دراسة تقنيات استخراج الميزات المستخدمة في تحليل المشاعر

طالب ماجستير: أيهم جمال شعار

علوم الويب في الجامعة الافتراضية السورية

كلية الهندسة المعلوماتية–قسم البرمجيات–جامعة البعث

المشرف: د. محمد مازن المصطفى

أستاذ في الهندسة المعلوماتية–قسم البرمجيات–الجامعة الدولية للعلوم والتكنلوجيا

الملخص:

أدّى ظهور الويب2.0 إلى تغيير طريقة تواصل الأفراد وتشجيعهم على مشاركة آرائهم وأفكارهم وخبراتهم في المدوّنات والمنتديات والشبكات الاجتماعية. تشكّل هذه البيانات معلومات قيّمة لكل من الأفراد والمؤسسات وتساعدهم في اتخاذ قراراتهم. يحتاج هذا الكم الهائل من البيانات إلى تقنيات خاصّة لمعالجتها وتحليلها تُعرف عملية استخدام معالجة اللغات الطبيعية واللغويات الحاسوبية للكشف عمّا يحمله النص من مشاعر باسم تحليل المشاعر. يعتمد تحليل المشاعر على أربع مهام رئيسية وهي: تحديد الرأي، واستخراج الميزات، ومن ثم تصنيف المشاعر لتحديد قطبية الرأي، وأخيراً عرض النتائج وتلخيصها. يعدّ استخراج الميزات أحد أكثر المهام تعقيداً في تحليل المشاعر. نجري في هذه الورقة البحثية دراسة حول أداء أربع تقنيات لاستخراج الميزات من مجموعة بيانات نصيّة وهي BOW وTF-IDF وWord2vec وBERT. استخدمنا الميزات المستخرجة باستخدام هذه التقنيات لتدريب واختبار خمس مصنّفات تعلّم آلي وهي: خوارزميات الانحدار اللوجستي وآلات متجه الدعم وبايز وشجرة القرار والغابة العشوائية. قمنا بتقييم أداء هذه المصنّفات مع جميع التقنيات لاكتشاف أثر تقنيات الاستخراج عليها والمقارنة بينها وفق عدة مقاييس تقييم.

الكلمات المفتاحية: تحليل المشاعر، استطلاع الآراء، استخراج الميزات، خوارزميات التصنيف، مقاييس التقييم.

Study of Feature Extraction Techniques Used in Sentiment Analysis

Abstract:

The development of Web 2.0 has changed the way people communicate and encourage them to share their opinions, ideas, and experiences in blogs, forums, and social networks. This data is valuable information for both individuals and organizations and helps them to make decisions. This huge amount of data needs special techniques to process and analyze. These techniques are known as sentiment analysis techniques. Sentiment analysis relies on four main tasks: defining opinion, extracting features, then classifying sentiments to determine opinion polarity, and finally visualizing and summarizing the results. Feature extraction is one of the most complex tasks in sentiment analysis. In this paper, we made a study about the performance of four feature extraction techniques, which are BOW, TF-IDF, Word2vec and BERT. We used the features extracted using these techniques to train and test five machine-learning classifiers: logistic regression (LR), support vector machine (SVM), Naïve Bayes (NB), decision tree (DT), and random forest (RF). We evaluated the performance of these classifiers with all techniques to discover the impact of extraction techniques on them and compare them according to several evaluation metrics.

Keywords: sentiment analysis, opinion mining, feature extraction, classification algorithms, evaluation metrics.

1-المقدمة

أدّى نمو تطبيقات الويب الحديثة مثل Facebook وTwitter وسهولة استخدامها إلى إقناع الناس بالتعبير عن آرائهم حول المنتجات والخدمات والقضايا المختلفة. تشكّل هذه الآراء مصادر قيّمة للمعلومات التي تؤثّر على حياتنا وكذلك القرارات التي نتخذها. على سبيل المثال، نقوم بتثبيت تطبيق أو استخدام خدمة معينة إذا كانت مشاعر الناس إيجابية تجاهها. هذه الكمية الهائلة من البيانات غير منظمة وذات أبعاد عالية وبالتالي تتطلب الكثير من الجهد لتحليلها باستخدام تقنيات معالجة اللغات الطبيعية واستخراج النصوص. تُعرف عملية استخدام معالجة اللغات الطبيعية واللغويات الحاسوبية للكشف عما يحمله النص من مشاعر باسم تحليل المشاعر وهو عملية تصنيف تهدف إلى تحديد ما إذا كان النص إيجابياً أم سلبياً حول موضوع ما. تؤثر العديد من العوامل على عملية تحليل المشاعر. تحتاج خوارزميات التعلّم الآلي إلى مجموعة الميزات التي تحمل أكبر قدر من المعلومات حول خصائص البيانات من أجل تحسين أداء نموذج التعلّم الآلي وتقليل التعقيد الحسابي، لذا فإن استخراج الميزات يلعب دوراً مهماً في تحليل المشاعر. تتضمن الدراسة المقترحة مقارنة بين أداء أربع تقنيات وهي: تقنية حقيبة الكلمات(BOW) وتقنية TF-IDF وتقنيةWord2vec وتقنيةBERT. تُستخدم هذه التقنيات لاستخراج الميزات من البيانات النصية ومن ثم تُستخدم هذه الميزات لتدريب خوارزميات التصنيف من أجل تحديد قطبية النص.

2-هدف البحث:

يهدف هذا البحث إلى دراسة تأثير تقنيات استخراج الميزات على أداء خمسة من الخوارزميات المستخدمة في تصنيف المشاعر من خلال إجراء مقارنة تجريبية على مجموعة بيانات حقيقية بين أربع تقنيات لاستخراج الميزات ومعرفة مدى تأثير اختيارها على أداء خمس خوارزميات تعلّم آلي من خلال تقييم الأداء وفق عدة مقاييس تقييم.

3-الدراسات السابقة

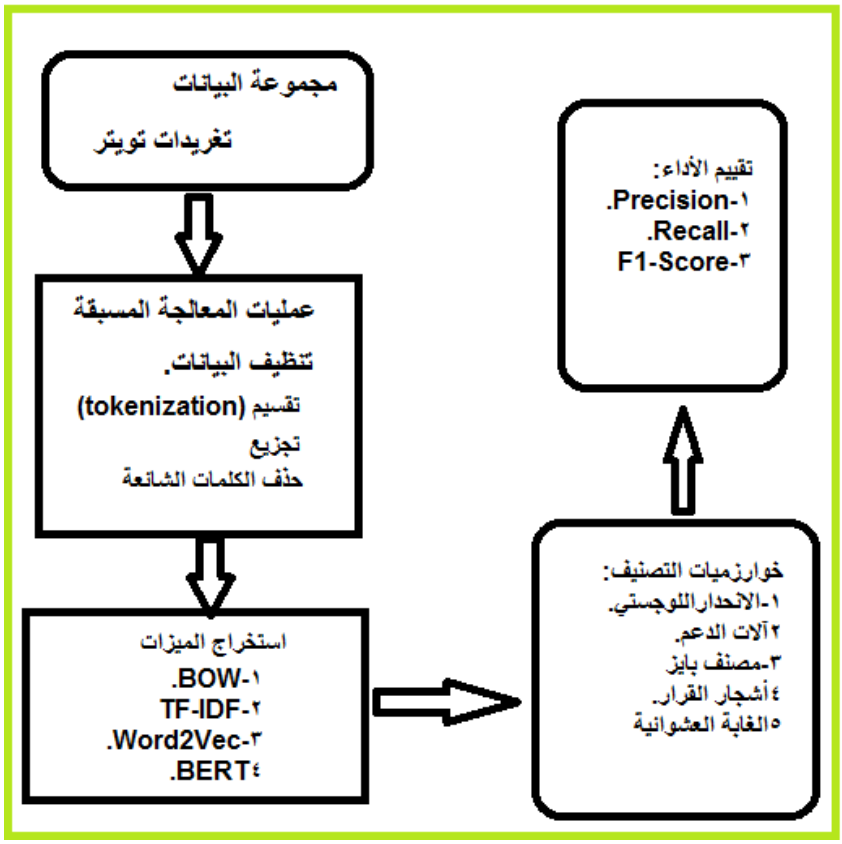
أجرى الباحثون في [1] تنقيباً عن الآراء في تعليقات(reviews) الطلاب باستخدام الخوارزميات الخاضعة للإشراف (K-NN وNN وNB وSVM). بعد المعالجة المسبَقة، قاموا بتوليد تمثيل شعاعي للكلمات باستخدام TF-IDF. تمّ استخراج الميزات الرئيسية في مجال التعليم وتصنيفها حسب قطبيتها. أظهرت النتائج تفوّق مصنّف NB من حيث مقياسي الاسترجاع(Recall) والدقة(Accuracy) بينما تفوّق مصنّف K-NN من حيث مقياس Precision. في [2] أجرى الباحثون تحليلاً للمشاعر اعتماداً على تعليقات على منتجات مختلفة من أمازون. أولاً قاموا بتطبيق المعالجة المسبقة للنص ثم قاموا بتوليد تمثيل شعاعي للنص باستخدام حقيبة الكلمات. قاموا بحساب قطبية كل جملة باستخدام قواميس الرأي ثم تطبيق الخوارزميات (SVM، NB). أظهرت النتائج تفوق NB على SVM من حيث مقياس الدقة في التعليقات على منتجات الكاميرا. نفّذ الباحثون في [3] تحليلاً للمشاعر على مجموعة بيانات كبيرة من أمازون. في البداية، تم تصنيف مجموعة البيانات بطريقة آلية، ثم قاموا بالمعالجة المسبقة للبيانات، بعدها تم استخراج الميزات بعدة طرق، وهي حقيبة الكلمات وTF-IDF، ثم طبقوا عدة مصنّفات. أظهرت النتائج حصول مصنف SVM على أفضل دقة في التعليقات على منتجات الآلات الموسيقية. استعرض الباحثون في [4] العديد من الدراسات التي استخدمت نموذج word2vec في تحليل المشاعر. معظم الدراسات استخدمت النموذجين CBOW وSkip-Gram، وقارنت نتائجهما. يعد Skip-gram أفضل للكلمات النادرة من CBOW؛ بينما يعد CBOW أسرع ويعمل بشكل جيد مع الكلمات المتكررة. نفّذ الباحثون في [5] تحليلاً للمشاعر على رسائل تويتر باستخدام نموذج word2vec، وقارنوا نتائج نموذجين مختلفين: 1) حقيبة الكلمات المستمرة (CBOW) و2) تخطي الجرام (SG)، أظهرت النتائج أن نموذج SG تفوق في الأداء على نموذج CBOW مع جميع المصنفّات. أعلى دقة تمَّ الحصول عليها كانت باستخدام مصنف آلات الدعم مع نموذج SG.طوّر الباحثون في [6] نهجاً لتحليل المشاعر في بيانات الشبكة الاجتماعية

باستخدام نموذجي Word2vec وBERT كطرق لاستخراج الميزات. بيّنت النتائج أن أعلى دقّة تم الحصول عليها كانت عند استخدام نموذج BERT.وجد الباحثون في [7] أنّ دمجَ متجهات الكلمات الثابتة التي تم تدريبها بشكل خاص على وسائل التواصل الاجتماعي يمكن أنْ يُحسّنَ أداء نماذج اللغة القائمة على BERT للغات منخفضة الموارد، حيثُ من غير المرجّح أنْ تُصبح نماذج اللغة المتخصصة (على سبيل المثال المدربة على وسائل التواصل الاجتماعي) متاحة لمثل هذه اللغات. قارن الباحثون في [8] بين طريقتين لاستخراج الميزات وهما TF-IDF وN-Gram من حيث تأثيرهما على أداء ستة خوارزميات تعلّم آلي في عملية تصنيف المشاعر حيث وجدوا أن استخدام TF-IDF زاد الأداء بنسبة 3-4%.

4-المنهجية المقترحة

تم إجراء هذه الدراسة على مجموعة بيانات حصلنا عليها من موقع Kaggle[1]. تحتوي قاعدة البيانات حوالي 14000 تغريدة عن شركات الطيران من مدونة تويتر. تمَّ تصنيف هذه التغريدات إلى ثلاث فئات (إيجابية، سلبية، ومحايدة). في البداية، قمنا بتحليل التغريدات وطبقنا عليها مجموعة من تقنيات المعالجة المسبقة لتنظيفها وإعدادها لمرحلة استخراج الميزات. بعد ذلك، قسمنا قاعدة البيانات إلى مجموعة تدريب ومجموعة اختبار. قمنا باستخراج الميزات من التغريدات باستخدام أربع تقنيات مختلفة ومن ثم قمنا بتدريب خمس خوارزميات للتعلم الآلي من أجل تصنيف قطبية التغريدات. بعد ذلك قمنا بتقييم أداء هذه المصنفات على مجموعة الاختبار مع جميع تقنيات الاستخراج لمقارنة تأثير هذه التقنيات على عملية تحليل المشاعر باستخدام عدة مقاييس لتقييم الأداء. يوضح الشكل (1) خطوات العمل.

[1] https://www.kaggle.com/crowdflower/twitter-airline-sentiment



الشكل 1 مخطط العمل

5-المعالجة المسبقة للبيانات

تهدف المعالجة المُسبَقة للبيانات إلى تسهيل عملية تدريب واختبار المصنّفات من خلال إجراء مجموعة تحويلات مناسبة على البيانات. تعتبر المعالجة المسبقة للبيانات عملية ضرورية قبل تدريب نماذج التعلّم الآلي؛ حيث تقوم بإزالة القيم المتطرفة والبيانات غير الضرورية من مجموعة البيانات بالإضافة لتقليل الأبعاد. بدأنا المعالجة المسبقة عن طريق تقسيم التغريدات على أساس الفراغات ومن ثم أزلنا بعض علامات التصنيف المرتبطة بالتغريدات مثل@ و#. قمنا أيضاً بإزالة الروابط التشعبية والأرقام وعلامات

الترقيم (احتفظنا بالأرقام وبعض علامات الترقيم مثل النقاط وعلامات الاستفهام عند استخدام نموذج BERT لتمييز نهاية الجمل)، وقمنا بتطبيق التجزيع(stemming) باستخدام خوارزمية بورتر (Porter Stemmer) وتحويل جميع الكلمات إلى أحرف صغيرة بالإضافة إلى حذف الفراغات الزائدة.قمنا بعمليات المعالجة المسبقة باستخدام مكتبة NLTK في لغة بايثون وبذلك أصبحت البيانات نظيفة وجاهزة لاستخراج الميزات.

6-استخراج الميزات:

قمنا باستخدام أربع تقنيات لاستخراج الميزات من البيانات النصية وهي:

- **حقيبة الكلمات (Bag of words)**

في نموذج حقيبة الكلمات (BOW)، يتم تمثيل النص كمجموعة غير مرتبة من كلماته دون مراعاة القواعد النحوية أو ترتيب الكلمات. يتم إعطاء وزن لكل كلمة في المستند وفقاً لتكرارها في المستند وتكرارها في المستندات الأخرى في مجموعة البيانات. يعتبر من أبسط نماذج التمثيل المستخدمة في معالجة اللغة الطبيعية واسترجاع المعلومات. لتطبيق طريقة BOW قمنا باستخدام CountVecotorizer من مكتبة sklearn في لغة Python.

- **نموذج TF-IDF (Term Frequency-Inverse Document frequency)**

TF-IDF هو مقياس ترجيح يقوم بحساب أهمية الكلمة في مستند نصي. هو عبارة عن جداء مقدارين إحصائيين [9]:

1. تكرار الكلمة (**Term Frequency**): هو عدد المرات التي يتكرر فيها مصطلح معين t في مستند، وهو نسبة تكرار الكلمة في المستند على إجمالي عدد الكلمات في ذلك المستند.

2. تردد المستند العكسي **(Inverse Document frequency)**: يستخدم لقياس أهمية كلمة معينة في مستند نصي. يكون تكرار المستند العكسي للكلمة منخفضاً إذا تكررت في العديد من المستندات ويكون مرتفعاً إذا وردت الكلمة في مستند واحد فقط. أي أنه يعطي أهمية أكبر للكلمات النادرة في المستند. يتم حساب الوزن النهائي للكلمة t في الوثيقة d حسب المعادلة التالية:

$$TF - IDF(t,d) = TF(t,d) * IDF(t)$$

لتطبيق تقنية TF-IDF؛ استخدمنا الصف Tfidfvectorizer من مكتبة sklearn وقمنا بضبطه باستخدام المعلمات: ngram_range، max_features.

❖ **نموذج word2vec**

يعتبر Word2Vec نموذجاً تنبؤياً لحساب التمثيل الشعاعي المستمر للكلمات في مجموعات البيانات الكبيرة [10]. يستخدم النموذج المقترح شبكات عصبية ضحلة(shallow) للحصول على متجه عالي الأبعاد لكل كلمة ويتضمن نموذجين بديلين. 1) حقيبة الكلمات المستمرة (CBOW) هي طريقة للتنبؤ بالكلمات باستخدام سياق الكلمات المجاورة لها 2) في المقابل، يستخدِم Skip-gram معلومات الكلمة للتنبؤ بالكلمات المجاورة لها. يحتوي كلا النموذجين على ثلاث طبقات: طبقة إدخال وطبقة إسقاط وطبقة إخراج. لإنشاء متجهات الكلمات باستخدام word2Vec استخدمنا مكتبة gensim حيث تمَّ تطبيقه في نموذجين مختلفين: CBOW وSkip-Gram وقمنا بمقارنة النتائج. تم ضبط Word2Vec باستخدام المعلمات: min_count وsize وwindow. استخدمنا متوسط متجهات الكلمات في التغريدة لتمثيل التغريدة بمتجه واحد فقط.

❖ **نموذج BERT (Bidirectional Encoder Representations from Transformers)**

BERT هو نموذج لغة مدرَّب مسبقاً مصمَّم للتدريب المسبق على التمثيلات العميقة ثنائية الاتجاه (deep bidirectional) من نص غير مصنّف عن طريق التكييف المشترك على كل من السياق الأيمن والأيسر في جميع الطبقات [11]. يمكن ضبط BERT بدقة باستخدام طبقة إخراج إضافية واحدة فقط لإنشاء نماذج حديثة لمجموعة واسعة من المهام، مثل الإجابة على الأسئلة والاستدلال اللغوي، دون تعديلات معمارية خاصة بالمهمة المحددة. قمنا باستخدام BERT بطريقتين: أولاً قمنا بتقسيم كل تغريدة إلى كلمات ثم حساب تضمين BERT لكل كلمة ومن ثم حساب متوسط جميع متجهات الكلمات كتمثيل للتغريدة. ثانياً، قمنا بتقسيم بيانات التدريب إلى جمل وحساب متوسط تضمين BERT للجمل لتمثيل كل بيانات التدريب.

7-خوارزميات التصنيف:

قمنا باستخدام خمس خوارزميات تصنيف وتطبيقها باستخدام مكتبة sklearn في لغة بايثون وهي:

❖ **خوارزمية الانحدار اللوجستي (Logistic Regression):**

إحدى أقدم طرق التصنيف، تم تطويرها بواسطة الإحصائي ديفيد كوكس عام 1958. يُستخدم الانحدار اللوجستي للتنبؤ باحتمالية وقوع حدث عن طريق إدخال البيانات لتابع لوجستي حيث يتنبأ بالاحتمالات بدلاً من الفئات [12]. يتم نمذجة الاحتمالات التي تصف نتائج التجربة باستخدام الانحدار اللوجستي. تسمى هذه الخوارزمية أيضاً بالانتروبي الأعظم. تمَّ تطوير التابع اللوجستي لوصف خصائص النمو السكاني في البيئة حيث ترتفع بسرعة وتتجاوز القدرة الاستيعابية للبيئة.

❖ خوارزمية آلات متجه الدعم (Support Vector Machines)

تم تطوير خوارزمية متجه الدعم بواسطة Cortes و Vapnik للعمل في مجال التصنيف الثنائي عام 1995 [13]. يمكن وصف طريقة عملها على النحو التالي: المهمة هي فصل الفئات حيث يتم البحث عن المستوى الأمثل للفصل المفرط بين الفئتين من خلال تعظيم الهامش بين النقاط الأقرب للفئات –تسمى النقاط الموجودة على الحدود متجهات الدعم، ووسط الهامش هو مستوى الفصل الأمثل.

❖ خوارزمية بايز (Naïve Bayes)

خوارزمية تعلّم خاضعة للإشراف وهي خوارزمية رياضية تعتمد على مبدأ الإحصاء (الاحتمالات) [14]. مصنف Naïve Bayes عبارة عن مجموعة من المصنفات الاحتمالية البسيطة يعتمد على الفرضية العامة التي مفادها أن جميع السمات مستقلة عن بعضها البعض. يعتبر المصنف الأساسي في فلترة النصوص والمستندات. سهولة تدريب النموذج باستخدام هذه الخوارزمية، خاصة عند التعامل مع البيانات الضخمة، بالإضافة إلى بساطته أثبتت كفاءته مقارنة ببعض الخوارزميات المتقدمة. تُستخدم هذه الخوارزمية في اكتشاف المشاعر وكذلك في فلترة البريد الإلكتروني.

❖ خوارزمية شجرة القرار (Decision Tree)

تعد شجرة القرار من أقوى الخوارزميات المستخدمة في استخراج البيانات، يمكنها التعامل مع أنواع مختلفة من البيانات، وهي عبارة عن شجرة موجهة مع عقدة تسمى الجذر ليس لها حواف واردة [15]. جميع العقد الأخرى لها حافة واردة واحدة بالضبط. العقدة ذات الحواف الصادرة تسمى عقدة داخلية. تسمى جميع العقد الأخرى أوراق أو عقد القرار. تقسم كل عقدة داخلية مساحة المثيل إلى مساحتين فرعيتين أو أكثر وفقاً لتابع منفصل محدد لقيم سمات الإدخال. يتم تخصيص كل ورقة لفئة واحدة تمثل القيمة المستهدفة

الأكثر ملاءمة. يتم تصنيف المثيلات عن طريق التنقل بينها من جذر الشجرة وصولاً إلى الورقة، وفقاً لنتائج الاختبارات على طول المسار.

❖ **خوارزمية الغابة العشوائية (Random Forest)**

يتكون مصنف الغابة العشوائية من مجموعة من المصنفات ذات الهيكلية الشجرية حيث يتم إنشاء كل مصنف باستخدام متجه عشوائي يتم أخذ عينات منه بشكل مستقل عن متجه الإدخال وتصوّت كل شجرة على الفئة الأكثر شيوعاً لتصنيف متجه الإدخال [16]. تحتفظ الغابة العشوائية بالعديد من فوائد أشجار القرار مع تحقيق نتائج أفضل من خلال استخدام ملء العينة (sample filling)، ومجموعات عشوائية من المتغيرات، ونظام تصويت الأغلبية (majority voting)، ولا تتطلب تقنيات تقليم الأشجار (pruning techniques).

8-النتائج التجريبية

قمنا بمقارنة أداء أربع تقنيات لاستخراج الميزات وهي BOW وTF-IDF وWord2Vec وBERT. أجريت التجربة على مجموعة بيانات حقيقية تحتوي على تغريدات عن شركات الطيران من تويتر (9178 تغريدة سلبية، 2363 إيجابية، 3099 محايدة). قمنا بتقسيم مجموعة البيانات بنسبة 70٪ لتدريب المصنّفات و30٪ لاختبارها بطريقة Hold-out، بعد تطبيق عمليات المعالجة المسبقة على البيانات باستخدام تقنيات معالجة اللغات الطبيعية، بدأنا عملية استخراج الميزات باستخدام التقنيات الأربعة لتوليد المتجهات الرقمية ومن ثم تُستخدم هذه الميزات لتدريب خمس خوارزميات تصنيف لتحديد قطبية النص.

من أجل مقارنة أداء تقنيات استخراج الميزات مع خوارزميات التصنيف، استخدمنا عدة مقاييس تقييم أخذين بالاعتبار أن قاعدة البيانات غير متوازنة. استخدمنا مقاييس: Precision وRecall وF-Score. يتم استخدام مقياس precision لقياس نسبة

الأصناف الإيجابية التي تمّ توقعها بشكل صحيح من إجمالي الأصناف المتوقعة في فئة إيجابية ويحسب من مصفوفة الارتباك الشكل(2) وفق المعادلة التالية:

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

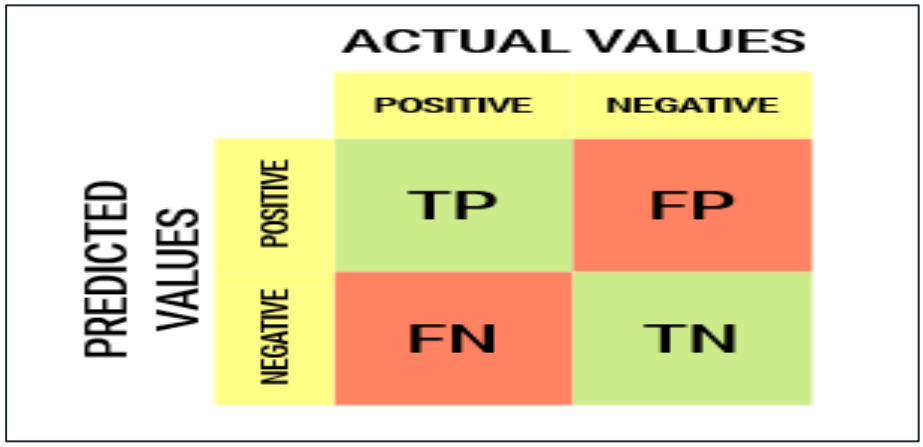
يتم استخدام مقياس Recall لقياس النسبة المئوية للأصناف الإيجابية التي تم توقعها بشكل صحيح من جميع الأصناف الإيجابية الفعلية ويحسب وفق المعادلة التالية:

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

مقياس F1−Score هو المتوسط التوافقي لمقياسي Precisionو Recallويحسب وفق الصيغة التالية:

$$F - score = 2(\frac{Precision * Recall}{Precision + Recall})$$

تم تطبيق هذه المقاييس برمجياً في لغة بايثون باستخدام مكتبة sklearn.



الشكل 2 مصفوفة الارتباك

حيث(TP[2]) تعني القيمة المتوقعة تطابق القيمة الفعلية أي كانت القيمة الفعلية إيجابية وتوقع النموذج قيمة إيجابية و (TN[3]) تعني القيمة المتوقعة تطابق القيمة الفعلية أي كانت القيمة الفعلية سلبية وتنبأ النموذج بقيمة سلبية. بينما (FP[4]) تعني أنّه تم التنبؤ بالقيمة المتوقعة بشكل خاطئ أي كانت القيمة الفعلية سلبية لكن النموذج تنبأ بقيمة إيجابية. كذلك(FN[5]) تعني أنّه تم التنبؤ بالقيمة المتوقعة بشكل خاطئ أي كانت القيمة الفعلية إيجابية لكن النموذج توقع قيمة سلبية.

توضح الجداول (1) و (2) و (3) المتوسط المرجّح لمقاييس Precision و Recall و F–Score لجميع المصنّفات لكل تقنية من تقنيات استخراج الميزات المستخدمة في التجربة.

الجدول1 المتوسط المرجح لل Precision مع جميع المصنفات

Classifier	BOW	TF-IDF	W2V (CBOW)	W2V Skip-Gram	Bert-Word	Bert-Sentence
LR	**0.77**	0.79	0.68	**0.73**	**0.79**	**0.81**
SVM	0.76	**0.80**	**0.70**	**0.73**	**0.79**	0.79
NB	0.75	0.76	0.60	0.66	0.72	0.78
DT	0.70	0.70	0.60	0.64	0.63	0.70
RF	0.75	0.77	0.68	**0.73**	0.74	0.80

[2] True Positive
[3] True Negative
[4] False Positive
[5] False Negative

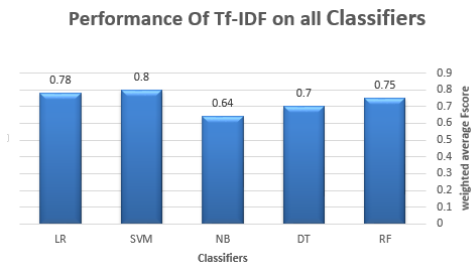
الجدول2 المتوسط المرجح لل Recall مع جميع المصنفات

Classifier	BOW	TF-IDF	W2V (CBOW)	W2V Skip-Gram	Bert-Word	Bert-Sentence
LR	**0.78**	0.79	0.69	**0.75**	0.79	**0.81**
SVM	0.77	**0.81**	0.69	0.74	**0.80**	0.80
NB	0.76	0.70	0.62	0.68	0.72	0.74
DT	0.70	0.69	0.59	0.64	0.63	0.70
RF	0.76	0.77	**0.71**	**0.75**	0.75	**0.81**

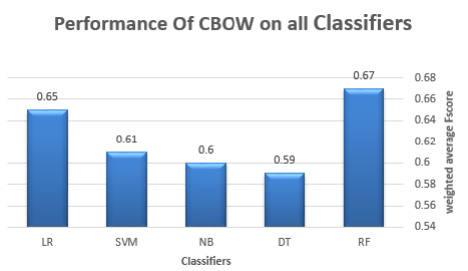
الجدول 3 المتوسط المرجح ل F1-Score مع جميع المصنفات

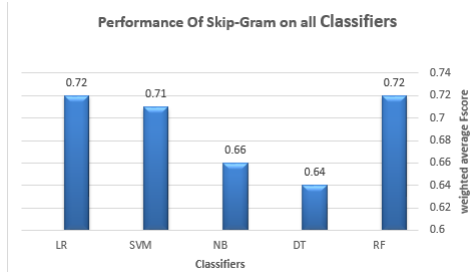
Classifier	BOW	TF-IDF	W2V (CBOW)	W2V Skip-Gram	Bert-Word	Bert-Sentence
LR	**0.78**	0.78	0.65	**0.72**	0.79	**0.81**
SVM	0.77	**0.80**	0.61	0.71	**0.80**	0.79
NB	0.73	0.64	0.60	0.66	0.72	0.75
DT	0.70	0.70	0.59	0.64	0.63	0.70
RF	0.75	0.75	**0.67**	**0.72**	0.71	0.80

يوضح الشكل (3) مخططات مقياس f-score لجميع المصنّفات عند استخدام نموذجي BOW وTF-IDF. يوضح الشكل (4) مخططات مقياس f-score لجميع المصنفات باستخدام نموذجي CBOW وSkip-gram. يوضح الشكل (5) مخططات مقياس f-score لجميع المصنفات باستخدام نموذجي BERTبناءً على قيم المتوسط المرجَّح لمقياس f-score من الجدول (3).

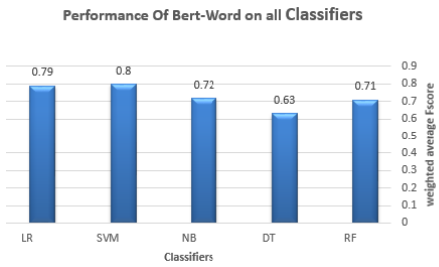


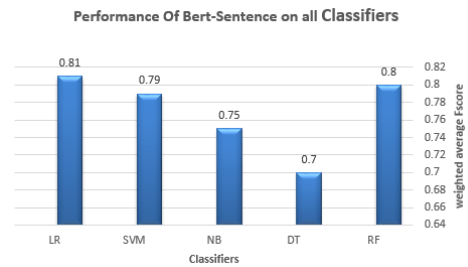
الشكل 3 نتائج درجة F1 لنموذجي BOW و TF-IDF مع جميع المصنفات





الشكل 4 نتائج درجة F1 لنموذجي CBOW و Skip-Gram مع جميع المصنفات





الشكل 4 نتائج درجة F1 لنموذجي Bert-sentence وBert-word مع جميع المصنفات

9-تحليل النتائج

1. بالنسبة لمقياس Precision: حقق نموذج Bert sentence أعلى أداء مع مصنف الانحدار اللوجستي (LR).
2. بالنسبة لمقياس Recall: تم تحقيق أعلى محصلة عند استخدام نموذج TF-IDF مع مصنف SVM وكذلك باستخدام نموذج Bert sentence مع المصنفات LR و RF.
3. بالنسبة لمقياس F-score: حقّق نموذج Bert sentence أعلى أداء مع مصنّف LR.
4. كان أداء TF-IDF أفضل من أداء BOW مع مصنّف SVM، بينما كان أداء BOW أفضل من أداء TF-IDF مع مصنّف NB. حصل النموذجان على أداء مشابه مع المصنّفات LR وDT و RF.
5. تفوّق نموذج TF-IDF على نموذج CBOW مع جميع المصنفات. بالإضافة إلى ذلك، كان أداء TF-IDF أفضل من Skip-Gram مع جميع المصنفات باستثناء مصنّف Naïve Bayes.
6. تفوّق نموذج Skip-Gram على نموذج CBOW مع جميع المصنفات ويرجّح هذا التفوق إلى كون نموذج Skip-Gram أفضل في حالة قواعد البيانات الصغيرة.
7. كان أداء نموذج Bert sentence أفضل من Bert-word مع جميع المصنفات باستثناء SVM حيث كان أداء Bert-word أفضل.
8. كان أداء نموذج Bert sentence أفضل من أداء TF-IDF مع المصنفات: LR وNB وRF بينما كان أداء TF-IDF أفضل مع مصنف SVMو حققوا أداءً مماثلاً مع مصنف DT.

9. حقق نموذج Bert sentence أداءً أفضل من أداء نموذجي (CBOW وSG) مع جميع المصنفات. كان أداء نموذج Bert-word أفضل من skip-gram مع المصنفات: LR وSVM وNB بينما كان أداء Skip-gram أفضل من Bert-word مع مصنفات DT و RF.

بعد تحليل أداء جميع المصنفات مع جميع تقنيات الاستخراج، نستنتج أن مصنف الانحدار اللوجستي حقق أعلى أداء مع نموذجي Bert sentence و BOWلاستخراج الميزات، بينما حصل مصنف SVM على أعلى أداء عند استخدام Bert-word وTF-IDF. حقق مصنف الغابة العشوائية الأداء الأعلى عند استخدام نموذج CBOW لاستخراج الميزات.

10-الخاتمة:

هدفت هذه الدراسة إلى إعطاء نظرة أعمق عن أداء أربع تقنيات لاستخراج الميزات BOW وTF-IDF وWord2vec وBERT المستخدمة في تحليل المشاعر. على الرغم من أنها حققت أداءً مرضياً، إلا أنَّ أياً منها لم يتفوق على جميع التقنيات الأخرى مع جميع المصنّفات. استناداً إلى المتوسط المرجّح لمقياس F-score، حقق نموذج BERT-sentenceأفضل أداء مع المصنّفات LR وNB و RF. بينما حقّق كل من النموذجين TF-IDF وBert-word أفضل أداء باستخدام مصنف SVM. عند استخدام مصنّف DT، حققت النماذج الثلاثة BERT sentence وTF-IDF وBOW أعلى محصلة. يمكن توسيع هذا العمل باستخدام تقنيات استخراج أخرى واختبارها على مجموعات بيانات من مجالات ولغات مختلفة.

11-المراجع

[1] D. B. A. M. Dhanalakshmi V., "Opinion mining from student feedback data using supervised learning algorithms," in *MEC International Conference on Big Data and Smart City*, 2016.

[2] S. V. D. S. Jagdale R.S., "Sentiment Analysis on Product Reviews Using Machine Learning Techniques.," in *Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing*, 2019.

[3] N. N. S. M. S. Tanjim Ul Haque, "Sentiment Analysis on Large Scale Amazon Product Reviews," in *IEEE International Conference on Innovative Research and Development 11-12 May 2018,Bangkok Thail*, 2018.

[4] A. A. Samar Al-Saqqa, "The Use of Word2vec Model in Sentiment Analysis: A Survey," in *Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control*, (2019).

[5] N. L. M. L. E. F. A. C. Joshua Acosta, "Sentiment Analysis of Twitter Messages Using Word2Vec," in *Proceedings of Student-Faculty Research Day CSIS*, (2017).

[6] V. M. ,. N. Y. Andrey Konstantinov, "Approach to the Use of Language Models BERT and Word2vec in Sentiment Analysis of Social Network Texts," in *O. Dolinina et al. (Eds.): ICIT 2020, SSDC 337*, (2021).

[7] L. E.-A. S. S. Israa Alghanmi, "Combining BERT with StaticWord Embeddings for Categorizing Social Media," (2020).

[8] A. C. K. G. A. Ravindar Ahuja, "The Impact of Features Extraction on the Sentiment Analysis," in *International Conference on Pervasive Computing Advances and Applications – PerCAA 2019*, 2019.

[9] S. E. Avinash M, "A Study of Feature Extraction techniques for Sentiment Analysis," in *arXiv:1906.01573v1 [cs.CL] 4 Jun 2019*, 2019.

[10] K. C. C. D. Tomas Mikolov, "Efficient Estimation of Word Representations in Vector Space," in *arXiv:1301.3781v3. 7 SEP2013*, 2013.

[11] M.-W. C. K. L. K. T. Jacob Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *arXiv:1810.04805v2* , (2019).

[12] K. J. M. H. M. Kamran Kowsari, "Text Classification Algorithms:A Survey," in *Information (2019), 10, 150; doi:10.3390/info10040150*.

[13] V. V. Corinna Cortes, "Support-vector networks," in *Machine Learning volume 20, pages273–297(1995)*.

[14] M. K. P. Jiawei Han, Data Mining Concepts and techniques, Morgan Kaufmann Publishers is an imprint of Elsevier., 2011.

[15] O. M. Lior Rokach, "Top-Down Induction of Decision Trees Classifiers—A Survey," in *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 35, NO. 4, NOVEMBER (2005)*.

[16] L. BREIMAN, "Random Forests," in *Machine Learning, 45, 5–32, (2001)*.