**تقرير حول دراسة الدكتوراه**

**للطالب مدين عبد الحميد**

**عنوان البحث**

**كشف الشذوذ في النصوص العربية في شبكات التواصل الاجتماعي**

**Anomaly Detection in Arabic Text at Social Networks**

**مقدمة**

تتيح مواقع التواصل الاجتماعي للمستخدمين التعبير عن آرائهم وتحقيق التواصل فيما بينهم وتبادل المعلومات من خلال التغريدات والتعليقات والمحادثات. تتيح خوارزميات التنقيب عن المعطيات النصية إمكانية استخلاص المعلومات والمعرفة المخبأة في هذه النصوص. ولكن كون هذه النصوص قصيرة وذات ضجيج وبنيتها متناثرة يجعل عملية التنقيب عن المعطيات في النصوص صعبة.

تعتبر الأبحاث الخاصة باللغة العربية قليلة مقارنة باللغات الأخرى واسعة الانتشار وذلك بسبب صعوبة اللغة العربية وتعقيداتها في التشكيل وطبيعتها الخاصة في الاشتقاق والإعراب. تتعامل معظم هذه الأبحاث مع الوثائق بشكل offline من خلال معطيات مخزنة مسبقًا ولا تتعامل معها كمعطيات دفقية stream data. مع ذلك، فالأبحاث التي عالجت النصوص كبيانات دفقية، عالجت الموضوع على وثائق كبيرة الحجم، واستخدمت مفاهيم مختلفة وذلك بسبب طبيعة المعطيات الدفقية فهي بيانات ذات تدفق مستمر وتتطلب استجابة سريعة، ما يجعل آليات التنقيب التقليدية غير قابلة للتطبيق في هذا الحالة.

تتعامل معظم الأبحاث على الوثائق العربية من خلال خوارزميات التعلم التقليدية Machine Learning، وعدد قليل جداً منها يتم تتناول المسألة من منظور التعلم العميق Deep Learning.

تقدم وسائط التواصل الاجتماعي كمية هائلة من المعطيات عن سلوك الأشخاص الذين يتفاعلون معها من خلال أسلوب تواصلهم فيما بينهم وتواصلهم مع المصادر المفتوحة كالكتب والصور والفيديو وغيرها. كذلك مع ازدياد استخدام أنظمة المراقبة واستخدام الأجهزة الذكية أصبح من السهل الحصول على معطيات كبيرة ومتنوعة. تعتبر مصادر التواصل الاجتماعي مفتوحة، ما يجعلها عرضة للاستخدام السيء من قبل بعض الأشخاص المشبوهين، حيث باتت تستخدم للتخطيط للجرائم والأعمال الإرهابية وتهديد أمن المجتمع. لذلك أصبح من الضروري العمل على تطوير نظم تعتمد على المعطيات الكبيرة من أجل كشف الأعمال المشبوهة وحذف التهديدات التي تمثلها. تستخدم هذه النظم معطيات شبكات التواصل الاجتماعي والمعطيات المحصلة من أنظمة المراقبة وخاصة الموجودة على الحدود ومعطيات شبكات الاتصال الخليوي لكشف السلوك المشبوه وحذف التهديد الذي يمثله وكذلك كشف التوجهات الطبيعية ضمن المجتمع للتفاعل الصحيح معها واستخدامها من أجل تلبية حاجات المجتمع بأفضل طريقة ممكنة.

انتشر خطاب الكراهية خلال الشبكات الاجتماعية مثل تويتر وفيسبوك، وأصبح له تأثير واضح مما يقتضي كشفه وتحديد مصدره للحد من خطره في المجتمع، إن التحديد المبكر للمستخدمين الذين يروجون لمثل هذا النوع من الرسائل يمكن أن يمنع التصعيد من الكلام إلى الفعل. يوجد في منصات وسائل التواصل الاجتماعي عدد لا يمكن السيطرة عليه من الرسائل الصادرة في كل ثانية مما يجعل من المستحيل تتبع أو التحكم في محتوى هذه المنصة يدويًا، لذلك يجب مواجهته من خلال تسخير قوة الذكاء الاصطناعي وخوارزميات التعلم الآلي لأتمتة كشف خطاب الكراهية في وسائل التواصل الاجتماعي وخاصةً عندما لا يحتوي النص على كلمات صريحة. ويجب الأخذ بعين الاعتبار المعلومات السياقية التي تحملها هذه الرسائل، وتواجه المنصات الاجتماعية مشكلة في الحد من هذه الرسائل مع موازنة حرية التعبير، ونظرًا لأن الأشخاص الذين ينشرون خطاب الكراهية قد يتم حظرهم أو معاقبتهم أو مراقبتهم لعدم وضع تهديداتهم في أفعال، فإن وثوقية الكشف الآلي هامة في صنع القرار وتساعد المراقب البشري بتوضيح الرؤية.

تعتبر اللغة العربية واحدة من أكثر اللغات انتشارًا في العالم فهي خامس أكثر اللغات استخدامًا في العالم وخامس أكثر اللغات استخدامًا على الإنترنت. يتحدث أكثر من 6.0٪ من سكان العالم اللغة العربية وهناك نمو ملحوظ في استخدام منصات التواصل الاجتماعي في المنطقة العربية، ويُنشر الكثير من خطاب الكراهية باللغة العربية على تويتر وغيرها ولا يوجد إلا عدد قليل من الدراسات لكشفه وذلك لصعوبات في اللغة العربية منها وجود كثير من اللهجات مثل (المصرية، الخليجية، المشرقية، المغربية، وغيرها...) بالإضافة للغة الكلاسيكية، واللغة الحديثة المنتشرة في وسائل الإعلام، وتتميز اللغة العربية بخواص تختلف بها عن اللغات الأخرى فلديها مورفولوجيا غنية وبنية نحوية معقدة ومما يميزها أيضًا الكم الهائل من المفردات المترادفة.

تعد مسألة كشف خطاب الكراهية إحدى مسائل التصنيف النصي الثنائي للمحتوى النصي، وفي حالتنا تعتبر الرسالة (أو التغريدة) إما رسالة عادية (طبيعية normal) أو كراهية (شذوذ anomaly).

يعتبر التصنيف الثنائي أحد المواضيع الأساسية في معالجة اللغات الطبيعية ويوجد العديد من الطرق المعروفة، والتوجه الحالي هو لتقنيات تعلم الآلة لتتعلم النماذج كشف الكراهية من المعطيات أثناء التعلم دون قواعد مسبقة.

**تعريف المسألة**

تعتبر عملية الكشف عن خطاب الكراهية في النصوص المكتوبة من المسائل الصعبة في تعلم الآلة، وذلك بسبب عدة عوامل نذكر منها: نصوص قصيرة، شخصية المؤلف والمتلقي، النوايا غير الواضحة، استخدام العامية، والدمج بين اللغات، الأخطاء الإملائية، ...الخ). ولا سيما عند الاعتماد على السياق ، حيث يمكن أن تختلف معاني الكلمات إلى حد كبير باستخدام الفكاهة والسخرية والتلميحات والاستعارة.

لذلك، من المهم تقديم تعريف واضح وموجز لخطاب الكراهية. ووفقاً للتعريف الذي قدمه المؤلفان [1] فإن خطاب الكراهية هو أي تواصل يستخدم للتعبير عن الكراهية تجاه شخص أو جماعة على أساس بعض الخصائص كالعرق أو الجنس أو التوجه الجنسي أو الجنسية أو الدين أو أي صفة أخرى. وهذا أمر مهم جداً عندما نتعامل مع مجموعة بيانات غير مصنفة حيث لا بد من مشاركة بشرية في عملية تصنيف هذه النصوص.

**تم اعتبار خطاب الكراهية هي حالة الشذوذ التي تتم دراستها. يشمل خطاب الكراهية نشر رسائل للتحريض على العنف أو القتل أو النبذ أو لتشويه سمعة شخص أو مجموعة أشخاص بناءً على العرق، اللون، الجنس، العقيدة أو الدين. كما تم اعتماد التغريدات المدونة في تويتر كمصدر للبيانات المطلوب دراستها.**

**الأعمال المنجزة**

* **بناء قاعدة بيانات:**

تعتبر عملية تأمين البيانات المرحلة الأولى في جميع عمليات التصنيف classification، لذلك تم الاعتماد على خدمة تويتر لتأمين مجموعة كبيرة من التغريدات باستخدام Twitter API. وبغية الحصول على مجموعة بيانات مترابطة ومتوافقة مع الهدف المنشود من الدراسة، تمت عملية البحث من خلال إضافة مجموعة المحددات التالية:

* نقطة إحداثيات: للحصول على بيانات تخص المنطقة الجغرافية الخاصة بسوريا وجوارها.
* اللغة: للحصول على التغريدات المكتوبة باللغة العربية.
* استبعاد المواقع الإخبارية: مؤقتاً وذلك من أجل استبعاد النصوص الإخبارية المتواترة بكثرة.
* **المعالجة الأولية**

تحوي التغريدات رموزاً وأرقاماً بالإضافة لأخطاء إملائية وتكرار لحرف ضمن كلمة (لتعزيز معنى) وكلمات بغير اللغة العربية مثل عناوين URL. تتضمن مرحلة المعالجة الأولية إلغاء الكلمات غير العربية وعلامات الترقيم والأرقام وحذف كلمات لا تضيف معنى وكثيرة التكرار. هذه المرحلة مهمة لأنها تنعكس على المراحل الأخرى فكلما كانت التغريدات أنظف وأوضح كانت أسهل في وضع علامات لها تحدد نوعها عادي أو خطاب كراهية من قبل المصنفين. تضمنت هذه المرحلة الخطوات التالية:

* إبقاء تغريدة واحدة من التغريدات المكررة.
* تصحيح إملائي بسيط يدوي.
* حذف الأجزاء المكتوبة بغير اللغة العربية من التغريدات.
* تحويل (إ أ آ) الى (ا).
* تحويل (ة) الى (ه).
* تحويل (ى) الى (ي).
* حذف علامات الترقيم والأرقام.
* **وسم قاعدة بيانات:**

بعد الحصول على قاعدة البيانات، يجب القيام بوسم قاعدة البيانات وذلك من أجل تصنيفها إما رسالة عادية (طبيعية normal) أو كراهية (شذوذ anomaly) من خلال قراءة هذه التغريدات ومحاولة معرفة طبيعة هذه التغريدة ووضعها في أحد الصفين.

يمكن أن يختلف الأشخاص (المُصَنِّفون) فيما بينهم على اعتبار تغريدة ما من هذا الصنف أو ذاك لعدة اعتبارات. وبما أن عملية التصنيف يمكن أن تؤثر بشكل كبير على النتائج المرجوة، فقد اعتمدنا على تقنية Crowd Sourcing. حيث قام مجموعة من ثلاثة أشخاص بمستوى تأهيل دراسي جيد ومن اتجاهات فكرية مختلفة بقراءة هذه التغريدات وتصنيفها كل على حدة بمعزل عن المصنفين الباقين. بعد الحصول على التقييمات من المصنفين، تم استبعاد التغريدات التي حصلت على تقييمات مختلفة، وتم الإبقاء على التغريدات التي حصلت على إجماع المصنفين.

* **تمثيل قاعدة بيانات:**

تتم هذه العملية عبر استخرج الخصائص feature extraction من خلال تحويل التغريدة إلى تمثيل رقمي، وتستخدم الطريقة التقليدية نموذج حقيبة الكلمات bag-of-words (تُمثَّل كل كلمة بتكرارها في التغريدة).

يتم تحويل كل تغريدة إلى مجموعة من الكلمات المفردة والكلمات المتتالية n-grams، كل عنصر (كلمة، مجموعة كلمات متتالية) هو خاصة مميزة للتغريدة، يتم تحديد قيمتها من خلال تكرار ظهورها في التغريدة يعبر عن هذا التردد بالرمز tf (term frequency).

هناك قيمة أخرى، يتم قياسها فيما يتعلق بالتغريدة، تسمى تردد التغريدة المعكوس ويعبر عنها بالرمز idf (inverse document frequency). تميز هذه القيمة مدى تميز كل خاصة، أي إذا كانت الخاصة موجودة في جميع التغريدات، فلا يمكن التمييز بينها، لذا يجب أن يكون وزنها منخفضًا. عندما تكون الميزة موجودة بالضبط في تغريدة واحدة، يعتبر لك الوضع المثالي لـ خاصية الـ idf، ما سيسمح بالعثور على هذه التغريدة فورًا. من الشائع استخدام تركيبة هاتين القيمتين، tf وidf، والمعروفة باسم قياس tf-idf، وتحدد قيمتها كما هو موضح في الشكل-1.

|  |
| --- |
|  |
| الشكل-1 معادلة tf-idf |

ظهرت حديثًا طرق أخرى لاستخراج الخصائص تستخدم تضمين الكلمات word embeddings، يجيز هذا التمثيل أن يكون للكلمات المتشابهة في المعنى تمثيلات رقمية متشابهة، مما يحسن كفاءة نماذج تعلم الآلة.

يعتبر تدريب متجهات تضمين الكلمات سهلًا لكنه يحتاج لكمية كبيرة من البيانات، وهناك متجهات للكلمات المدربة مسبقًا (قليلة في اللغة العربية) ومن أهمها Aravec وهي متاحة للعموم، دربت Aravec مسبقًا باستخدام بيانات كبيرة (67 مليون تغريدة) من (تويتر، ويكبيديا، ويب) ونفذ بواسطة Word2Vec مع الأخذ بعين الاعتبار (unigram, bigram, trigram). وبأطوال مختلفة لمتجه التضمين، بعد اختبار كل النماذج من Aravec، قررنا تطبيق النموذج full-skip-gram للبعد d=300 (full\_grams\_sg\_300\_twitter.mdl).

كما يوجد نموذج أخر Mazajak مسبق التدريب على بيانات أكبر (250 مليون تغريدة فريدة) ويستخدم word2vec بموديل skip-gram (لم نتمكن من استخدامه في تجاربنا لعدم توفره).

* **تدريب واختبار المصنفات**

يمكن توزيع تقنيات التصنيف السائدة في:

* Naïve Bayes عائلة من الخوارزميات الاحتمالية التي تستخدم نظرية بايز في تصنيف النصوص.
* Linear Regression الانحدار الخطي وهو أسلوب إحصائي للتوقع بقيمة ما (Y) بدلالة مجموعة من الخصائص (X).
* Support Vector Machines أشعة الدعم الآلي وهي نموذج غير احتمالي يستخدم تمثيل النصوص كنقاط في مجال متعدد الأبعاد. تُحدد النصوص في المجال حيث النصوص المنتمية لنفس التصنيف (خطاب كراهية) تقع في أجزاء مستقلة من المجال. ثم نحدد النصوص الجديدة إلى نفس المجال وتتوقع تصنيفها حسب المنطقة التي تقع فيها داخل المجال.
* Deep Learning التعلم العميق وهو مجموعة من الخوارزميات المختلفة التي تحاول تقليد طريقة عمل دماغ الإنسان باستخدام الشبكات العصبية الصناعية لمعالجة البيانات.

تم استخدام مجموعة متنوعة من المصنفات نذكر منها:

* RandomForestClassifier
* SGDClassifier
* SVC
* XGBClassifier
* LinearSVC
* LogisticRegressionCV
* CatBoostClassifier
* KNeighborsClassifier
* RNN-LSTM

1. **التقييم**

* **التحقيقات المتقاطعة cross-validation**

هنالك العديد من الطرق للحصول على معايير أداء لتقييم النموذج وفهم مدى دقته في الكشف. واحدة من أشهر هذه الطرق هي التحقيقات المتقاطعة cross-validation. ما تفعله التحقيقات المتقاطعة هو تقسيم بيانات التعلم training data إلى عدد من الأقسام للتعلم training folds (باستخدام 80% فقط من بيانات التعلم) ونفس عدد الأقسام للاختبار testing folds (باستخدام المتبقي 20% من بيانات التعلم)، ونستخدم أقسام التعلم لتعليم النموذج ونختبر كفاءة النموذج باستخدام أقسام الاختبار لنحصل على معايير لأداء النموذج. نكرر هذه العملية عدة مرات ونأخذ متوسط معايير الأداء.

إذا كانت بيانات الاختبار نفسها في كل مرة، قد يحدث ما يسمى بالتعلم الزائد overfitting لبيانات الاختبار. هذه يعني أننا قد نضبط تجاربنا بصورة مفرطة على البيانات المعطاة لدرجة أنها لا تعمم على بيانات أخرى ويفشل النموذج. تساعد التحقيقات المتقاطعة على منع ذلك، وكلما زادت كمية البيانات كلما تمكنا من زيادة عدد الأقسام.

* **المعايير (الدقة precision والإرجاع recall  والضبط accuracy والمقياس F1)**

تعتبر الدقة والإرجاع والضبط معايير أساسية لتقييم أداء أي نموذج.

تقيس الدقة كم عدد التغريدات التي توقع النموذج تصنيفها صحيحاً خطاب كراهية، من كل التغريدات التي توقع (صحيحة وغير صحيحة) أنها خطاب كراهية.

يقيس الإرجاع كم عدد التغريدات التي توقع النموذج تصنيفها صحيحاً خطاب كراهية، من كل التغريدات التي هي خطاب كراهية كلما زاد حجم بيانات التعلم للنموذج كلما تحسن الإرجاع.

يقيس الضبط كم عدد التغريدات التي توقع النموذج تصنيفها صحيحاً على مستوى كل التغريدات.

يعرف المقياس F1 بعلاقةبين الضبط والإرجاع ويعتمد في التقييم كمرجع أساسي.

كثيراً ما يستخدم الدقة والإرجاع لقياس الأداء لأن الضبط وحده لا يكفي لمعرفة تفاصيل جودة النموذج.

في مشكلة صعبة مثل كشف خطاب الكراهية، غالبًا ما تكون قيم الدقة والإرجاع منخفضة في البداية. لكن كلما غذينا النموذج بالمزيد من البيانات كلما تحسن الأداء. ولكن (كما سنرى لاحقًا) بما أن بيانات التعلم مصنفة بطريقة يدوية فغالباً تحتوي بعض الأخطاء البشرية، لهذا فغالباً لا نرى قيمة مرتفعة للدقة. ولكن إذا استخدمنا بيانات مصنفة يدوياً بصورة متناسقة، سنرى نتائج جيدة مثل أي مشكلة تصنيف في تعلم الآلة، وهذه من تحديات معالجة اللغات الطبيعية.

يبين الشكل-2 علاقات معايير التقييم للنموذج

|  |
| --- |
|  |
| الشكل-2 علاقات معايير التقييم |

* **المنحنيات**

**(roc\_ curve وroc\_auc\_score)**

**(precision\_recall\_curve وprecision\_recall\_auc\_score)**

يعتبر المقياس roc\_auc\_score من المعايير المهمة لتقييم النماذج وهو يمثل المساحة تحت المنحني roc\_curve الممثل للعلاقة بين:

* fpr = the false positive rate (FP / (FP + TN)) for each threshold
* tpr = the true positive rate (TP / (TP + FN)) for each threshold

يظهر المنحنى فعالية المصنف الذي يجري اختباره في ترتيب الحالات الإيجابية بالنسبة للحالات السلبية. تشير النقطة (0, 1) إلى المصنف المثالي، وفيه معدل القيم الإيجابية الصحيحة هو 1، ومعدل القيم الإيجابية الخاطئة هو 0. وبالمثل، تمثل النقطة (1, 1) أن المصنف يتنبأ جميع الحالات بأنها إيجابية بينما تمثل النقطة (0, 0) أن المصنف الذي يتوقع جميع الحالات سلبية. يبين الشكل 3 المنحني وأفضل أداء هو بالقرب من الزاوية اليسارية العليا.

يعتبر المقياس precision\_recall\_auc\_score من المعايير المهمة لتقييم النماذج وهو يمثل المساحة تحت المنحني precision\_recall\_curve الممثل للعلاقة بين الإرجاع والدقة.

يبين الشكل 3 مثال على المنحني roc\_curve والمنحني precision\_recall\_curve

|  |
| --- |
|  |
| الشكل 3 مثال على نتيجة التوقع وprecision\_recall\_curve وroc\_curve |

يبين الشكل 4 قيم التمييز فيroc\_auc\_score

|  |
| --- |
|  |
| الشكل 4 قيم التمييز فيroc\_auc\_score |

1. **تجارب**

معطيات تجميع محلي:

تم إجراء التجارب على مجموعة البيانات التي تم تحصيلها من تويتر ضمن منطقة (بلاد الشام، العراق) خلال الشهرين الأخيرين 2019 وتصنيفها يدويًا (عادي أو خطاب كراهية)، ثم تقسيمها إلى مجموعتين واحدة للتدريب وأخرى للاختبار، وبعد المعالجة الأولية كان لدينا البيانات التالية:

|  |  |  |  |
| --- | --- | --- | --- |
| Type | normal | anomaly | Total |
| Training | 2044 | 89 | 2133 |
| Test | 493 | 69 | 562 |

الجدول- 1توزع البيانات بين التدريب والاختبار

التجربة الأولى:

إجراء تجارب لاستنتاج التمثيل الأفضل للتغريدات ونبين مقارنة بين الطرق الأساسية المعتمدة لتضمين الكلمات بعد تطبيقها على مصنفين عاديين.

ويبين الجدول-2 النتائج حسب Accuracy Score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tf** | **tf-idf** | **Word2vec** | **LDA-td** | **LDA-wd** |  |
| 87.722 | 86.476 | 85.765 | 83.629 | 83.629 | KNN |
| 87.900 | 87.722 | 87.722 | 87.722 | 87.722 | SVM |

الجدول- 2 Accuracy Score

ويبين الجدول-3 النتائج حسب F1 score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tf** | **tf-idf** | **Word2vec** | **LDA-td** | **LDA-wd** |  |
| 0.0 | 11.627 | 14.893 | 4.166 | 4.166 | KNN |
| 8.108 | 0.0 | 0.0 | 0.0 | 0.0 | SVM |

الجدول- 3 F1 score

قمنا بعملية تعزيز بسيطة يدوية لمجموعة التدريب من خلال إضافة تغريدات تمثل خطاب الكراهية أو بمعنى آخر تنتمي إلى الصف anomaly، وذلك من خلال العمليات التالية:

* إضافة تغريدات تمثل خطاب كراهية للمرأة.
* تكرار بعض تغريدات خطاب الكراهية بعد حذف كلمة أو اثنتين من التغريدة الواحدة.
* تكرار بعض تغريدات خطاب الكراهية بعد تبديل ترتيب بعض الكلمات في التغريدة الواحدة.

أصبح لدينا البيانات التالية:

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Normal | anomaly | total |
| Training | 2066 | 148 | 2214 |
| Test | 493 | 69 | 562 |

الجدول- 4 توزع البيانات بين التدريب والاختبار بعد التعزيز

ويبين الجدول - 5 النتائج حسب Accuracy Score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tf** | **tf-idf** | **Word2vec** | **LDA-td** | **LDA-wd** |  |
| 95.907 | 93.772 | 94.128 | 88.434 | 86.298 | KNN |
| 96.975 | 96.619 | 88.078 | 87.722 | 87.722 | SVM |

الجدول- 5 Accuracy Score بعد التعزيز

ويبين الجدول-6 النتائج حسب F1 score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tf** | **tf-idf** | **Word2vec** | **LDA-td** | **LDA-wd** |  |
| 80.672 | 77.987 | 77.852 | 53.237 | 43.795 | KNN |
| 86.821 | 85.271 | 8.219 | 0.0 | 0.0 | SVM |

الجدول- 6 F1 score بعد التعزيز

نستنتج من الجدول أن تمثيل التغريدات LDA-td و LDA-wdلم يعطِنتائج أفضل لا قبل التعزيز ولا بعده

يبين الشكل-1 مقارنة النتائج حسب ROC score على المعطيات قبل التعزيز وبعد التعزيز.

|  |  |
| --- | --- |
|  |  |
| بعد تعزيز مجموعة التدريب Cleantrain2300NRU1.csv | مجموعة التدريب Cleantrain2300NR.csv |
| الشكل-1 مقارنة النتائج حسب ROC score | |

**التجربة الثانية**

تم اعتماد التمثيل full\_grams\_sg\_300\_twitter.mdl من AraVec لتمثيل الكلمات في Word2vecوتدريب عدد من المصنفات ثم اختبارها ويبين الشكل 2 مقارنة بين النتائج حسب ROC score قبل التعزيز وبعده

|  |  |
| --- | --- |
|  |  |
| بعد تعزيز مجموعة التدريب Cleantrain2300NRU1.csv | مجموعة التدريب Cleantrain2300NR.csv |
| الشكل-2 مقارنة النتائج حسب ROC score | |

**التجربة الثالثة**

تم تطبيق RNN-LSTM بالمواصفات المبينة في الجدول 7:

|  |  |  |
| --- | --- | --- |
| Layer (type) | Output Shape | Param# |
| embedding\_3 (Embedding) | (None, 500, 300) | 1500000 |
| dropout\_5 (Dropout) | (None, 500, 300) | 0 |
| conv1d\_3 (Conv1D) | (None, 500, 128) | 192128 |
| max\_pooling1d\_3 (MaxPooling1) | (None, 250, 128) | 0 |
| lstm\_3 (LSTM) | (None, 100) | 91600 |
| dropout\_6 (Dropout) | (None, 100) | 0 |
| dense\_3 (Dense) | (None, 1) | 101 |
| Total params | 1,783,829 |  |
| Trainable params | 1,783,829 |  |
| Non-trainable params | 0 |  |

الجدول- 7 تطبيق RNN-LSTM

وقد حصلنا على النتائج المبينة في الشكل 3

|  |  |
| --- | --- |
| 3-2 | 3-1 |
| 3-4 | 3-3 |
| الشكل-3 (3-1 ،3-2 منحنيات التدريب و3-4،3 3-منحنيات الاختبار) | |
| 4-2 | 4-1 |
| 4-4 | 4-3 |
| الشكل-4 (4-1 ،4-2 منحنيات التدريب و4-4،4 3-منحنيات الاختبار) | |

التجربة الرابعةعلى معطيات جاهزة L-HSAB

​​L-HSAB هي أول مجموعة بيانات Levantine متاحة للجمهور لنظام كشف الكراهية واللغة المسيئة. تهدف مجموعة البيانات المقترحة إلى أن تكون مجموعة بيانات مرجعية للكشف التلقائي عن المحتويات السامة في بلاد الشام على الإنترنت. اعتمدت L-HSAB على تحصيل تغريدات من تويتر وقام 3 من المختصين بتسمية التغريدات يدويًا باتباع مجموعة من القواعد. جمعت 5846 تغريدة وصنفت ضمن 3 فئات: عادي، مسيء وكراهية. تم تحقيق قيم عالية بالاتفاق دون تصحيح واتفاق بين المعلقين مما يدل على موثوقية التعليقات التوضيحية. لاختبارات التصنيف الثنائي جمعت تغريدات مسيء وكراهية (1) وعادي (0) وأجريت التجارب التالية:

تجربة 1

يبين الجدول 8 النتائج حسب Accuracy Score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tf** | **tf-idf** | **Word2vec** | **LDA-td** | **LDA-wd** |  |
| 69.457 | 75.0 | 72.759 | 62.853 | 53.773 | KNN |
| 83.490 | 83.018 | 84.080 | 72.051 | 60.377 | SVM |

الجدول- 8 Accuracy Score

يبين الجدول 9 النتائج حسب F1 score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tf** | **tf-idf** | **Word2vec** | **LDA-td** | **LDA-wd** |  |
| 42.825 | 68.545 | 69.158 | 53.879 | 40.963 | KNN |
| 77.987 | 77.358 | 80.0 | 68.609 | 14.720 | SVM |

الجدول- 9 F1 score

يبين الشكل-5 النتائج حسب ROC score

|  |
| --- |
|  |
| الشكل5 النتائج حسب ROC score |

التجربة 2

اعتمد النموذج full\_grams\_sg\_300\_twitter.mdl من AraVec لتمثيل الكلمات في Word2vec ويبين الشكل 6 نتائج تطبيق معطيات الاختبار على عدة مصنفات مدربة على معطيات التدريب

|  |  |
| --- | --- |
|  | MacAvg. 80.71% F1. 74.96% P. 86.77 R. 65.98 Acc. 82.43%: RandomForestClassifier  MacAvg. 82.98% F1. 78.88% P. 83.01 R. 75.15 Acc. 83.96%: SGDClassifier  MacAvg. 39.25% F1. 59.66% P. 42.52 R. 100.00 Acc. 46.11%: SVC  MacAvg. 82.37% F1. 77.92% P. 83.45 R. 73.08 Acc. 83.49%: XGBClassifier  MacAvg. 81.82% F1. 77.38% P. 81.85 R. 73.37 Acc. 82.90%: CatBoostClassifier  MacAvg. 83.29% F1. 79.38% P. 82.69 R. 76.33 Acc. 84.20%: LogisticRegressionCV |
| ROC score | MacAvg, F1, Precision, Recall, Accuracy Score |
| الشكل-6 النتائج | |

التجربة 3

تطبيق RNN-LSTM بالمواصفات المبينة في الشكل 3 وتكون المرحلة الأولى Embedding تمثيل عشوائي للكلمات وثم تدريب من رسائل التدريب للوصول لتمثيل الكلمات رقميًا ونتائج التدريب والاختبار مبينة في الشكل 7

|  |  |
| --- | --- |
| 7-2 | 7-1 |
| 7-4 | 7-3 |
| الشكل-7 (7-1 ،7-2 منحنيات التدريب و7-4،7 3-منحنيات الاختبار) | |

وعند اعتماد النموذج full\_grams\_sg\_300\_twitter.mdl من AraVec لتمثيل الكلمات تحسنت نتيجة الاختبار كما هو مبين في الشكل 8

|  |  |
| --- | --- |
| 8-2 | 8-1 |
| 8-4 | 8-3 |
| الشكل-8 (8-1 ،8-2 منحنيات التدريب و8-4،8 3-منحنيات الاختبار) | |

**تجربة التحقيقات المتقاطعة cross-validation**

في تقييم المصنف RandomForestClassifier طبقت 4 طرق للتحقيقات المتقاطعة على البيانات بعد التعزيز ويبين الجدول التالي المنحنيات لبيانات (validation, test)

وللمقارنة طبقت الطريقة الرابعة على البيانات قبل التعزيز وتمثلها منحنيات السطر الأخير**.**

|  |  |  |
| --- | --- | --- |
|  |  | KFold(n\_splits=5)  الطريقة 1: (K-fold Cross-Validation) وهي تقسيم البيانات إلى (K) قسم عادة 5 أو 10. عملية التدريب تتم على (K-1) من هذه الأقسام والاختبار على القسم المتبقي. |
|  |  | StratifiedKFold(n\_splits=5,random\_state=100)  الطريقة 2 : (Strafified) وهي تشبه الطريقة 1 فقط الأقسام التي نحصل عليها من بيانات التدريب تحتوي على (labels) بشكل متوازن، يعني كل قسم يحتوي بالتقريب على أمثلة لهم (labels) متنوعة بالتقريب قدر تنوع كل (labels) الموجودة. |
|  |  | الطريقة 3 : (Leave One Out) في هذا الطريقة وبعد تقسيم البيانات إلى (K) قسم فإن كل قسم يحتوي على عنصر واحد فقط وبالتالي (K) يكون بعدد الأمثلة الموجودة في البيانات. كل مرة تتغير مجموعة الاختبار حتى تشارك جميع أمثلة البيانات في الاختبار وهي مكلفة للغاية وتستغرق وقتًا طويلًا أثناء التنفيذ ولكنها فعالة عندما تكون البيانات صغيرة. |
|  |  | ShuffleSplit(n\_splits=5,random\_state=100)  الطريقة 4 : (Repeated Random Test-Train Splits  وهي مزيج بين الطريقة العادية والطريقة 1 (K-fold ) وتعتمد على تقسيم البيانات إلى قسم للتدريب وقسم للاختبار وتستمر في تقسيم البيانات بشكل عشوائي وإجراء التدريب والاختبار في كل مرة بعدد محدد. |
|  |  | ShuffleSplit(n\_splits=5,random\_state=100)  الطريقة 4 : (Repeated Random Test-Train Splits )  البيانات قبل التعزيز |

الجدول- 10 التحقيقات المتقاطعة cross-validation

**الأعمال القادمة:**

**نشر مقالة:**

* سيتم العمل على نشر مقالة بخصوص الاختبارات والنتائج التي تم تحصيلها وذكرها سابقاً**.**

**المحولات Transformers:**

* تعرف المحولات بأنها نموذج تعلم عميق ظهر في عام 2017، واستخدم بشكل أساسي في مجال معالجة اللغات الطبيعية NLP، وهي على غرار الشبكات العصبونية المتكررة recurrent neural networks (RNNs) مصممة للتعامل مع البيانات التسلسلية كاللغات الطبيعية. ولكن تختلف هذه المحولات عن الشبكات RNNs في أنها لا تتطلب معالجة البيانات التسلسلية بالترتيب، ما يتيح للمحولات العمل على التوازي ما يعني توفيراً كبيراً في وقت التدريب، وبالتالي إمكانية معالجة بيانات ضخمة.
* كذلك سيتم تطبيق نماذج المحولات التالية: mBERT, XLM-RoBERTa, AraBERT, GigaBERT. وسيتم العمل على تطوير أحد هذه النماذج.