu et al. 2019). وتوسيم الكينونات الاسمية مهم في مجالات متعددة مثل بناء أنظمة الإجابة الآلية على الاستفسارات (Kandasamy & Cherukuri 2020) وأنظمة المحادثة الآلية (Wu et al. 2020) واكتشاف العلاقات بين الأحداث ومسبباتها (Perera et al. 2020) وغير ذلك من التطبيقات. وهناك نوع آخر شبيه بتوسيم الكينونات الأسمية وهو توسيم المصطلحات العلمية (Ma et al. 2019) حيث يتم تحديد المصطلحات العلمية في النص لاستخدامها لاحقا في بناء المعاجم أو رفع كفاءة محركات البحث.

4.2 تصنيف الأحرف

يكون هذا النوع من التصنيف على مستوى الأحرف وأشهر أنواعه التشكيل. وفيه يتم وضع التشكيل لمناسب مثل الفتحة والكسرة والضمة أمام كل حرف في الجملة (Al-Thubaity et al. 2020). والتشكيل مهم جدا في تطبيقات وأنظمة متعددة مثل دفع اللبس (Alqahtani et al. 2019)، وفي تطبيقات تحويل النص إلى كلام (Ali et al. 2019)، وفي زيادة كفاءة التوسيم بأقسام الكلام (Kadim & Lazrek 2018). وهناك نوع آخر أتمنى أن ينال نصيبا من الاهتمام ألا وهو تحديد البنية الصرفية للكلمات بحيث يكون أمام كل حرف وسما يدل على أنه عين الكلمة أو فائها أو لامها وغير ذلك من الوسوم المناسبة.

2. نماذج من عمليات التوسيم

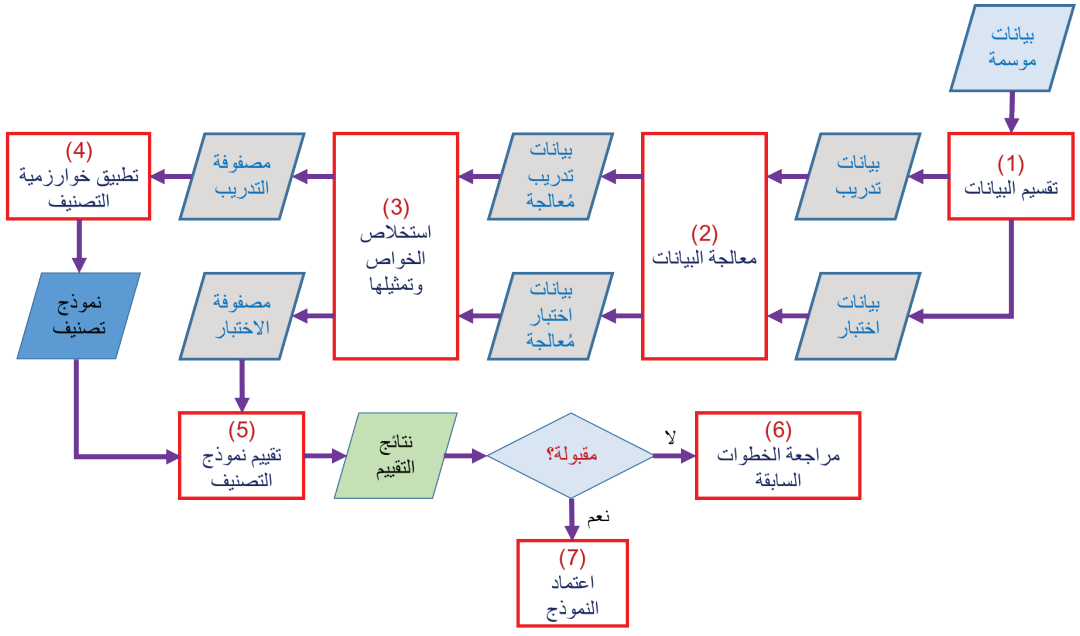
5.2 التوسيم غير المباشر

في الأمثلة أعلاه كان التوسيم مباشرا يغطي نطاقا واضحا (جملة، عبارة، كلمة، حرف) ولكن هناك عمليات توسيم غير مباشرة مثل استخراج إجابات أسئلة معينة من داخل النص وقد تكون إجابة أي سؤال موجودة في جملة بعينها أو في جمل متتابعة أو قد تكون الإجابة في جمل متفرقة داخل النص (Rajpurka et al. 2016). وتستخدم مثل هذه البيانات في أنظمة الإجابة عن الاستفسارات والأنظمة التعليمية والمحادث الآلي (Chen & Yih 2020). ومن الأمثلة الأخرى البيانات المستخدمة في تقييم أنظمة التلخيص الآلي. وتنقسم هذه البيانات إلى نوعين أولهما صياغة ملخص لنص أو لمجموعة من النصوص من متخصص بالموضوع وتكون هذه الصياغة بأسلوب هذا الخبير وكلماته. وقد تتعدد الصياغات والملخصات لنفس النص. وثانيهما - وهو الأسهل في الإعداد – أن يتم اختيار الجمل الأهم من النص التي تعبر عن الموضوع وتلخصه (Hailu et al. 2020). ويستخدم التلخيص الآلي في بناء أنظمة أخرى مثل متابعة وسائل الإعلام ومنصات التواصل الاجتماعي وتلخيص مخرجات أنظمة الإجابة عن الاستفسار (El-Kassas et al. 2020).

وكل عملية من عمليات التوسيم المذكورة أعلاه - وغيرها من عمليات التوسيم الممكنة - يمكن أن تستخدم منفردة لحل مشكلة ما، أو أن تتضافر لحل المشكلة نفسها، كما أن بعضها -كما رأينا- قد تساهم في زيادة كفاءة أنظمة توسيم أخرى.

3. خطوات بناء نموذج حاسوبي للتوسيم

يوضح الشكل 1 الخطوات الرئيسية لبناء نموذج حاسوبي لعملية التصنيف/التوسيم/التحشية. ويتطلب بناء النموذج توفر بيانات موسومة بادئ ذي بدء. وتجهيز هذه البيانات وتوفيرها هو ما يسعى إلى توضيحه هذا الدليل. ولكننا - قبل أن نخوض في هذا الموضوع - بحاجة إلى فهم الخطوات الأساسية التي يمر بها بناء أي نموذج حاسوبي خاص بالتصنيف لأن وجود مثل هذا النموذج حاسوبي المعتمد على البيانات الموسمة شرط لازمٌ للتحقق من جودة البيانات الموسمة قبل اطلاقها واستخدامها بشكل عملي. وكل خطوة من هذه الخطوات تؤثر بشكل كبير على نتائج الأداء الخاصة بنظام التصنيف. وفيما يلي شرح مبسط ومختصر لهذه الخطوات مع ذكر لبعض المراجع ما أمكن عند الحديث عن كل خطوة:



الشكل 1: الخطوات الرئيسية لبناء نموذج حاسوبي للتصنيف

1.3 تقسيم البيانات

عند بناء نموذج آلي للتصنيف (للتوسيم/للتحشية) نحتاج إلى أن نقسم البيانات الموسمة إلى قسمين رئيسين: بيانات التدريب وبيانات الاختبار. نستخدم بيانات التدريب لتعليم/تدريب خوارزمية التصنيف على تحديد أصناف الكائنات المعروضة عليها -مثل النصوص والجمل والكلمات - بحسب الأصناف المحددة سلفا. ونستخدم بيانات الاختبار لفحص واختبار نموذج التصنيف الحاسوبي الناتج بعد عملية التدريب. ويكون حجم بيانات التدريب في العادة بين 70% إلى 80% في المئة من حجم البيانات الموسمة ويكون المتبقي (30% إلى 20%) مخصصا للاختبار (Khorsheed & Al-Thubaity 2013).

3. خطوات بناء نموذج حاسوبي للتوسيم

ويكون اختيار النصوص أو الجمل المنضوية تحت كل من القسمين عشوائيا لمنع الانحياز بقدر الإمكان ولمنع شبهة التلاعب بالبيانات للحصول على نتائج معينة. ويجب أن تكون البيانات المستخدمة في كلا القسمين ثابتة لا تتغير مع كل تغير في العمليات التالية لتقسيم البيانات (معالجة البيانات واستخلاص الخواص وتمثيلها وتطبيق خوارزمية التصنيف) حتى يمكن المقارنة بين أداء الخوارزميات المستخدمة في التصنيف والاختيارات الخاصة بالمعالجة واستخلاص الخواص وتمثيلها بشكل صحيح ومنصف.

ويمكن تقسيم البيانات أيضا – وهو السائد- إلى ثلاثة أقسام هي بيانات التدريب وبيانات التحقق وبيانات الاختبار. حيث تستخدم بيانات التحقق لفحص أداء النموذج بعد كل دورة من دورات التدريب على نفس البيانات وعند استقرار الأداء يتم إيقاف التدريب ثم النظر في النتائج فإن كانت مناسبة تم اختبار النموذج باستخدام بيانات الاختبار وإلا تم تغيير ومعايرة العوامل المؤثرة في عمل خوارزمية التصنيف. فلو كنا نستخدم إحدى طرق التعلم العميق على سبيل المثل، فإن عددا من العوامل يمكن أن تؤثر على أداء النموذج مثل عدد طبقات الشبكات العصبية أو عدد الخلايا العصبية في كل طبقة أو معدل الفقد بعد كل دورة تدريب وغيرها من العوامل. وتشكل بيانات التدريب في العادة 60% من حجم البيانات الموسمة الكلي، بينما يكون حجم بيانات التحقق والاختبار 20% من حجم البيانات الموسمة لكل منهما (Al-Thubaity et al. 2020; Diab et al. 201).

ومن عيوب المنهج الثاني في تقسيم البيانات أنه يقلل من حجم البيانات المخصصة للتدريب مما قد يؤثر على أداء النموذج مما يجعلنا نلجأ إلى منهج ثالث -وقد يكون مفضلا - وهو أن نقسم البيانات إلى أقسام متساوية (في الأغلب تكون 10 أقسام) ثم نستخدم الأقسام من 1 إلى 9 للتدريب والقسم العاشر للاختبار. ثم نبدل بين الأقسام بحيث في كل مرة يكون هناك قسم للاختبار من الأقسام العشرة لم يستخدم من قبل ثم نأخذ متوسط نتائج الاختبار العشرة. وقد نلجأ لهذا المنهج عندما لا يكون هناك أقسام محددة ومتعارف عليها بالنسبة للتدريب والتحقق والاختبار (Madi & Al-Khalifa 2020).

وينطبق على المنهجين الأخيرين لتقسيم البيانات ما ينطبق على المنهج الأساسي من حيث عشوائية الجمل أو النصوص التي تكون في كل قسم لنفس الأسباب المذكورة أعلاه. ومن الأولى أن يكون تقسيم البيانات مع وقت إصدار البيانات وإتاحتها للاستخدام بحيث يكون هذا التقسيم مرجعيا ومستخدما دائما عند المقارنة بين الخوارزميات أو طرق بناء النظام. وبهذا تكون البيانات مرجعية وتقسيمها مرجعيا أيضا.

2.3 معالجة البيانات

في الأغلب لا تكون البيانات الموسمة في صورة تصلح معها للاستخدام المباشر بحسب رغبة الشخص القائم على بناء نظام التصنيف. ويمكننا تقسيم عمليات معالجة البيانات إلى قسمين رئيسيين. الأول يقوم بتغيير في البيانات تفرضه بيئة الاستخدام الفعلي لنموذج التوسيم. فعلى سبيل المثال لو كانت البيانات الموسمة ملتزمة بهمزة القطع أو بوضع النقط في الياء المتطرفة مثلا، وكان من يقوم ببناء نموذج التصنيف يرى أن الواقع – وهو في الحقيقة كذلك - وما سوف يواجه نظام التصنيف هو عدم الالتزام بهمزة القطع أو تنقيط الياء المتطرفة في الكتابة العربية، فإنه قد يرى من المناسب إزالة همزة القطع وتبديل الياء المتطرفة بألف مقصورة مثلا. ومن أشهر أنواع هذا القسم من المعالجة - إضافة إلى ما سبق ذكره – تغيير التاء المربوطة إلى هاء، وإزالة التشكيل، وإزالة الفراغات الزائدة، وفصل علامات الترقيم والأقواس عن الكلمات (Al-Thubaity et al. 2015).

ويندرج تحت هذا النوع حذف بعض بيانات التوسيم (ليس النص الأصلي) من البيانات الموسمة عندما تكون بيانات التوسيم متعددة - مثلما هو الحال في بيانات البنك الشجري العربي (أنظر شكل 2)- مثل حذف التحليلات الصرفية الخاطئة أو حذف المعلومات الخاصة بالتصريف أو جذر الكلمة من التحليل الصرفي الصحيح عند بناء نموذج خاص بالتوسيم بأقسام الكلام فقط.

وعملية معالجة البيانات ليست إلزامية ولكنها اختيارية متى ما رأي القائمون على بناء نظام التصنيف الحاجة لها. وما يتم في عملية معالجة البيانات سواء حذفا أو إضافة يجب توثيقه وتطبيقه على البيانات التي سوف يتعامل معها النظام عند تطبيقه فعليا على بيانات لم يتدرب عليها من قبل. 2.3 معالجة البيانات

```
INPUT STRING: لونغ
LOOK-UP WORD: lwng
     Comment:
       INDEX: P1W1
* SOLUTION 1: (luwnog) [luwnog_1] luwnog/NOUN_PROP
     (GLOSS): Long
  SOLUTION 2: (luwnog) [luwuwnog_1] luwnog/NOUN_PROP
     (GLOSS): Luong
  SOLUTION 3: (lwng) [DEFAULT] lwng/NOUN_PROP
     (GLOSS): NOT_IN_LEXICON
  SOLUTION 4: (liwng) [DEFAULT] li/PREP+wng/NOUN_PROP
     (GLOSS): for/to + NOT_IN_LEXICON
  SOLUTION 5: (lawng) [DEFAULT] la/EMPH_PART+wng/NOUN_PROP
     (GLOSS): indeed/truly + NOT_IN_LEXICON

INPUT STRING: بيتش
LOOK-UP WORD: byt$
     Comment:
       INDEX: P1W2
* SOLUTION 1: (biyt$) [biyt$_1] biyt$/NOUN_PROP
     (GLOSS): Beach
  SOLUTION 2: (byt$) [DEFAULT] byt$/NOUN_PROP
     (GLOSS): NOT_IN_LEXICON
  SOLUTION 3: (biyt$) [DEFAULT] bi/PREP+yt$/NOUN_PROP
     (GLOSS): by/with + NOT_IN_LEXICON

INPUT STRING: (
     Comment:
       INDEX: P1W3
* SOLUTION 1: () [DEFAULT] (/PUNC

INPUT STRING: الولايات
LOOK-UP WORD: AlwlAyAt
     Comment:
       INDEX: P1W4
* SOLUTION 1: (AlwilAyAtu) [wilAyap_1] Al/DET+wilAy/NOUN+At/NSUFF_FEM_PL+u/CASE_DEF_NOM
     (GLOSS): the + states/provinces + [fem.pl.] + [def.nom.]
  SOLUTION 2: (AlwilAyAti) [wilAyap_1] Al/DET+wilAy/NOUN+At/NSUFF_FEM_PL+i/CASE_DEF_ACC
     (GLOSS): the + states/provinces + [fem.pl.] + [def.acc.]
```

الشكل 2: نموذج من بيانات البنك الشجري العربي بعد توسيمها. علامة * تدل على التحليل الصرفي الصحيح. كما نلاحظ أن كل تحليل صرفي يحتوي على نقحرة الكلمة وعلى ترتيبها في النص وعلى ترجمة الكلمة للإنجليزية. الصورة أعلاه مأخوذة من أحد نصوص وكالة الأنباء الفرنسية (الملف AFP_ARB.0001_20000715).

3. خطوات بناء نموذج حاسوبي للتوسيم

3.3 استخلاص الخواص وتمثيلها

نحتاج في هذه الخطوة الإجابة على السؤال التالي: هل تكفي البيانات بتوسيمها الحالي لبناء نموذج تصنيف كفوء؟ قد تكون الإجابة المباشرة غير دقيقة بدون تجريب خوارزمية التصنيف على البيانات المتاحة. إن كانت الإجابة بلا فقد نحتاج إما إلى إضافة معلومات إضافية أو حذف وتغيير في البيانات الأساسية. فبالنسبة للحالة الأولى فإنه يتم إضافة معلومات ليست ضمن بيانات التوسيم الأصلية لتعزيز عملية التدرب على التصنيف. وهذه المعلومات المضافة تكون في العادة نتائج لتوسيم آلي بعكس البيانات الأصلية التي كان توسيمها وتحقيقها يدويا. مثل إضافة وسوم أقسام الكلام عند التعرف على الكينونات الأسمية حيث يكون للكلمة وسمين الأول الوسم الخاص بالكينونات الأسمية والثاني خاص بقسم الكلام الذي تنتمي له الكلمة (Alonso et al. 2021). أو قد يكون الحال إضافة التشكيل الكامل للكلمة عند بناء نظام للتوسيم بأقسام الكلام (Kadim & Lazrek 2018). ويمكن إضافة إي معلومات لغوية تختص بالكلمة في النص الأصلي لكي تساعد في جودة عملية التصنيف. كما يمكن إضافة معلومات غير لغوية مثل موقع الكلمة في الجملة أو تكرارها في النص أو ماهي الكلمات التي تسبقها والتي تليها.

وفي الحالة الثانية يتم حذف بعض مكونات النص الأصلي مثلما هو الحال في عمليات تصنيف النصوص أو تحليل الآراء. حيث يتم حذف الكلمات التي ليس لها تأثير على دقة عملية التصنيف والتي يؤدي وجودها إلى بطء تدريب الخوارزمية وتعقيد نموذج التصنيف مثل الكلمات الوظيفية (أحرف الجر، النسخ، النصب، أسماء الإشارة والضمائر وغيرها من الأدوات) والكلمات قليلة التكرار في بيانات التدريب. وقد يتطلب الحال أيضا استبدال الكلمات الأصلية في النص بجذورها أو جذوعها (Khorsheed & Al-Thubaity 2013).

فمثلا عند عرض الجملة المراد تصنيفها أو تصنيف مكوناتها على خوارزمية التصنيف، تمرر الكلمات واحدة تلو الأخرى بحسب ترتيبها في الجملة على نموذج التدريب ويكون تمثيلها في هذه الحالة هو الكلمة والمعلومات اللغوية وغير اللغوية المتعلقة بالكلمة ثم الوسم الصحيح لهذا التمثيل. وفي حالة تصنيف النصوص قد يكون التمثيل قيمة عددية لكل كلمة في النص اعتمادا على تكرارها في النص. ومع ظهور خوارزميات التعلم العميق، ظهر ما يسمى «تضمين الكلمات» الذي يضع لكل كلمة تمثيلا عدديا عبارة عن متجه ثابت الطول لجميع الكلمات بحيث تتقارب قيم مكونات هذا المتجه للكلمات ذات الدلالات المتشابهة. وهذا لتمثيل إما أن يكون مبنيا سابقا على بيانات كبيرة مثل « تمثيلات التشفير ثنائي الاتجاه من المحولات» BERT (Devlin et al. 2018) أو مبنيا على بيانات التدريب. وقد أثبت التمثيل باستخدام تضمين الكلمات أو الأحرف تفوقه على الطرق التقليدية في تمثيل البيانات (Al-Smadi et al. 2019).

4.3 تطبيق خوارزمية التصنيف

بشكل عام ومبسط فإن هدف خوارزميات تعلم التصنيف هو أن تنتج نموذجا يمكن استخدامه لاحقا في تصنيف بيانات لم تتدرب عليها الخوارزمية من قبل. ويمكن تمثيل هذا النموذج بدالة رياضية تحول المدخلات (متغيرات الدالة) إلى قيمة يمكن إرجاعها أو تقريبها إلى أحد الأصناف التي تدربت عليها الخوارزمية.

وعلى الرغم من معرفتنا للمتغيرات فإن هناك عوامل أخرى في هذه المعادلة - يتم بناؤها خلال عملية التدريب - تؤثر في عملية التحويل من الصعب علينا في كثير من الأحيان معرفتها فهي قد تصل إلى مئات الآلاف بل الملايين في بعض الأحيان. وهذا ما يجعلنا عاجزين في أحيان كثيرة عن تفسير وصول النموذج لنتيجة معينة. وعادة ما يتم إيقاف عملية تدريب خوارزمية التصنيف عند وصول النموذج الناتج إلى دقة معينة باستخدام بيانات التدريب أو بعد عدد معين سلفا من مرات مشاهدة خوارزمية التصنيف لبيانات التدريب.

وهناك العديد من خوارزميات التصنيف التي تتفاوت في دقتها وقابليتها للتعامل مع بيانات كثيرة خلال التدريب وفي صلاحيتها لبعض البيانات دون غيرها. ومن أشهر هذه الخوارزميات المتعلقة بالنصوص: آلات المتجهات الداعمة (Support Vector Machines) وبايز البسيط (Naive Bayes) وشجرة القرار (Decision Tree) وأقرب الجيران (K-Nearest Neighbors) والشبكات العصبية التكرارية (Recurrent Neural Networks) في صورها المختلفة.

5.3 تقييم نموذج التصنيف

بعد الانتهاء من تطبيق خوارزمية التصنيف ينتج لدينا نموذج حاسوبي للتصنيف يمكن تطبيقه على بيانات جديدة بحيث تكون هذه البيانات هي المدخلات للنموذج ومن ثم يقوم النموذج بتوقع التصنيف المناسب لهذه المدخلات. ولكن قبل تطبيقه على بيانات جديدة في الواقع، فإننا نقوم بتقييم هذا النموذج باستخدام بيانات الاختبار (لا تنس أن بيانات الاختبار هي جزء من البيانات الموسمة يدويا مسبقا).

نستخدم في تقييم نموذج التصنيف أربعة مقاييس هي الصحة (Accuracy) وسوف نرمز لها بالرمز Acc، والدقة (Precision) وسوف نرمز لها بالرمز P، والاستدعاء (Recall) وسوف نرمز له بالرمز R، ومقياس ف1 (F1 score) وسوف نرمز له بالرمز F1. ونستطيع تعريف هذه المقاييس على النحو التالي:

- **الصحة:** نسبة التصنيفات الصحيحة من كامل مجموعة الاختبار.
- **الدقة:** تحتسب لكل تصنيف على حدة. فلو فرضنا أننا نحسب الدقة للصنف «أ» فتكون الدقة هي نسبة التصنيفات الصحيحة للصنف «أ» من مجموع بيانات الاختبار التي وضعها نموذج التصنيف ضمن التصنيف «أ».
- **الاستدعاء:** تحتسب لكل تصنيف على حدة. فلو فرضنا أننا نحسب الاستدعاء للصنف «أ» فيكون الاستدعاء هو نسبة التصنيفات الصحيحة للصنف «أ» من مجموع بيانات الاختبار التي تنتمي للصنف «أ» في بيانات الاختبار.
- **مقياس ف1:** هو مقياس يجمع بين الدقة والاستدعاء باستخدام المتوسط التوافقي بينهما (harmonic mean).

وحساب المقاييس المذكورة أعلاه يتطلب إيجاد القيم التالية لكل تصنيف. فلو افترضنا أننا نحسب هذه المقاييس للتصنيف «أ»، فإننا نوجد القيم التالية:

أ. **التوقع الإيجابي الصحيح (True Positive):** يشير إلى عدد المرات التي استطاع نموذج التصنيف أن يحدد أن البيانات المدخلة تنتمي للتصنيف «أ» بصورة صحيحة. وسوف نرمز له بالرمز «TP».

3. خطوات بناء نموذج حاسوبي للتوسيم

ب. **التوقع السلبي الصحيح (True Negative):** يشير إلى عدد المرات التي استطاع نموذج التصنيف أن يحدد أن البيانات المدخلة التي لا تنتمي للتصنيف «أ» أنها لا تنتمي إليه. وسوف نرمز له بالرمز «TN».

ج. **التوقع الإيجابي الخاطئ (False Positive):** يشير إلى عدد المرات التي لم يستطع نموذج التصنيف تحديد أن البيانات المدخلة تنتمي للتصنيف «أ» وهي في الحقيقة تنتمي للتصنيفات الأخرى. وسوف نرمز له بالرمز «FP».

د. **التوقع السلبي الخاطئ (False Negative):** يشير إلى عدد المرات التي لم يستطع نموذج التصنيف أن يحدد أن البيانات المدخلة تنتمي للتصنيف «أ» وتوقع أنها تنتمي للتصنيفات الأخرى. وسوف نرمز له بالرمز «FN».

ولتسهيل حساب هذه القيم ننشئ ما يسمى بمصفوفة الدقة (confusion matrix). لنفترض أن لدينا نموذج للتصنيف يقوم بتصنيف البيانات إلى ثلاثة تصنيفات «أ»، و «ب»، و «ج». وكان حجم بيانات الاختبار 45 عينة، 12 عينة منها تنتمي للصنف «أ»، و16 تنتمي للصنف «ب»، و 17 تنتمي للصنف «ج». وكانت نتائج التصنيف كما تمثلها مصفوفة الدقة الموضحة في الجدول 1 أدناه:

جدول 1: مصفوفة الدقة الافتراضية لبيانات الاختبار.

		التصنيفات الصحيحة		
		أ	ب	ج
النموذج	أ	7	8	9
	ب	1	2	3
	ج	4	6	5

فلو أخذنا الصنف «أ» على سبيل المثال فإن قيمة التوقع الإيجابي الصحيح (TP) تساوي 7، وقيمة التوقع السلبي الصحيح (TN) تساوي 16(2+3+6+5)، وقيمة التوقع الإيجابي الخاطئ (FP) تسوي 17 (8+9)، وقيمة التوقع السلبي الخاطئ (FN) تسوي 5 (1+4). وبنفس الطريقة يتم حساب قيم التوقعات بالنسبة للصنفين الآخرين. الجدول 2 يوضح قيم التوقعات للتصنيفات بحسب مصفوفة الدقة الموضحة في الجدول 1.

الجدول 2: قيم التوقعات حسب مصفوفة الدقة الموضحة في الجدول 1.

	التوقعات			
الصنف	TP	TN	FP	FN
أ	7	5+6+3+2	9 +8	4 +1
ب	2	5+4+9+7	3 +1	6 +8
ج	5	2+1+8+7	6 +4	3 +9

وبالتالي فإنه يمكننا حساب الصحة من خلال جمع قيم التوقع الإيجابي الصحيح لكل تصنيف (7+2+5) ثم قسمة هذا الجمع على العدد الكلي للعينات (45) ليكون الناتج (0.31). وهذه القيمة للصحة متدنية جدا ولا يمكن القبول بها ولكنها في هذا السياق أتت كمثال مصطنع فقط.

أما مقاييس الدقة (P) والاستدعاء (R) ومقياس ف1 (F1) فإنها تحسب لكل تصنيف على حدة كالتالي:

$$P = \frac{TP}{TP+FP}$$

$$R = \frac{TP}{TP+FN}$$

$$F1 = \frac{TP}{2TP+FP+FN}$$

ولتقييم أداء نموذج التصنيف بشكل عام نحسب المتوسط الحسابي لكل قيمة من قيم الدقة والاستدعاء ومقياس ف1. فكلما اقتربت قيم هذه المقاييس من الواحد الصحيح دل هذا على دقة نظام التصنيف. وكلما كانت القيم قريبة من النصف أو أقل منه دل هذا على أن عملية التصنيف تكاد تكون عشوائية.

6.3 مراجعة الخطوات السابقة

إن مراجعة خطوات بناء النموذج الحاسوبي الخاص بالتصنيف تعتمد بشكل رئيس على نتائج تقييم نموذج التصنيف وخصوصا عندما تكون نتائج التقييم الناتجة من خطوة تقييم النموذج غير مرضية أو عندما يكون أداء النموذج بعد تطبيقه في الواقع غير مرض (حتى لوكان أداءه على بيانات الاختبار مقبولا) . في هذه الخطوة يتم النظر أولا في مقاييس الأداء الخاصة بالنموذج بشكل عام فلو كانت أقل من المأمول فقد يكون المتسبب في ذلك أحد العوامل التالية:

أ. **البيانات:** فهي إما أن تكون غير كافية للتدريب في مجملها أو لا يوجد عينات تغطي بعض التصنيفات بشكل كاف. أو أن توسيم البيانات غير متسق وبه أخطاء. ووجود هذه المشاكل في البيانات هو أصعب المشاكل حلا لما يتطلبه من جهد ووقت وتكلفة لذلك يجب الحرص على توسيم بيانات كافية منذ البداية وتدقيق البيانات خلال عمليات التوسيم. وقد يكون السبب أن خوارزمية التصنيف تحتاج بطبيعتها إلى بيانات أكبر حجما. وقد يكون الحل في هذه الحالة هو تجربة خوارزميات أخرى للتصنيف والنظر في أدائها.

ب. **تقسيم البيانات:** يمكن أن يتأثر نموذج التصنيف بخطوة تقسيم البيانات حينما لا يكون التقسيم عشوائيا بحيث ينحاز التقسيم إلى وضع بيانات أصناف بعينها في بيانات التدريب ولا يكون هناك تمثيل كاف لبقية الأصناف. وهذا بدوره يؤدي إلى تعثر نموذج التصنيف في التعرف على هذه الأصناف قليلة التمثيل في بيانات التدريب كما إنه قد يؤثر أيضا في قدرة النموذج على تصنيف الأصناف الأخرى وإن كان بصورة أقل. وحل هذه المعضلة يسير جدا باستخدام الطرق الآلية التي تقسم البيانات عشوائيا كما ذكرنا في خطوة تقسيم البيانات آنفا.

3. خطوات بناء نموذج حاسوبي للتوسيم

ج. **معالجة البيانات:** قد تكون الخطوات التي أجريناها في خطوة معالجة البيانات غير كافية أو أنها تسببت في حذف بعض المعلومات الضرورية لتعلم خوارزمية التصنيف. أو قد تكون خاطئة ولم تتم بالشكل الصحيح. وهنا يجب التأكد أن نفس خطوات معالجة البيانات التي تمت على بيانات التدريب قد تمت على بيانات الاختبار. وتجاوز هذه المشكلة متيسر إن حصلت بسبب أنها تتم بصورة آلية في العادة.

د. **استخلاص الخواص وتمثيلها:** هذه الخطوة من أهم الخطوات المؤثرة على أداء نموذج التصنيف. فالخواص المختارة لعرض بيانات الكائنات المطلوب تصنيفها على نموذج التصنيف قد تكون غير كافية فنحتاج إلى إضافة خواص أخرى. أو أن المعلومات التي أضفناها للبيانات - مثل أقسام الكلام - قد تمت بصورة آلية ويشوبها الكثير من الأخطاء. وهنا نحتاج إلى جهد أكبر في البحث عن السبب واختيار خواص أخرى أو إضافية. وقد تكون الخواص كافية ولكن طريقة تمثيلها غير مناسبة أو غير كافية للخوارزمية المستخدمة.

ه. **خوارزمية التصنيف:** يؤثر اختيار خوارزمية التصنيف لتأدية مهمة معينة بشكل كبير على أداء نموذج التصنيف الناتج عنها. فبعض خوارزميات التصنيف تنجح في أداء مهام تصنيف معينة دون غيرها من المهام. فلذلك فإن البحث والتأكد من أن خوارزمية التصنيف المختارة مناسبة منذ البداية للمهمة وللبيانات أيضا. ولكن حتى لو كانت الخوارزمية مناسبة لمهمة التصنيف المعطاة فإن عملية التعلم الخاصة بهذه العملية تحتاج إلى معايرة المتغيرات الخاصة بالتعلم واختيار المناسب منها لهذه العملية. ففي خوارزميات التعلم العميق قد نحتاج إلى إضافة طبقة إضافية من الشبكات العصبية أو اختيار معدل أقل أو أكبر للتعلم. وفي آلات المتجهات الداعمة قد يكون استخدام الدالة الخطية أفضل أداء لعملية التصنيف التي نعمل عليها من استخدامنا لمعادلة من الدرجة الثانية. والنظر في مثل هذه الأمور وحل المشاكل المتعلقة بها تخصصٌ حاسوبي بشكل خاص وهو متيسر بسبب توافر المكتبات الجاهزة لهذه الخوارزميات وسهولة تغيير المتغيرات الخاصة بها.

7.3 اعتماد النموذج

بعد أن يتم التأكد من أن نموذج التصنيف يحقق الأداء المطلوب يتم اعتماد النموذج وتطبيقه على بيانات غير مصنفة من قبل. فلو تم اعتماد نموذج للتوسيم بأقسام الكلام مثلا فإنه يمكن تطبيقه على نصوص المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية فتنتج مدونة موسمة بأقسام الكلام في وقت قصير وبكلفة قليلة جدا وتزداد بذلك الفائدة من المدونة. ولكن الرحلة لا تنتهي هنا فقد نجد بعد تطبيق النموذج على بيانات مختلفة أخطاء في التصنيف تقلل من فائدة استخدام النموذج فعندها يجب أن ننظر في بيانات التدريب فقد تكون بحاجة لإضافة بيانات جديدة أو أن ننظر في خطوات بناء نموذج التصنيف مرة أخرى. فعملية التحسين يجب أن تستمر والا تتوقف.

4. توسيم البيانات يدويا

يجب النظر إلى بناء مدونة موسومة يدويا على أنه مشروع مهم يتطلب وقتا وجهدا وتكلفة - غير متيسرة في بعض الأحيان لكل الراغبين بهذا المشروع - وأنه مشروع ممتد لفترة طويلة لا ينتهي حال تجهيز البيانات بل يمتد إلى ما بعد إطلاق هذه البيانات الموسمة يدويا وتطبيقها في عدة مجالات. فلكل مجال من هذه المجالات تحدياته وما ينتج عنه من ملاحظات قد تستدعي إعادة النظر في البيانات وفي جودتها وتصحيح ما هو بحاجة للتصحيح ثم إعادة إطلاق نسخة جديدة منها.

إن التهاون وعدم الحرص الكافي في التخطيط والتنفيذ لهذا المشروع مرده خسارة الوقت والجهد والمال وبيانات غير ذات قيمة. وسوف يركز هذا الجزء من الدليل على أهم العوامل المؤثرة في بناء البيانات الموسمة يدويا وهي: البيانات الخام، والوسوم المستخدمة، والأسلوب المستخدم في التوسيم، والقائمون بعملية التوسيم.

ويجب التنبيه – في هذا المقام - على أهمية التوثيق لكل ما يتم خلال رحلة بناء البيانات الموسمة وما يتعلق بالملاحظات عليها بعد استخدامها وما يتعلق بتصحيحها. فلو تم التعديل على قائمة الوسوم يجب توثيق هذا التغيير وأن تكون هناك نسخة جديدة من الوسوم برقم تسلسلي يوضح مراحل التغيير. مع الحرص دائما على حفظ نسخ احتياطية بصورة دورية لكل ما يتم العمل عليه سواء للبيانات أو للوثائق. فالمشاكل التقنية التي قد تفقدنا البيانات محتملة دائما والعاملون في المشروع أو القائمين عليه قد يتغيرون أو تخونهم الذاكرة بخصوص ما يتعلق بالمشروع بعد فترة من الزمن.

ولكن قبل أن نبدأ ببناء بيانات موسمة يدويا نحتاج الإجابة على السؤال التالي:

هل نحن بحاجة لبناء مدونة موسومة لبناء نموذج حاسوبي للتصنيف؟

تكون الإجابة بنعم في الحالات التالية:

أ. أن نظام التوسيم الذي نريد بنائه سوف يخدم غرضا لا يوجد له بيانات موسمة. على سبيل المثال -وبحد علمي - فإنه لا يوجد بيانات موسمة للأوزان الصرفية أو لأوزان بحور الشعر أو لجذور الكلمات. حينها من الضروري بناء مثل هذه البيانات لو كان الحال يتطلب ذلك.

ب. لو كانت هناك بيانات موسومة ولكن الوسوم لا تنطلق من النظرية اللغوية التي نريد دراستها. فعلى سبيل المثال تتوفر بيانات البنك الشجري العربي ولكنها لا تخدم النظرية المستقرة في النحو العربي. أو حتى لو أردنا دراسة نظرية تمام حسان في أقسام الكلام على سبيل المثال. مثل هذه الحالات تستدعي بناء بيانات موسومة بحسب النظرية التي ننطلق منها.

ج. لو أننا جربنا بناء نموذج جديد على بيانات موسمة تخدم غرضنا وتتوافق مع النظرية التي ننطلق منها ولكن عندما طبقنا النموذج على البيانات التي نرغب في دراستها كان أداء النموذج غير مرض بالنسبة لنا إما بسبب اختلاف طبيعة البيانات التي ندرسها عن البيانات التي تدرب عليها النموذج أو بسبب قلة بيانات التدريب وعدم كفايتها لإنتاج نموذج تصنيف جيد. في هذه الحالة يمكن توسيم بيانات جديدة تضاف للبيانات القديمة الموسمة.

د. لو كان هناك بيانات تحقق الغرض وتنطلق من نفس النظرية التي نعمل عليها ولكن يتعذر حصولنا عليها لأي سبب، أو لأننا نريد أن نبني نماذجا ونستخدمه في أغراض تجارية ولا يمكن أن نستفيد من البيانات المتوفرة لأغراض البحث أو لأننا جهة بحثية من مهامنا أن نوفر بيانات موسمة للباحثين مفتوحة المصدر لتثري البحث في مجال حوسبة اللغة العربية ومعالجتها مثلا.

4. توسيم البيانات يدويا

وبطبيعة الحال ستكون الإجابة بلا لو وجدنا نموذجا جاهزا مناسبا لتوسيم البيانات التي لدينا أو أن هناك بيانات موسمة تفي بالغرض وتنطلق من نفس النظرية التي ننطلق منها ونستطيع من خلالها بناء نموذجنا الخاص بالتصنيف. أو لو كان هناك نموذج تصنيف نرتضي دقته ولكنه لا ينطلق من نفس النظرية التي نريد تطبيقها ولكننا يمكن أن نتصرف في النتائج بحيث تتناسب مع حاجتنا مثل أن ندرج وسما معينا ليكون مع آخر أو أن نفصل وسما واحدا لعدة وسوم باستخدام القواميس والقواعد. من أمثلة هذه الحالة لو أن نموذج التصنيف الخاص بأقسام الكلام يضع جميع الحروف تحت وسم واحد فإننا بسهولة نستطيع التعامل مع الوسم ونفصله لوسوم متعددة بحسب الحاجة مثل تحديد وسم لحروف الجر وآخر لحروف النصب وغيرهما.

لو كانت الإجابة على السؤال السابق بنعم نحتاج الإجابة على الأسئلة الخاصة بكل عامل من العوامل المؤثرة في بناء البيانات الموسمة يدويا كما يلي:

1.4 البيانات الخام

وهي البيانات التي نحتاج توسيمها. هل تتوفر لدينا البيانات التي نريد توسيمها يدويا لنبني نموذج التصنيف؟ وهل يتحقق فيها معياري التوازن والتمثيل؟ وهل هي منحازة لنوع معين من النصوص مثلا؟ هل هناك تكرار في النصوص أو الجمل؟ وهل هذه البيانات كافية بحسب الأبحاث السابقة لنفس المشكلة من التوسيم؟ وهل هناك حقوق ملكية فكرية لهذه البيانات؟ أم أننا بحاجة لبناء بيانات جديدة غير موسمة تفي بالغرض من حيث التمثيل والتوازن والحجم ولدينا الأذن باستخدامها؟

فلو احتجنا لبناء مدونة جديدة يجب أن نضع لها خطة مناسبة للجمع تتضمن معايير تصميم المدونة ومصادر الجمع وطريقة حفظ النصوص وترميزها. إن موضوع بناء المدونات موضوع يطول شرحه لذا أحيل القارئ الكريم الى الثبيتي (2015) لمعلومات أكثر تفصيلا حول هذا الموضوع.

2.4 الوسوم المستخدمة

لوكان بناء نموذج للتصنيف يلبي حاجة عملية ليس لها نظرية علمية تنطلق منها مثل تصنيف أنواع الخدمات التي ترد لجهة ما من مستخدميها بناء على رسائل الواتساب أو المحادث الآلي، فإننا في هذه الحالة نكون أكثر حرية، ومجال الاجتهاد واسع لوضع الوسوم المناسبة لهذه المهمة. ولكن لو كانت الوسوم تنطلق من نظرية علمية مثل أقسام الكلام المتعارف عليها في النحو العربي فنحن هنا ملزمون باتباع هذه التقسيمات ويكون مجال اجتهادنا في تقليص الوسوم مثل الاكتفاء بالاسم والفعل والحرف مثلا أو توسيعها لتشمل جميع الوسوم الرئيسية وما يتبعها.

وفي كلتا الحالتين يجب أن يكون لدينا قائمة كاملة بهذه الوسوم مع المراجع المناسبة لها. ومن المفترض أن يكون هناك مختص بالموضوع يكون مسؤولا عن تحديد هذه الوسوم بمشاركة أعضاء الفريق. إن تحديد الوسوم المستخدمة بشكل واضح يساعد في تحديد مدى الجهد المبذول في التوسيم وما يحتاجه من خبرات وأعداد.

بعد إعداد قائمة الوسوم يجب أن يكون هناك دليل مفصل - قدر الإمكان - للتوسيم يشرح كل وسم من الوسوم ويعرض أمثلة متعددة عليه ويوضح أي لبس قد يحصل مع غيره من الوسوم. ويجب أن يوضح كيف يقوم الموسم بوضع الوسم واختياره بحسب الطريقة التي اخترناها للتوسيم. هذا الدليل يجب أن يكون بلغة بسيطة ومباشرة لا تحمل لبسا للمتلقي. وقد يكون من المستحب وضع اختبار بسيط للموسم بعد انتهاء شرح كل وسم وتكون اجاباته الصحيحة في آخر الدليل لاختبار معرفة الموسم بهذا الوسم.

لنفرض أننا نعد دليلا للتوسيم بأقسام الكلام وفي القسم الخاص بوسوم الحروف نريد الحديث عن حروف العطف وأحدها هو حرف الفاء. هنا نضع مثالا لاستخدام الفاء كحرف عطف ثم نضع أنواعه الأخرى مثل فاء التعليل وفاء السببية والفاء التعليلية وغيرها من الأنواع مع الأمثلة المناسبة لها مع توضيح الفرق بين هذه الأنواع وفاء العطف. ويوفر دليل التوسيم بأقسام الكلام الخاص بالبنك الشجري العربي مثالا جيدا لأدلة التوسيم [6] .

3.4 الأسلوب المستخدم في التوسيم

ماهي الطريقة التي سوف نستخدمها لتوسيم البيانات؟ بشكل عام نستطيع اتباع ثلاثة أساليب للتوسيم. أولها أن نعطي الموسمين النصوص كل نص على حدة ونطلب منهم إضافة الوسوم على هذه النصوص بالطريقة التي نرغبها. ولكن يعيب هذه الطريقة كثرة الأخطاء التي قد يقع فيها الموسمون في كتابة الوسوم والتفاوت الذي يحصل بينهم في طريقة إضافة الوسوم علاوة على التغيرات التي قد يحدثونها في النصوص بقصد أو بغير قصد.

الأسلوب الثاني أن نعطي الموسم النصوص في ملفات إكسل ونضع فيها شفرات برمجية تجعله يختار الوسوم من قوائم منسدلة. هذا بطبيعة الحال يقلل التفاوت بين الموسمين ويقلل الأخطاء ولكنه لا يمنع التعديل في البيانات. ويعيب هذا الأسلوب كما يعيب الأسلوب الأول أن البيانات ليست بمنأى عن التسرب ويزداد هذا الخطر لو كانت البيانات حساسة أو لها حقوق ملكية فكرية. ويمكن تقليل خطر تسرب البيانات من خلال حضور فريق التوسيم لمقر المشروع وعملهم على الأجهزة المتوفرة ولكنه يتطلب تكلفة إضافية تختص بتجهيز أماكن وأجهزة خاصة بالتوسيم.

أما الأسلوب الثالث فهو الأكثر أمانا والأقل أخطاء ولكنه أكثر تكلفة وهو استخدام منصة خاصة بالتوسيم على شبكة الإنترنت تمكننا من التحكم بعملية التوسيم بشكل كامل فنضمن التوافق والاتساق في عملية التوسيم وكذلك منع تسرب البيانات كما تساعدنا أيضا في معرفة أداء الموسمين والتحكم بالأوقات التي يقومون فيها بالتوسيم وعدد النصوص التي يوسمونها تقليلا لأخطاء التوسيم. كما تضمن هذه المنصة وصول القائمين على مشروع التوسيم وكذلك الموسمين للبيانات التي تخصهم في أي وقت ومن أي مكان تتوفر فيه خدمة الإنترنت. كما يُمَكن هذا الأسلوب من جعل عملية التوسيم أكثر متعة من خلال أساليب التلعيب والمسابقات أثناء عملية التوسيم. وتقل تكلفة هذا الخيار عندما تستخدم هذه المنصة في مشاريع متعددة للتوسيم فيكون الاستثمار في مثل هذه المنصة مجديا.

---

(6) https://www.yumpu.com/en/document/view/11808334/arabic-treebank-morphological-pos-guidelines-v37-ldc-projects

4. توسيم البيانات يدويا

كل هذه الخيارات متاحة لفريق المشروع وله اختيار ما يناسبه منها أو من غيرها إن وجدها أكثر ملائمة بحسب ما يتوفر له من إمكانات. ولكن ماهي الطريقة التي نتأكد بها من جودة التوسيم؟ نستطيع استخدام طريقتين الأولى أن يكون هناك شخص أو فريق أكثر خبرة من الموسمين يقوم بمراجعة أعمالهم وتصحيحها إذا أقتضى الأمر. وفي هذه المرحلة يجب أن نحدد ما إذا كانت المراجعة تتم بالتوازي مع عملية التوسيم أو بعد انتهائها. إن المراجعة خلال عملية التوسيم تمكننا من تقليل الأخطاء وتوجيه الموسمين حيال اختياراتهم لو كانت خاطئة فيقومون بتجنب مثل هذه الأخطاء عند توسيم البيانات الأخرى التي لم توسم. أما الطريقة الثانية للتأكد فهي أن يتشارك ثلاثة موسمين أو أكثر في توسيم النص الواحد بدون اطلاع أحدهم على توسيم الآخرين ثم نأخذ ما استقر عليه الغالبية من الموسمين. وهذا الخيار أبطاء ولكنه أكثر جودة في أغلب الأحيان. كما أن تلقي ملاحظات مستخدمي البيانات الموسمة لاحقا وكذلك نتائج تقييم نموذج التصنيف سوف تحسن من جودة البيانات لاحقا.

ولتقليل الوقت المستغرق في التوسيم وكذلك التكلفة يمكننا أن نوسم بعض البيانات آليا باستخدام قواعد منطقية مباشرة وقائمة بالكلمات التي تحوي وسوما فريدة. ففي التوسيم بأقسام الكلام على سبيل المثال يمكننا أن نسم أكثر الحروف مباشرة لأنها معروفة ومحددة سلفا. كما أن الكلمات الأكثر تكرار في العربية والتي لها وسوم ثابتة يمكن إضافتها للقائمة مثل «قال»، «بن»، «رئيس» وغير ذلك. استخدام مثل هذا الأسلوب ممكن أن يوفر حوالي 60% من الوقت والتكلفة (الثبيتي وآخرون 2012). ولا شك أن توفر قوائم لأسماء الأعلام والمنظمات والدول والمواقع الجغرافية بمختلف أنواعها سوف يساعد في البناء السريع لبيانات موسمة تختص بالتعرف على الكينونات الأسمية.

4.4 القائمون بعملية التوسيم

بناء على ما يتقرر بخصوص الوسوم ومنهجية التوسيم يتم اختيار الموسمين. وأول سؤال هو ماهي المواصفات التي يجب أن تتوفر في الموسمين. فالوسوم التي تنطلق من نظرية علمية بحاجة إلى موسمين متخصصين في المجال ليطبقوها على البيانات الخام مثل التوسيم بأقسام الكلام الذي يحتاج إلى متخصصين وعارفين بالنحو. بينما توسيم الكينونات الأسمية لا يحتاج إلى تخصص معين بل بحاجة إلى قدر معقول من التعليم والثقافة العامة. وقد لا يكون التخصص والمعرفة هي المطلب الوحيد فقد يكون العمر وجنس الموسم مؤثر في عملية التوسيم مثلما هو الحال في توسيم التغريدات في تويتر. ففي هذه الحالة اختلاف الأعمار وإشراك موسمين وموسمات من مناطق وبيئات مختلفة يزيد من جودة البيانات ودقتها. ومن أهم المواصفات التي يجب تحديدها في الموسمين: المستوى التعليمي، السن، الجنس، والمنطقة، وإمكانية اتصاله بالإنترنت.

ولا يمكن أن يبدأ الموسمون عملهم بدون تدريب مسبق ومكثف نتأكد من خلاله من فهمهم للوسوم ولأسلوب التوسيم الذي اخترناه مع مراقبة مستمرة لإنجازهم وتصحيحه إن لزم الأمر. ويجب أيضا وضع بروتوكول للتواصل مع الموسمين وكيفية تبليغهم عن أي ملاحظات يجدونها على البيانات وأي استفسارات لديهم حول التوسيم وكيفية إبلاغنا لهم بأي مستجدات تخص عملية التوسيم.

بعد انتهائنا من توسيم البيانات يجدر بنا التأكد من جودتها باتباع الخطوات التالية:

أ. أخذ عينات عشوائية من البيانات الموسمة وفحصها بدقة للتأكد من جودتها. فلو كانت نسبة الأخطاء مقبولة (3% على سبيل المثال) نعدل هذه الأخطاء ونقبل بجودة البيانات بشكل عام. ولو كانت نسبة الأخطاء أكبر من ذلك فنحن بحاجة لمراجعة كاملة للبيانات لتحقيق الجودة.

ب. التأكد من وجود عينات كافية لكل وسم من الوسوم. فالوسوم غير الموجودة في البيانات أو نادرة الوجود لا يمكن لنموذج التصنيف أن يتعرف عليها. ولذلك يتوجب علينا إضافة عينات كافية لتمثيل هذه الوسوم في البيانات.

ج. عند استخدام أكثر من موسم لتوسيم نفس البيانات، يجب التأكد من اتساق عملية التوسيم بين الموسمين وأنهم يقومون بتوسيم البيانات بنفس الطريقة من خلال عدة طرق إحصائية من ضمنها مقياس فلييس-كابا Fleiss› kappa (Fleiss 1971).

د. تدريب خوارزميات متعددة للتصنيف على البيانات الموسمة والنظر في نتائج التقييم هل هي مناسبة أم لا. لو كانت نتائج التصنيف غير مناسبة نحتاج لمراجعة خطوات بناء نموذج التصنيف المذكورة آنفا.

ه. إن قبلنا بنتائج التصنيف فإننا نقوم بتطبيق النموذج على بيانات غير موسمة في مجالات واقعية أو أن نتيح البيانات لجهات أخرى لاستخدامها وبناء نماذج تصنيف خاصة بهم ونتقبل أي ملاحظات نشاهدها أو تردنا على أداء نموذج التصنيف أو البيانات الموسمة ونقوم بتصحيح ما يجب تصحيحه.

ومن المستحسن أيضا أن نقوم بخطوات التحقق من جودة البيانات -المذكورة أعلاه – أثناء عملية التوسيم بحيث نقوم بتوسيم البيانات على مراحل كل مرحلة تتضمن 20% من البيانات وبعد كل مرحلة نقوم بالتحقق من جودة التوسيم. اتباع هذا الأسلوب يساهم في تقليل الوقت والجهد اللازمين لرفع جودة البيانات. كما أنها يمكن أن تدلنا في وقت مبكر على حجم البيانات المناسب لتدريب خوارزمية التصنيف. فقد يكون أداء نموذج التصنيف مناسبا بعد توسيم 60% من البيانات الخام. وبهذا نوفر وقتا وجهدا وتكلفة قد لا يكون لإهدارها داع.

5. حقوق الملكية الفكرية

يجب أن نأخذ موضوع حقوق الملكية الفكرية في مشروع توسيم البيانات على محمل الجد، فقد يتسبب خرقنا للقوانين المرعية في هذا المجال في إيقاف المشروع وخسارتنا للوقت والجهد والتكلفة المبذولة في إنجازه ناهيك عما يترتب علينا من عقوبات لوثبت خرقنا لقوانين حماية الملكية الفكرية. علاوة على ذلك فإن انتهاك حقوق الآخرين الفكرية قد يتسبب في كثير من الأحيان في ضياع مباشر وغير مباشر لحقوقهم المالية والأدبية. وكما أنك يجب أن تراعي حقوق الآخرين فيجب عليك أيضا أن تراعي حقوقك أنت وتحمي حقوق الملكية الخاصة ببياناتك الموسمة وبالنماذج الخاصة بها، وإن أردت إتاحتها للعموم فحاول أن تكون ضمن إحدى رخص المشاع الابداعي [7].

نستطيع حصر ما يتعلق بحقوق الملكية الفكرية عند بناء المدونات الموسمة يدويا في أربع مجالات. المجال الأول يتعلق بالبيانات الخام التي نريد توسيمها، والثاني في البيانات الموسمة يدويا، والثالث في الأدوات والبرامج المستخدمة في التوسيم، والرابع في النموذج الحاسوبي الناتج من استخدام البيانات الموسمة.

أ. البيانات الخام: يمكننا تصنيف أنواع البيانات الخام من ناحية حقوق الملكية الفكرية إلى ثلاثة أصناف. الصنف الأول أن تكون البيانات الخام التي نريد العمل عليها ملك للجهة التي تريد توسيمها يدويا أو أنها ضمن إحدى رخص المشاع الإبداعي ونعرف حدود استخدامها وبالتالي ليس هناك مانع قانوني من استخدامها. فالمحادثات التي تتم من خلال المحادث الآلي الخاص بجهة ما تعتبر ملكا لها وبالتالي فهي ليست بحاجة للحصول على أذن من أي جهة باستخدامها. ولكن قبل إتاحة هذه البيانات لفريق التوسيم وكذلك إتاحتها للعموم يجب التأكد من أن هذه البيانات لا تنتهك خصوصية الأفراد مثل الأسماء ومواقع السكن وغير ذلك.

والصنف الثاني بيانات لا نمتلكها، ولكنها غير محمية بموجب قانون حقوق الملكية الفكرية وبالتالي نستطيع استخدامها في مشروع التوسيم بدون تردد. فنظام حماية حقوق الملكية الفكرية في المملكة العربية السعودية [8] على سبيل المثال يستثني بعض المصنفات من الحماية (انظر المادة الرابعة) مثل الأنظمة والأحكام القضائية، وإجراءات العمل و «ما تنشره الصحف، والمجلات، والنشرات الدورية، والإذاعة من الأخبار اليومية، أو الحوا دث ذات الصبغة الإخبارية». وينطبق على هذا النوع الكتب التراثية مثل تاريخ ابن كثير على سبيل المثال وما تخطي حدود مدة الحماية (حياة المؤلف ولمدة خمسين عاما بعد وفاته بحسب النظام في المملكة العربية السعودية).

والصنف الثالث بيانات محمية بموجب القانون وبالتالي فنحن ملزمون بالحصول على إذن باستخدامها أو أن نتبع ما يتيحه القانون من استثناءات للاستخدام النظامي. فالمادة الخامسة عشر من نظام حماية حقوق الملكية الفكرية في المملكة العربية السعودية على سبيل المثال لا يمنع من الاستشهاد بفقرات من مصنف ما في مصنف آخر شريطة أن يكون الاستشهاد متماشيا مع العرف وشريطة ذكر المصدر واسم المؤلف (لطفا انظر بقية الاستثناءات مثل عدم التربح والاستخدام للأغراض التعليمية).

---

(7) https://creativecommons.org/licenses/?lang=ar

(8) https://www.saip.gov.sa/wp-content/uploads/2019/10/%D9%86%D8%B8%D8%A7%D9%85-%D8%AD%D9%85%D8%A7%D9%8A%D8%A9-%D8%AD%D9%82%D9%88%D9%82-%D8%A7%D9%84%D9%85%D9%88%D9%94%D9%84%D9%81L-LA-041-01.pdf

ولو كانت حاجتنا لا تتطلب نصوصا كاملة مثل التوسيم بأقسام الكلام أو التحليل الصرفي والتعرف على الكينونات الأسمية، فإننا في هذه الحالة نستطيع بناء البيانات غير الموسمة من خلال اختيار ثلاثة جمل من أول المصنف ووسطه وآخره. ونكرر هذا مع أي مصنف محمي حتى نصل لحجم البيانات التي نريدها.

وفي المجمل، فإننا نستطيع الحصول على البيانات اللازمة والتعامل مع قضية الملكية الفكرية بدون مشاكل خاصة وأن البيانات التي نحتاجها في عمليات التوسيم صغيرة الحجم في العادة.

ب. البيانات الموسمة: البيانات الموسمة يدويا ملك للجهة التي قامت بتوسيمها حتى ولو كانت البيانات الخام خاضعة للحماية (فيما يعتقده المؤلف) [9] ، ونظام حماية حقوق الملكية الفكرية في المملكة العربية السعودية على سبيل المثال يوضح هذا بشكل جلي في عدة مواضع من المادة الثالثة - كلها يمكن أن تنطبق على البيانات الموسمة يدويا - كما يلي:

- «مصنفات التلخيص، أو التعديل، أو الشرح، أو التحقيق، أو غير ذلك من أوجه التحوير».
- «قواعد البيانات سواء أكانت بشكل مقروء آلياً أم بأي شكل آخر، والتي تعد مبتكرة من حيث اختيار أو ترتيب محتوياتها».

اعتمادا على الفقرتين أعلاه يمكنا النظر للتوسيم على أنه شرح للمصنف الأصلي كما يمكننا اعتبار البيانات الموسمة والتي في العادة تكون على شكل بيانات مجدولة بصيغة الكترونية نوعا من قواعد البيانات المبتكرة من حيث الوسوم ورموزها والمعلومات الأخرى المرتبطة بها.

ج. الأدوات والبرامج المستخدمة في التوسيم: نحاول قدر المستطاع استخدام برامج مفتوحة المصدر نعرف حدود رخص استخدامها ولا نستخدم إلا نسخا أصلية من المنتجات التجارية متى ما احتجنا لذلك. فلا داعي أبدا لصرف مبالغ طائلة بلا فائدة أو حاجة ضرورية. فيمكننا على سبيل المثال بناء موقع لتوسيم البيانات باستخدام الأدوات والبرامج مفتوحة المصدر بداء من أدوات بناء واجهات المستخدم وليس انتهاء بنظام إدارة قاعدة البيانات.

د. نموذج التصنيف: نموذج التصنيف الناتج من عملية التدريب على البيانات الموسمة هو ملك للجهة المنتجة له ويحق لها استخدامه وتوزيعه كما تشاء. ولكن يجب البحث في رخصة استخدام البيانات عن حدود استخدام وملكية النماذج الناتجة. فبعض الجهات المالكة للبيانات قد توفر البيانات الموسمة مجانا، ولكنها تشترط الحصول على مقابل مالي في حال لو استخدام النموذج في الأغراض التجارية.

وعلى كل حال، وبغض النظر عن كل ما ذكرته آنفا بخصوص الملكية الفكرية، فمن المستحسن أن يكون لدى فريق التوسيم مستشار قانوني يتولى التأكد من حقوق الملكية الفكرية الخاصة بأي بيانات أو نماذج وأدوات ليست ملكا للجهة التي يعمل لها فريق التوسيم. كما يقوم بتجهيز الوثائق ورخص الاستخدام التي تحمي حقوق الملكية الفكرية للبيانات الخام وللبيانات الموسمة يدويا وللنماذج ولكل نواتج مشروع التوسيم الأخرى – إن كان ذلك ضروريا - مثل موقع التوسيم على سبيل المثال.

---

[9] يجب ألا يسلم بهذا الرأي بل على الجهة القائمة بالتوسيم استشارة جهة قانونية في هذا المجال. لا يتحمل المؤلف أي مسئولية قانونية تجاه هذا الرأي.

6. ثم أما بعد

أوضح هذا الدليل بشكل موجز عن أهم العوامل المؤثرة في عملية بناء مدونة موسمة يدويا وهي البيانات الخام، والوسوم المستخدمة، وأسلوب التوسيم والقائمين بعملية التوسيم. واستعرض الدليل في بدايته أهمية وجود البيانات الموسمة يدويا، ثم بعض أهم نماذج عمليات التوسيم مثل التقطيع وتصنيف الكلمات والأحرف. ثم شرح الخطوات التي تستخدم فيها البيانات الموسمة لبناء نموذج حاسوبي خاص بالتوسيم لأهمية وجود مثل هذا النموذج في تقييم جودة البيانات الموسمة يدويا. ثم انتهت بالحديث عن جانب مهم قد يغفل عنه بعض المشتغلين باللسانيات الحاسوبية العربية ومعالجة اللغة العربية ألا وهو حقوق الملكية الفكرية. وقد حاولت قدر الإمكان الاستشهاد بأمثلة توضيحية ومراجع علمية من مجالي اللسانيات الحاسوبية ومعالجة اللغة العربية.

وأعيد التذكير هنا بأن العاملين في مجالي اللسانيات الحاسوبية ومعالجة اللغة العربية -وبالتالي العاملين في مجال لسانيات المدونات اللغوية – يعانون من مشكلة عويصة هي شح البيانات المرجعية الموسمة يدويا وهذا بدوره انعكس سلبا على جودة الأبحاث في هذه المجالات وعلى عددها. ولتخطي هذه المشكلة اقترح بناء بيانات لغوية مرجعية موسمة يدويا تراعي حقوق الملكية الفكرية في بياناتها الخام كما أوضحت سابقا ويتم توسيم جملها وكلماتها بأكثر من وسم مثل تقطيع المتصلات والزوائد الصرفية، والتشكيل، والتوسيم بأقسام الكلام، والأوزان الصرفية والجذوع والجذور، وتوسيم الكينونات الأسمية. وحتى تسهل عملية التوسيم يمكننا إنشاء موقع على شبكة الإنترنت يستطيع كل من له رغبة المساهمة في توسيم البيانات بحسب خبرته ومعرفته. وبعد انتهاء التوسيم والتأكد من جودته تتاح البيانات للباحثين بشكل مجاني بحسب رخصة المشاع الإبداعي المناسبة.

شكر وتقدير

يشكر المؤلف كل من: الدكتور عبدالرحمن المحارب والدكتور وليد الصانع والأستاذة منيرة بنت زيد الحوشان من المركز الوطني لتحليل البيانات والذكاء الاصطناعي بمدينة الملك عبدالعزيز للعلوم والتقنية والدكتورة فاطمة بنت عبيد الثبيتي من معهد الحرم المكي الشريف على مراجعتهم لمسودة هذا الدليل وللملاحظات القيمة التي أبدوها.

المراجع العربية

- حسان، تمام. (2004). اللغة العربية معناها ومبناها . عالم الكتب، القاهرة.
- الثبيتي، عبدالمحسن. الطوالة، ندى. المرشدي، بشاير و العتيبي، سعد.(2012). طريقة تعتمد على المدونات اللغوية لتجهيز بيانات تدريب واختبار أنظمة الوسوم النحوية. المؤتمر الدولي لعلوم وهندسة الحاسوب باللغة العربية في دورته الثامنة. القاهرة. جمهورية مصر العربية.
- التميمي، أفراح. (2019). التوسيم النحوي للمدونات العربية: نماذج توسيمية مقترحة [رسالة دكتوراة، جامعة الإمام محمد بن سعود الإسلامية]، الرياض، المملكة العربية السعودية.
- المجدوب، عزالدين. الثبيتي، عبدالمحسن. اللاحم، إبراهيم. كرونة، سندس. (2019). الوسم النحوي الآلي للعربية في منهجية بنك المشجرات النحوية في جامعة بنسيلفانيا. مجلة اللسانيات العربية. ع 9، 6-78.
- الثبيتي، عبدالمحسن. (2015). تصميم المدونات اللغوية وبناؤها. في صالح فهد العصيمي (محرر.)، المدونات اللغوية العربية: بناؤها وطرق الإفادة منها. (ص ص 147-178) مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية. الرياض. المملكة العربية السعودية.

المراجع الأجنبية

- Al-Thubaity, A. O. (2015). A 700M+ Arabic corpus: KACST Arabic corpus design and construction. Language Resources and Evaluation, 49(3), 721751-.
- Ali, I. H., Mnasri, Z., & Laachri, Z. (2019, March). Gemination prediction using DNN for Arabic text-to-speech synthesis. In 2019 16th International Multi-Conference on Systems, Signals & Devices (SSD) (pp. 366370-). IEEE.
- Almuhareb, A., Alsanie, W., & Al-Thubaity, A. (2019). Arabic word segmentation with long short-term memory neural networks and word embedding. IEEE Access, 7, 12879.12887-
- Alnaied, A., Elbendak, M., & Bulbul, A. (2020). An intelligent use of stemmer and morphology analysis for Arabic information retrieval. Egyptian Informatics Journal, 21(4), 209217-.
- Alonso, M. A., Gómez-Rodríguez, C., & Vilares, J. (2021). On the Use of Parsing for Named Entity Recognition. Applied Sciences, 11(3), 1090.
- Alqahtani, S., Aldarmaki, H., & Diab, M. (2019, August). Homograph Disambiguation through Selective Diacritic Restoration. In Proceedings of the Fourth Arabic Natural Language Processing Workshop (pp. 49.(59-
- Al-Smadi, M., Talafha, B., Al-Ayyoub, M., & Jararweh, Y. (2019). Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. International Journal of Machine Learning and Cybernetics, 10(8), 21632175-.
- Al-Thubaity, A., Alhoshan, M., & Hazzaa, I. (2015). Using word N-grams as features in Arabic text classification. In Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (pp. 3543-). Springer, Cham.
- Al-Thubaity, A., AlKhalifa, A., Almuhareb, A., & Alsanie, W. (2020). Arabic Diacritization Using Bidirectional Long Short-Term Memory Neural Networks with Conditional Random Fields, IEEE Access, 8, 154984154996-.
- Al-Thubaity, A., Alkhalifa, A., Almuhareb, A., & Alsanie, W. (2020). Arabic Diacritization Using Bidirectional Long Short-Term Memory Neural Networks With Conditional Random Fields. IEEE Access, 8, 154984154996-.
- Al-Thubaity, A., Alqahtani, Q., Aljandal, A. (2018). "Sentiment lexicon for sentiment analysis of Saudi dialect tweets". Procedia Computer Science, 142, 301307-.
- Anderson, M. D., & Vilares, D. (2018). Increasing NLP parsing efficiency with chunking. Multidisciplinary Digital Publishing Institute Proceedings, 2(18), 1160.

المراجع الأجنبية

- Chen, D., & Yih, W. T. (2020, July). Open-Domain Question Answering. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts (pp. 3437-).
- Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A multilingual evaluation for online hate speech detection. ACM Transactions on Internet Technology (TOIT), 20(2), 1.22-
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Diab, M., Habash, N., Rambow, O., & Roth, R. (2013). LDC Arabic treebanks and associated corpora: Data divisions manual. arXiv preprint arXiv:1309.5652.
- El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2020). Automatic Text Summarization: A Comprehensive Survey. Expert Systems with Applications, 165 113679.
- Eshkol, I., Maarouf, M., Badin, F., Skrovec, M., & Tellier, I. (2020, May). Chunk Different Kind of Spoken Discourse: Challenges for Machine Learning. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 5164.(5168-
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5), 378.
- Freihat, A. A., Abbas, M., Bella, G., & Giunchiglia, F. (2018). Towards an Optimal Solution to Lemmatization in Arabic. Procedia computer science, 142, 132.140-
- Fügen, C., & Kolss, M. (2007, August). The influence of utterance chunking on machine translation performance. In Eighth Annual Conference of the International Speech Communication Association (pp. 2837. .(2840-
- Hailu, T. T., Yu, J., & Fantaye, T. G. (2020). A Framework for Word Embedding Based Automatic Text Summarization and Evaluation. Information, 11(2), 78.
- Hriez, S., & Awajan, A. (2020, July). Authorship Identification for Arabic Texts Using Logistic Model Tree Classification. In Science and Information Conference (pp. 656666-). Springer, Cham.
- Kadim, A., & Lazrek, A. (2018). Parallel HMM-based approach for arabic part of speech tagging. Int. Arab J. Inf. Technol., 15(2), 341351-.

- Kadim, A., & Lazrek, A. (2018). Parallel HMM-based approach for arabic part of speech tagging. Int. Arab J. Inf. Technol., 15(2), 341351-.

- Kandasamy, S., & Cherukuri, A. K. (2020). Query expansion using named entity disambiguation for a question answering system. Concurrency and Computation: Practice and Experience, 32(4), e5119.
- Khorsheed, M. S., & Al-Thubaity, A. O. (2013). Comparative evaluation of text classification techniques using a large diverse Arabic dataset. Language resources and evaluation, 47(2), 513538-.
- Kumar, N., & Sonowal, S. (2020, July). Email Spam Detection Using Machine Learning Algorithms. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 108113-). IEEE.
- Leech, G. (2013). Grammatical Tagging. In: Garsire, R., Leech, G., McEnery, A. (eds.) Corpus Annotation: Linguistic Information for Computer Text Corpora. Routledge, New York.
- Liu, L., Shang, J., & Han, J. (2019, August). Arabic Named Entity Recognition: What Works and What's Next. In Proceedings of the Fourth Arabic Natural Language Processing Workshop (pp. 60.(67-
- Ma, D., Li, S., Wu, F., Xie, X., & Wang, H. (2019, July). Exploring sequence-to-sequence learning in aspect term extraction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3538.(3547-
- Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004, September). The penn arabic treebank: Building a large-scale annotated arabic corpus. In NEMLAR conference on Arabic language resources and tools (Vol. 27, pp. 466.(467-
- Madi, N., & Al-Khalifa, H. (2020). Error Detection for Arabic Text Using Neural Sequence Labeling. Applied Sciences, 10(15), 5279.
- Manning, C. D. (2011, February). Part-of-speech tagging from 97% to 100%: is it time for some linguistics?. In International conference on intelligent text processing and computational linguistics (pp. 171189-). Springer, Berlin, Heidelberg.
- Mishra, G. S., & Raj, D. (2020). Language Translation Using Natural Language Processing. Language, 7(9), 2020.
- Ekman, P. (1992). An argument for basic emotions. Cognition & emotion, 6(3200-169 ,(4-.
- Perera, N., Dehmer, M., & Emmert-Streib, F. (2020). Named Entity Recognition and Relation Detection for Biomedical Information Extraction. Frontiers in Cell and Developmental Biology, 8, 673.
- Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

المراجع الأجنبية

- Soares, M. A. C., & Parreiras, F. S. (2020). A literature review on question answering techniques, paradigms and systems. Journal of King Saud University-Computer and Information Sciences, 32(6), 635646-.
- Wu, C., Wu, F., Qi, T., Liu, J., Huang, Y., & Xie, X. (2020). Detecting Entities of Works for Chinese Chatbot. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 19(6), 1.13-
- Zalmout, N., & Habash, N. (2020, July). Joint Diacritization, Lemmatization, Normalization, and Fine-Grained Morphological Tagging. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 82978307-).