استخدام القواميس والتعلم الآلي لتحليل المشاعر في اللغة العربية

طالبة الدراسات العليا: م. ريم ناصر كلية الهندسة المعلوماتية– جامعة البعث

اشراف الدكتور: كمال السلوم

مشرف مشارك : د. ناصر أبو صالح

الملخص:

بسبب تزايد مواقع التجارة الالكترونية وشبكات التواصل الاجتماعي بدأ الناس بالاعتماد على الآراء الواردة في هذه الشبكات كأحد المراجع الأساسية عند شراء منتج أو خدمة، حيث تساعدهم في اتخاذ القرار أثناء عملية الشراء أو التصفح، ولكن عند الولوج إلى هذه الشبكات نلاحظ كثرة التعليقات والآراء حول العديد من المواضيع وبذلك يصبح من الصعب الوصول لكل هذه التعليقات وتحليلها يدوياً. قُدمَت هذه الدراسة بهدف إجراء مراجعة لبعض النماذج السابقة في مجال تحليل المشاعر للغة العربية، واستخدام طرق تساعد في تحسين دقة التنبؤات وذلك باستخدام تقنيات أفضل لتجذيع الكلمات. تمت مقارنة نتائج تنفيذ هذه الطرق مع تقنيات أخرى وذلك بالنسبة لدقة التنبؤ. ثم تم تطبيق عدة خوارزميات للتعلم الالي ومقارنة نتائجها بالنسبة لمجموعتي البيانات وتحديد الخوارزمية الأفضل.

الكلمات المفتاحية: تحليل المشاعر، القواميس، التعلم الآلي، التجذيع، نماذج التصنيف، القطبية.

Using of lexicons and machine learning for sentiment analysis in the Arabic language

Eng. Reem Nasser **Dr. Kamal Alsalloum**
Dr. Nasser Abo Saleh

Abstract

Due to the increase in e-commerce sites and social networks, people began to rely on the opinions expressed in these networks as one of the basic references when purchasing a product or service, as it helps them in making decisions during the buying or browsing process, but when accessing these networks we notice the abundance of comments and opinions about many Topics thus making it difficult to access all of these comments and analyze them manually. This study was presented with the aim of conducting a review of some previous models in the field of sentiment analysis for the Arabic language, and using methods that help improve the accuracy of predictions by using better techniques for word stemming. The results of implementing these methods were compared with other techniques in terms of prediction accuracy. Then, several machine learning algorithms were applied and their results compared for the two data sets and the best algorithm was determined.

Keywords: sentiment analysis, lexicons, machine learning, stemming , classification models,polarity.

1. مقدمة:

انتشر تحليل المشاعر بكثرة في السنوات الأخيرة، حيث استخدم في كثير من المجالات معتمداً على تعليقات الأشخاص التي تعد عنصراً أساسياً لمعرفة مشاعرهم حول أي منتج.

يعرف مصطلح تحليل المشاعر sentiment analysis أو التنقيب في الآراء opinion mining بأنه تعريف أتوماتيكي للآراء المُعبر عنها في النص تجاه موضوع معين [1] .

يمكن القول أن السبب الرئيسي وراء ازدياد الاهتمام بتحليل المشاعر هو انتشار استخدام شبكة الانترنت التي شكلت بيئة مميزة لتفاعل البشر على اختلاف أماكنهم وأفكارهم، حيث زاد حجم البيانات المتبادلة بين مستخدمي الانترنت وهذا ما ساعد على زيادة التواصل وتبادل المعارف والمشاعر حول ظاهرة أو منتج معين، وبذلك أيضاً توفر للشركات وللمحللين بيانات خام لإجراء الإحصائيات اللازمة لمعرفة المنتجات الأكثر رغبة من قبل المستهلكين أو معرفة أراء الناس حول إحدى الظواهر بين إيجابي أو سلبي.

تجري عملية تحليل المشاعر على مستويات عدة من التقسيمات: الكلمة، العبارة، الجملة، المستند، أو المفهوم. كما توجد طريقتان أساسيتان لتطبيق تحليل المشاعر: الطرق التي تعتمد على القواميس وطرق التعلم الآلي.

تنصب معظم الجهود في مجال تحليل المشاعر على اللغة الانكليزية واللغات الأوربية الأخرى في حين حظيت اللغة العربية بالقليل من الأعمال. معظم طرائق تحليل المشاعر طُورت للغة الانكليزية ومن الصعب تطبيقها على باقي اللغات كاللغة العربية.

تعد اللغة العربية لغةً غنيةً من حيث المفردات والقواعد وهذا ما شكل تحدياً كبيراً لأنظمة معالجة اللغات الطبيعية NLP، حيث أن القواميس العربية المحددة القطبية قليلة نوعاً ما وليست متاحة للعموم إضافةً إلى ذلك ركزت الجهود المبذولة لتحليل المشاعر في اللغة

العربية على معالجة النصوص العامة دون تحديد المجال وتم تطوير القليل في مجالات محددة كالرياضة، القانون، وما إلى ذلك.

2. الهدف من البحث:

أصبح لدينا كم هائل من البيانات وذلك نظراً لتطور تكنولوجيا المعلومات والويب [2] ، وهذا ما سمح لنا بفهم مشاعر مستخدمي الانترنت من خلال تحليل هذه البيانات. على العموم يمكن القول أنه كلما كان حجم هذه البيانات كبيراً، كلما كان من الصعب تحليلها لاستخراج معلومات مفيدة منها.

أظهرت الدراسات أن وسائل التواصل الاجتماعي مثل فيسبوك تنتج ما يقارب 250 مليون منشور في الساعة، بينما ينتج مستخدمو تويتر حوالي 21 مليون تغريدة في الساعة.

يغطي هذا البحث فعلياً ما يلي:

1- دراسة شاملة لمختلف تقنيات تحليل المشاعر إضافةً إلى تقنيات معالجة البيانات.
2- تحليل الأعمال التي قام بها مجموعة من الباحثين في مجال تحليل المشاعر حتى وقتنا الحالي باستخدام مختلف التقنيات.
3- مقارنة تحليلية لعدة مجذعات stemmer بهدف الوصول للمجذع الأفضل، ومقارنة عمل مجموعة من خوارزميات التصنيف على مجموعتين من البيانات.

3. أدوات وطرائق البحث:

يتم التعامل في هذه الورقة مع نظام تشغيل windows 10/64 bit منصب على جهاز له المواصفات التالية:

- المعالج: intel core i7-10510U 2.30 GHz
- الذاكرة: 16 غيغا بايت.

4. تحليل المشاعر:

يعرف تحليل المشاعر sentiment analysis أو التنقيب في الآراء opinion mining على أنه تحليل لآراء الناس ومشاعرهم ومواقفهم تجاه منتج أو خدمة أو موضوع معين. و يعتبر فرعاً من التنقيب في محتويات الويب حيث يقوم بتطبيق تقنيات التنقيب في البيانات لاكتشاف الأنماط وتطبيق تقنيات معالجة اللغات الطبيعية لتتبع أراء الناس حول منتج أو خدمة معينة. [1]

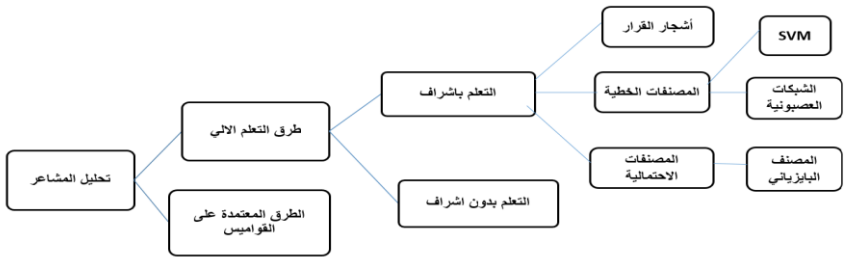
4-1 مستويات تحليل المشاعر:

يوجد ثلاث مستويات لتحليل المشاعر وهي: [1]

- **تحليل على مستوى Document:** مهمة هذا المستوى تصنيف الرأي الكلي للوثيقة فيما إذا كان شعوراً إيجابياً أو سلبياً. يفترض هذا المستوى أن كل وثيقة تعبر عن أراء حول غرض واحد (منتج مثلاً) وبالتالي غير قابل للتطبيق على الوثائق التي تقيّم أو تقارن أغراض متعددة.
- **تحليل على مستوى Sentence:** مهمة هذا المستوى من التحليل تحديد فيما إذا كانت الجملة تعبر عن رأي إيجابي أو سلبي أو حيادي (لا رأي).
- **تحليل على مستوى Aspect:**يعتمد هذا المستوى على فكرة أن الرأي يتألف من شعور(إيجابي ،سلبي، حيادي) وغرض(ميزة) مستهدف بالشعور. وبدلاً من النظر إلى التراكيب اللغوية(الوثائق أو العبارات أو الفقرات) يتم النظر مباشرة إلى الرأي بالغرض بحد ذاته.

4-2 تقنيات تحليل المشاعر:

تقسم تقنيات تحليل المشاعر إلى نوعين مختلفين: تدعى الأولى بتقنيات التعلم الآلي أما الثانية فهي الطرائق التي تعتمد على القواميس، وهذا ما نلاحظه في الشكل(1).

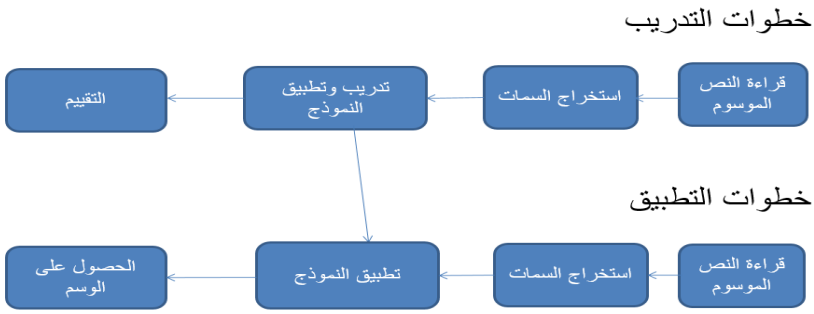


الشكل –1– تقنيات تحليل المشاعر.

4-2-1 طرائق التعلم الآلي:

1- التعلم الخاضع للإشراف Supervised Learning: هناك العديد من التقنيات الشائعة المستخدمة لتحليل المشاعر والتي تندرج ضمن طرائق التعلم بإشراف نذكر منها مايلي: [2]

- المصنف البايزياني Naïve Bayes : يُعرف بأنه مصنف احتمالي بسيط يعتمد على نظرية بايز وفرضية الاستقلال (كل المميزات مستقلة عن بعضها البعض)، حيث يقوم بحساب الاحتمالات المسبقة لكل وسم اعتماداً على تكرار ظهوره ضمن مجموعة بيانات التدريب.
- Maximum entropy(ME): التوزيع الاحتمالي الأفضل لتمثيل الحالة الراهنة من المعرفة هو التوزيع الذي يحتوي على الأنتروبي الأعظم.
- Support vector machines (SVM): أحد نماذج التعلم بإشراف المستخدمة لتحليل البيانات والتعرف على الأنماط للقيام بعمليات التصنيف والانحدار regression، ويعتبر أحد المصنفات الخطية حيث يفصل مجموعات البيانات إلى فئتين أو أكثر بحيث يكون التباعد بين الفئتين أعظمياً وذلك وفق قانون رياضي.
- أشجار القرار: يتم تقسيم البيانات عودياً باستخدام شرط معين حتى نصل للأوراق التي تنتمي لأحد الصفوف، أشهرها: C4.5 ، ID3



الشكل –2– : آلية تحليل المشاعر باستخدام طرائق التعلم الآلي.

يوضح الشكل (2) خطوات تحليل المشاعر باستخدام طرائق التعلم الالي، حيث يوضح عملية تدريب النموذج المقترح بدايةً ثم اختباره ومعرفة فاعليته وأداءه .

2– التعلم بلا إشراف Unsupervised Learning :

يندرج التجميع في حقل التعلم من دون إشراف، حيث تقوم هذه الخوارزميات بتقسيم البيانات إلى عدد من المجموعات الفرعية. كل من هذه المجموعات الفرعية تحتوي على بيانات مشابهة لبعضها البعض. تُعرف هذه المجموعات الفرعية باسم المجموعات أو العناقيد.

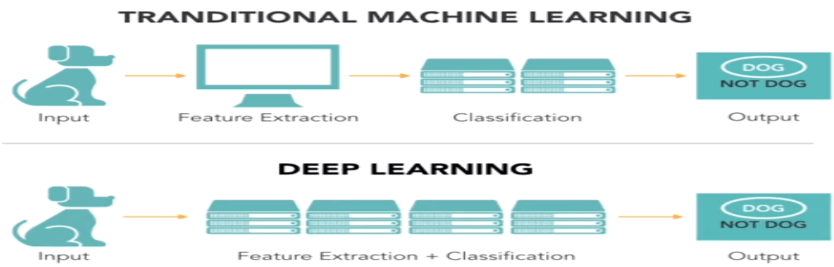
4–2–2 التقنيات التي تعتمد على القواميس (المنهجيات الدلالية):

تعتمد هذه الطرق على قواميس الكلمات ذات القطبية، حيث تقوم الأنظمة بمعالجة النص وتقسيمه لمجموعة من الكلمات ثم إهمال كلمات التوقف وإعادة الكلمات إلى أصلها stemming ثم مقارنة كل كلمة مع القاموس الذي يعطي أوزاناً محددة لكل كلمة وجمع هذه القيم للحصول على قيمة الشعور العام للنص. أيضاً تتضمن القواميس معالجة لبعض الكلمات التي تزيد أو تنقص القطبية مثل (very, little) أو التي تعكس الشعور الذي تؤثر عليه مثل (no, not).

4–2–3 استخدام التعلم العميق في تحليل المشاعر: [3]

أحد أشكال التعلم الآلي، يتضمن خوارزميات تتيح للآلة التعلم بنفسها عن طريق محاكاة الخلايا العصبية، أثبتت الدراسات تقدماً كبيراً في مختلف المجالات أهمها معالجة اللغات الطبيعية. يمكن أن يقدم التعلم العميق دقة عالية قريبة من دقة الأداء البشري وذلك لأنه يحاكي الخلايا العصبية، وتستخدم نماذج التعلم العميق لمجموعات البيانات الكبيرة المحددة الصفوف حيث تتعلم المميزات مباشرةً من البيانات دون الحاجة لعملية الاستخراج اليدوي لها وهذا مايمكن ملاحظته من خلال الشكل(3).

علماً أن أداء شبكات التعلم العميق يزداد بازدياد حجم البيانات بينما أداء خوارزميات التعلم الآلي التقليدية يزداد إلى حد معين ثم يصبح ثابتاً مهما ازداد حجم البيانات.



الشكل –3– : الفرق بين التعلم الالي التقليدي والتعلم العميق.

يوضح الشكل(3) أن بنى التعلم العميق ماهي إلا شبكات عصبونية ذات طبقات مخفية كثيرة العدد (كلما كان عددها أكبر كلما كانت الشبكة أعمق)، يتم معالجة المستندات وتمرر إلى طبقة الدخل للشبكة العصبونية حيث يعبر عن كل كلمة بشعاع مستخدماً موقع كل كلمة في الجملة لإيجاد التشابه العاطفي بين الكلمات وهذا ما يسمى Word2vec أو يتم حفظ ترتيب الكلمات في الجملة إضافةً إلى ما سبق (Doc2vec) وغيرها الكثير من الأساليب لأنظمة تضمين الكلمات word embedding. ثم ترسل للطبقات المخفية ومنها لطبقة الخرج التي تعطي الخرج النهائي (الشعور، التصنيف، الخ).

3–4 التحديات التي تواجه مجال تحليل المشاعر: [4]

هناك العديد من التحديات التي تواجه تحليل المشاعر بشكل عام والتي توصف كما يلي:

- المشاعر الضمنية والسخرية: من المحتمل أن تحوي الجملة على مشاعر ضمنية حتى لو لم يكن هناك كلمة ظاهرة تدل على نوع هذه المشاعر، أو أن تحوي على نوع من السخرية وهذا ما يشير إلى نوع من السلبية غير الظاهرة بشكل علني.
- الاعتماد على المجال: تختلف قطبية الكلمة من مجال لآخر، فمن الممكن أن تعطي شعوراً إيجابياً في مجال ما ومشاعراً سلبيةً في مجال اخر.
- الآراء المزيفة: يعتبر الرأي المزيف تضليل للمستخدمين والقراء من خلال تزويدهم برأي إيجابي أو سلبي أو غير صادق يتعلق بأي موضوع، وهو التحدي الاجتماعي الذي يواجهه مجال تحليل المشاعر.
- النفي: يؤدي استخدام أحد أدوات النفي إلى تغيير قطبية الرأي بشكل كامل.

أما فيما يتعلق باللغة العربية بشكل خاص فهناك العديد من التحديات الإضافية وذلك بسبب تنوعها وغناها الصرفي، نذكر منها ما يلي:

- تنوع اللغة العربية: تعد اللغة العربية لغةً رسمية لحوالي 300 مليون شخص في 22 دولة، وبالتالي هناك العديد من اللهجات المستخدمة إضافةً للغة الفصحى.
- قواعد الإملاء في اللغة العربية: تتألف اللغة العربية من 28 حرف، وتكتب الكلمات من اليمين إلى اليسار وهناك العديد من الإشارات التي توضع فوق أو تحت الأحرف ليتم تهجئة الكلمة بشكل صحيح ولتوضيح معناها.
- قواعد الصرف في اللغة العربية: يمكن اشتقاق العديد من الكلمات من نفس الكلمة بإضافة لاحقة أو سابقة على الكلمة، مثلاً من الكلمة "شعر" نشتق الكلمات "شاعر"، "شعور" ،وغيرها الكثير. أيضاً كل فعل له أكثر من شكل (الماضي والحاضر) ويختلف أيضاً حسب الفاعل(أنا، هو، هي، الخ.).

- قلة مجموعات المفردات Corpora : ليس هناك أي مجموعة مفردات للغة العربية يمكن أن تغطي جميع الكلمات وبكل اللهجات، علماً أن الدقة تزداد حسب نوعية وحجم مجموعة المفردات المستخدمة لتدريب مصنف المشاعر.
- قلة قواميس المشاعر sentiment lexica : هناك العديد من الأعمال الحالية لإنتاج قواميس المشاعر، وعلى الرغم من ذلك لا يزال عددها قليلاً مقارنةً بقواميس اللغة الانكليزية.
- تعدد اللهجات العربية: يستخدم الناس لهجاتهم للتعبير عن آرائهم على شبكات التواصل الاجتماعي، إلا أن أدوات معالجة اللغات الطبيعية مصممة خصيصاً للغة العربية الفصحى وهذا ما يجعل مهمة المعالجة أكثر تعقيداً.

4-4 خطوات تحليل المشاعر:

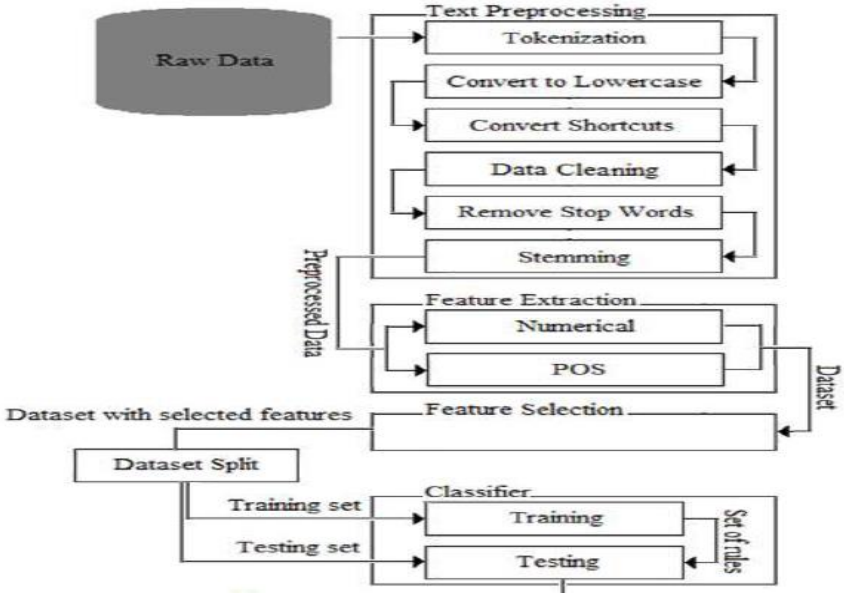
يعتبر نظام تحليل المشاعر نظاماً أوتوماتيكياً لتحديد قطبية النص المدخل. حيث يعتمد على تقنيات التعلم الآلي ويمر بعدة خطوات: الخطوة الأولى هي جمع البيانات يليها معالجة النص بهدف التخلص من الضجيج في النص باستخدام مجموعة من العمليات كالتقطيع، حذف كلمات التوقف، التجذيع (إعادة الكلمة لجذرها). فيما بعد تأتي خطوة استخراج المميزات من البيانات، ومن ثم اختيار أكثرها فائدة من المجموعة الكلية، وهذا مايوضحه الشكل(4). فيما يلي سنتحدث عن الخطوات السابقة بالتفصيل: [5]

- اختيار البيانات Data Selection: وهي المرحلة الأولى من أي نظام تصنيف، بهدف اختيار البيانات الأنسب لهذه المهمة. قمنا باختيار مجموعتي بيانات في هذا البحث.
- معالجة البيانات Text processing: وهي المرحلة الأهم في أي نظام تصنيف، وتأتي أهميتها لكونها تقوم بفلترة النص المدخل وذلك عن طريق حذف الكلمات غير المرغوبة وتحويل الكلمات إلى تمثيل مناسب.

تغطي هذه العملية مجموعة من الخطوات بهدف المعالجة المسبقة للبيانات:

– التقطيع Tokenization: وهي عملية تقسيم النص المدخل إلى أجزاء أصغر كالكلمات أو العبارات وتدعى هذه العناصر بالرموز Tokens. وقد نالت عملية التقطيع أهميتها من الحاجة للتعامل مع كل كلمة بمفردها كدخل. تتم هذه الخطوة بمسح النص المدخل وحفظ كل جزء منفصل من النص كرمز (يتم الفصل من خلال الفراغات).

– تنظيف البيانات Data Cleaning: وهي عملية إزالة الضجيج من النص المدخل باستخدام الخطوات التالية:



الشكل–4– : بنية نظام تحليل المشاعر

1– التحويل للأحرف الصغيرة: الحصول على شكل واحد للنص وهذا ما يسهل التعامل مع الكلمات، تكمن أهمية هذه الخطوة لأنه بدونها يتم اعتبار الكلمات المتشابهة ذات الأشكال المختلفة للحرف على أنها كلمات مختلفة وبالتالي تسبب مشاكل للنظام عند حاجته للعمل على ذاكرة أو سرعة محددة.

2- تحويل الاختصارات: يتم تحويل كافة الاختصارات إلى كلماتها الأصلية، وهذا ما يقلل عدد الكلمات التي سيتعامل معها النظام.

3- إزالة الروابط: يتم إزالتها من النص لأنها جزء غير مهم لعملية التصنيف، ويتم حذفها من خلال فحص كل الرموز وحذف أي رمز يحوي URL .

4- حذف الرموز التعبيرية: وهي عبارة عن صور رقمية تستخدم في الشبكات الاجتماعية، عندما يدخل النص إلى النظام يتم تحويل أي رمز تعبيري إلى ثلاث علامات استفهام؟؟؟ وبالتالي تصبح عملية الحذف أسهل عن طريق إزالة أي رمز يتألف من ثلاث إشارات استفهام متتالية.

5- حذف إشارة (//) بين أي كلمتين: يتم البحث عن هذه الإشارة واستبدالها بفراغ في أي رمز من النص المدخل.

6- حذف الأقواس: يتم حذف الأقواس لأنها غير مهمة في عملية التحليل الحالية، حيث يتم فحص كل الرموز المدخلة وحذف أي قوس بداية أو نهاية.

7- حذف الأرقام: تعد الأرقام غير مهمة في مسألة تحليل المشاعر.

8- حذف إشارة Mention: هناك أداة في الشبكات الاجتماعية تستخدم لإخبار المستخدمين عن تعليق أو منشور معين من خلال ذكر اسمه باستخدام الإشارة @ (أي ذكر حسابه). وتعتبر حسابات المستخدمين وإشارة @ غير مهمة في عملية تحليل المشاعر وبالتالي يتم حذفها.

9- حذف إشارة الهاشتاغ: يعد الهاشتاغ أداة في الشبكات الاجتماعية تمثل كل المنشورات التي تستخدم #، وبما أن علامة الهاشتاغ غير مهمة في عملية تحليل المشاعر يتم حذفها ولكن نحافظ على كلمة الهاشتاغ لأنها من الممكن أن تحوي معلومات مفيدة حول قطبية النص.

- حذف كلمات التوقف: وهي الكلمات التي تظهر بشكل متكرر في النص ولكن لا تحوي أي معنى ويقتصر استخدامها في الجملة لربط الكلمات فقط.

- إعادة الكلمات إلى جذورها Word Stemming: وهي عملية تقييس لغوي تهدف لتحويل مختلف أشكال الكلمة إلى شكل شائع وحيد.

- استخراج المميزات: تعد مسألة اختيار المميزات أحد المهام الحرجة في مجال تحليل المشاعر، إذ أنَّ اختيار المميزات الصحيحة يحدد الأداء الكلي لعملية التصنيف[6] . فيما يلي سنشير إلى أكثر المميزات استخداماً في هذا المجال:

– N–grams: وهو متتالية من الكلمات المستخرجة من النص. أكثرها استخداماً unigrams والتي تحوي كلمة واحدة، bigrams والتي تحوي كلمتين متتاليتين، trigrams والتي تحوي ثلاث كلمات متتالية.

– TF–IDF (term frequency–inverse document frequency): وهو مقياس إحصائي يقيّم مقدار ارتباط الكلمة بالمستند في مجموعة من المستندات. حيث يتم إسناد الأوزان العليا للكلمات ذات التردد الأعلى في مختلف المستندات.

$$w_{x,y} = tf_{x,y} \log\left(\frac{N}{df_x}\right)$$

حيث أن w هو وزن الكلمة x في المستند y و tf هو تكرار الكلمة x في المستند y ، N هو عدد المستندات في المجموعة و df هو عدد المستندات التي تحوي الكلمة x . تكرار الكلمة للمستند ينبغي أن يتم تقييسه من خلال قسمته على طول المستند.

– أجزاء الكلام part of speech (POS): وهي عملية تحديد الكلمات في النص بالاعتماد على طبيعتهم وعلاقاتهم مع الكلمات المجاورة والمرتبطة في ذلك النص. عندما يكون النص مكتوباً باللغة الانكليزية يتم إسقاط كل كلمة إلى ثمانية فئات نحوية: الأفعال، الأسماء، الضمائر، الظروف، الصفات، حروف الجر، كلمات الربط كحروف العطف مثلاً، وصيغ التعجب. تتمتع اللغة الإنكليزية بغناها بالأدوات المخصصة للتقطيع الخاصة بتويتر –Twitter specific POS taggers ، أما بالنسبة للغة العربية فإن معظم هذه الأدوات مخصصة للغة الفصحى مع بعض الأعمال التمهيدية للهجة المصرية.

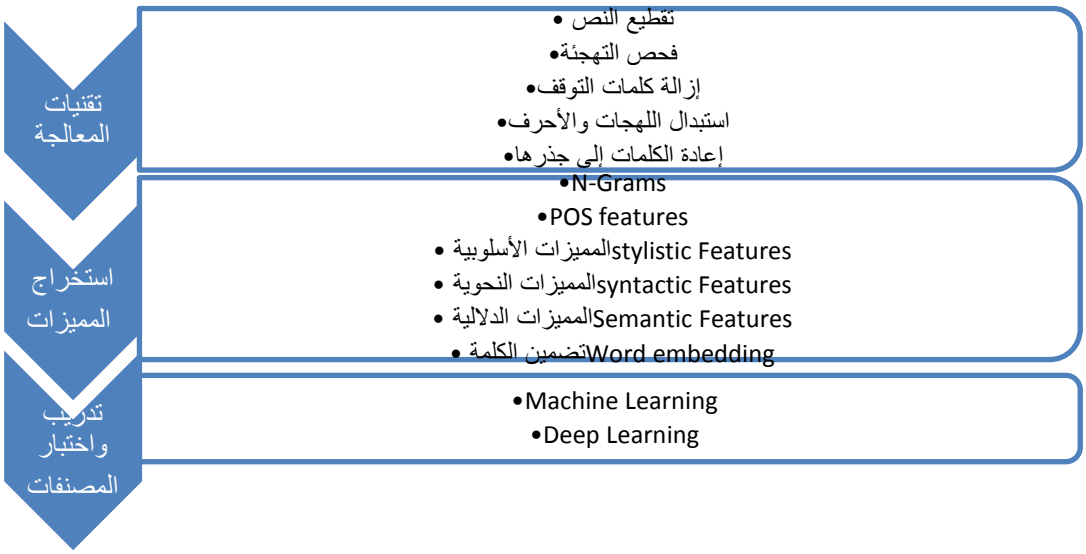
– الميزات الأسلوبية stylistic features: تهدف هذه الميزات إلى التحقق من وجود مؤشرات معينة للمشاعر في النص، مثل الرموز الإيجابية والسلبية ومقياس الغنى ببعض المفردات وتكرار محارف خاصة. تشمل المميزات الأسلوبية أيضاً فحص علامات الاستفهام والتعجب.

– الميزات النحوية Syntactic features: هي نماذج عبارات تشير إلى المشاعر مثل الأسماء التي تليها صفة سلبية. على سبيل المثال عبارة "يوم جيد". يستخدم عادةً لاستخراج مثل هذه الميزات n–grams، POS، وحقيبة الكلمات BOW. أو تُستخدم الأفعال المتعدية واللازمة.

– الميزات الدلالية Semantic features: تمثل الميزات الدلالية الاتجاه الدلالي للنص. من أجل قطبية معينة، تقيس هذه السمات ارتباط مجموعة من الكيانات من خلال مفاهيم الكيانات المختلفة.

– ميزات المعجم Lexicon features: تشتق هذه المميزات من قاموس المشاعر، مثل معدل المشاعر. في حين أن هنالك العديد من المعاجم للغة الإنكليزية، إلا أن اللغة العربية تمتلك القليل من المعاجم والتي تم إنشاؤها تلقائياً من وسائل التواصل الاجتماعي أوترجمتها من المعاجم المكتوبة باللغة الإنكليزية.

يوضح الشكل(5) خطوات تحليل المشاعر باستخدام تقنيات التعلم الالي من تقنيات المعالجة، إلى تقنيات اختيار المميزات وصولاً إلى تدريب واختبار المصنفات.



الشكل –5–: الطرائق المعتمدة على تقنيات التعلم الآلي لتحليل المشاعر باللغة العربية.

5. مصادر المشاعر باللغة العربية:

1– مجموعات المفرداتcorpus : تعد هذه المجموعات أساسية لتدريب نظم تصنيف المشاعر، حيث تحوي عدداً هائلاً من التعابير على شكل كلمات، عبارات، جمل.

بذل الباحثون جهد كبير لتطوير مصادر للغة العربية بهدف تحسين دقة عملية تحليل المشاعر، سنذكر بعض الأعمال التي اعتمدت على مجموعات المفردات المتاحة والتي سيتم تقييمها تبعاً لعدة خصائص منها: الكمية، النوعية، تغطيتها لأكثر من لهجة، توفره.

قدم الباحثون [7] ARSENTD–LEV وهو عبارة عن مجموعة مشاعر باللغة العربية تتألف من 4K تغريدة باللهجة الأردنية، اللبنانية، الفلسطينية والسورية. يتم اختيار الشعور الكلي للتغريدة (سلبي جداً، سلبي، محايد، إيجابي، إيجابي جداً).

قدم الباحثون [8] أكبر مجموعة بيانات تضم مراجعات لكتب باللغة العربية BRAD خاصة بتحليل المشاعر، حيث تضم 510K مراجعة بعدة لهجات أكثرها شيوعاً اللهجة المصرية.

قدم الباحثون [9] أكبر مجموعة بيانات تضم مراجعات لفنادق باللغة العربية HARD متاحة بالمجان للباحثين حيث تضم حوالي 370000 مراجعة باللغة العربية الفصحى إضافةً إلى بعض اللهجات، وتتوزع هذه المراجعات بين السلبية والايجابية.

2- القواميس: إن توفر القواميس يعد شيئاً أساسياً في تطوير نظم تحليل المشاعر، وذلك لأنها توفر الوقت والجهد. سنستعرض بعض القواميس العربية وأهم خصائصها.

لابد لنا أن نذكر أنَّ الباحثين [10] حولوا قاموس sentiWordNet (SWN) إلى اللغة العربية، حيث يعتبر SWNقاموساً متاحاً بشكل عام للباحثين يعتمد على كلمات مكتوبة باللغة الانكليزية. يتم إعطاء قيم رقمية سلبية أو إيجابية كشعور لكل كلمة من المفردات وتم اعتماد نفس الأسلوب في القاموس الناتج للغة العربية، الذي يحوي فقط 10K كلمة وليس متاحاً للعموم.

طور الباحثون [11] قاموس مشاعر ثنائي اللغة يدعى BiSaL ، حيث بنوا قاموسين SentiLEn للغة الانكليزية وSentiLAr للعربية. استخرجت كلمات المشاعر من 2000 منشور في منتدى العقاب مرتبطة بالتهديدات الأمنية، النزاعات، والتطرف، فيما بعد تم الاستعانة بثلاث خبراء باللغة العربية لإعطاء القطبية للمصطلحات المستخرجة (0،1،-1). على أي حال لم يتحقق الباحثون من القاموس الناتج باستخدامه في تطبيق حقيقي.

6. الدراسات السابقة:

1- دراسات باللغة الانكليزية:

قام كل من P. Bansal و R. Kaur [12] بتصنيف النصوص إلى سلبية أو إيجابية، قدموا طريقة هجينة باستخدام خوارزميات تحسين ذكية تدعى Swarm مع المصنفات . يتم معالجة كل تغريدة من خلال عدة مراحل : عملية التقطيع، إزالة كلمات التوقف، إعادة الكلمة لأصلها stemming. ثم يتم تشكيل شعاع المميزات من خلال حساب TF-IDF وpractical Swarm Optimization قبل عملية تصنيف النصوص. تم استخدام كل من المصنف البايزياني و SVM من أجل تصنيف التغريدات، وقد أظهرت

النتائج أن استخدام عملية التحسين مع التصنيف أعطت نتائج أفضل من استخدام المصنفات لوحدها.

قدم A. Ortigosa واخرون [13] طريقة جديدة لتحليل المشاعر في فيسبوك، حيث تقوم باستخراج معلومات حول قطبية مشاعر المستخدمين(إيجابي، سلبي، محايد) ونمذجة قطبية مشاعرهم الاعتيادية. استُخدم تطبيق خاص بالفيسبوك يدعى SentBuk مهمته استعادة الرسائل المكتوبة من قبل المستخدم وتصنيفها، وتدعم عملية اكتشاف التغيرات العاطفية.

إن عملية التصنيف في SentBuk تعتمد طريقة هجينة حيث تدمج تقنيات التصنيف المتعمدة على القواميس والتصنيف الآلي. أظهرت النتائج إمكانية تحليل المشاعر في فيسبوك بهذه الطريقة وبدقة عالية 83.27% . يمكن أن تستخدم هذه النتائج من قبل الأنظمة التعليمية لمعرفة مشاعر الطلاب تجاه مادة معينة وبذلك تقدم تلك المعلومات للمعلمين وخاصة في حالة التعليم عن بعد.

اقترح K. Gull وأخرون [14] طريقة لكشف المشاعر من بيانات مأخوذة من تويتر، حيث يتم معالجتها وإعطاء وزن لكل كلمة اعتماداً على POS ومن ثم حساب النتيجة النهائية واختبار المجال التي تنتمي إليه لتصنيفها. تمت مقارنة نتائج الطريقة المقترحة مع نتائج تطبيق المصنف البايزياني (يعتمد مبدأ الاحتمالات وبالتالي إذا ظهرت كلمة ولم تكن موجودة في مجموعة بيانات التدريب سيقوم بتجاهلها) باستخدام الأداة WEKA على نفس مجموعة البيانات، وقد أظهرت النتائج تفوق تقنية NLP على NB لأن الاخيرة تحتاج بيانات تجريب كثيرة والتي بدورها تتطلب ذاكرة أكبر. يمكن تطوير النظام المقترح من خلال استخراج التعليقات للزبائن من مواقع متعددة وتحليلها باستخدام خوارزميات مختلفة لمعرفة وجهة نظرهم بالمنتجات أو الأفراد، أيضاً يمكن تحسين الدقة من خلال حذف كلمات التوقف من التغريدات المُجمعة. نستطيع أن نعمم النموذج المقترح ليقبل الصوت والصورة كدخل له واستخراج النص منهم لتحليله، أو معالجة البيانات الضخمة باستخدام خوارزمية MapReduce التي يمكن أن تضاف للمعالجة بدون الانترنت.

اقترح الباحثون [15] طريقة لتحسين دقة التوقع في الأنظمة الناصحة اعتماداً على المشاعر. في البداية قدموا آلية لحساب شعور كل مستخدم تجاه المنتج، ثم أخذوا سمعة المنتج بعين الاعتبار والتي يمكن استنتاجها من قبل توزيعات عاطفية لمجموعات المستخدمين التي تعكس تقييم العملاء الشامل. أخيراً دمجوا ثلاثة عوامل: تشابه مشاعر المستخدمين، تأثير المشاعر بين الأشخاص، وتشابه سمعة المنتجات في نظام التوصية لجعل التنبؤ بالتقييم أدق.

قدم الباحثون [16] نظاماً يعتمد على خوارزميات التنقيب في البيانات لتصنيف مراجعات المستخدمين، حيث دمجوا المصنف البايزياني والشبكات العصبونية للقيام بهذه المهمة وتم الوصول إلى دقة أعلى بنسبة 80.65% من خلال هذه العملية وذلك باستخدام unigram للحصول على المميزات. يمكن تطبيق هذا العمل على عناقيد مراجعات الأفلام لتطبيقات التنقيب في الآراء حيث يستخدم العنقود لحل مشكلة نقص الآراء.

اقترح الباحثون [17] طريقة جديدة لدمج التعلم العميق مع النماذج التي تعتمد على الميزات باستخدام شبكة MLP لتحليل المشاعر المالية. لذلك طوروا نماذج عديدة للتعلم العميق CNN LSTM GRU ، التي تم تدريبها باستخدام مجموعة من القواميس. أظهرت النتائج أن النظام المقترح حقق تفوقاً على الأنظمة الموجودة عند اختباره على قاعدتي بيانات مأخوذة من (المدونات الصغيرة وعناوين الأخبار).

اقترح كل من N. Mittal وآخرون [18] طريقة لتصنيف التغريدات في تويتر إلى صنفين (إيجابي أو سلبي)، حيث اعتمدوا على نموذج يتكون من ثلاث مراحل. تعتمد المرحلة الأولى بالتصنيف على الرموز التعبيرية emoticons ، وفي حال عدم وجودها يتم تصنيف التغريدات تبعاً لقائمة من الكلمات المعرفة مسبقاً ذات مشاعر سلبية أو ايجابية، أخيراً يتم إعطاء أوزان للرموز token اعتماداً على القواميس وطريقة تعتمد على الاحتمالات(احتمال أن الكلمة تنتمي لصف إيجابي هو حاصل قسمة عدد المرات التي تظهر فيها الكلمة في صف إيجابي على العدد الكلي لمرات ظهور هذه الكلمة، وبنفس الأسلوب تحسب قيمة القطبية السلبية). درس هذا البحث أيضاً تأثير أدوات الربط في

مرحلة المعالجة الأولية، أظهرت النتائج العملية فعالية الطريقة الهجينة التي تم اعتمادها في تصنيف التغريدات. حيث اعتمدت الطريقة الهجينة على الطريقتين السابقيتين معاً لإعطاء القطبية الموزونة للكلمة.

اقترح الباحثون [19] تقنية لإنشاء قاموس مشاعر يعتمد على مجال محدد، حيث يتم دمج قاموس عام مع قاموس معتمد على المجال لتحليل المشاعر. تم تحقيق وتقييم هذه التقنية على تغريدات مأخوذة من تويتر لإظهار الفائدة من عملية دمج القاموسين، ومقارنة الأداء مع مصنفات التعلم الآلي التي أثبتت تفوقها، وأخيراً دراسة سلوك التقنية على مجموعات بيانات مختلفة من حيث الحجم.

اقترح الباحثون [20] طريقة هجينة تدمج التعلم بدون إشراف (K-means clustering) وذلك لتجميع التغريدات مع التعلم بإشراف (أشجار القرار، SVM ،..) للقيام بالتصنيف، وذلك لتحسين دقة توقع المشاعر في تويتر. أثبتت هذه الطريقة كفاءتها عند دمج خوارزمية K-means مع أشجار القرار، ويستطيع النموذج المقترح استيعاب كميات كبيرة من بيانات تويتر النصية وكلما كانت البيانات أضخم وأكثر تنوعاً كان أداء النموذج أفضل.

2- دراسات باللغة العربية:

هناك الكثير من الأعمال التي ركزت على اللغات الأوربية على عكس العربية التي شكلت حيزاً صغيراً من الاهتمام وبالتالي هناك القليل من الأعمال التي تتوجه لمعالجة مشاكلها.

-استخدام تقنيات التعلم الآلي فقط:

حلل الباحثون [21] تعليقات الفيسبوك المكتوبة باللغة العربية الفصحى واللهجة المغربية، بهدف اختبار ميزة إضافة تصنيف مجموعات المفردات العربية Arabic corpus تبعاً لنماذجها (لغة فصحى أو لهجات) وذلك قبل عملية تصنيف المشاعر. حيث ارتكزت فكرتهم الأساسية على تكييف عملية المعالجة الأولية للنص تبعاً للفئة اللغوية. على سبيل

المثال استخدم الباحثون الاشتقاق الخفيف للنصوص (إعادة الكلمة لأصلها) light stemming . طبقوا نهجهم مع المصنف البايزياني NB و SVM . سمح هذا التصنيف المكون من خطوتين بتقليل الأخطاء الناتجة عن اشتقاق الكلمات. ومع ذلك، فإن هذا النهج سيكون معقداً في حالة الحجم الكبير ومجموعات البيانات المتعددة اللهجات.

ركز المؤلفون [22] على اللغة العربية الفصحى واللهجة الأردنية. قاموا بتطبيق خوارزميات التعلم الالي الخاضعة للإشراف على وسائل التواصل الاجتماعي الخاصة بالعرب للمواد العامة والتي تتضمن إما MSA أو اللهجة الأردنية. استخدموا SVM و NB لتقييم تأثير مخططات الوزن المختلفة والاشتقاق وتقنيات n-grams. أظهرت التجارب أن مصنف SVM الذي يستخدم stemmer و TF-IDF للأوزان من خلال bigrams هو أفضل تركيبة، ولاحظ المؤلفون أن هذا المزيج حقق أعلى أداء وتفوق على مصنف NB.

-استخدام تقنيات التعلم العميق فقط:

اقترح الباحثون [23] إطاراً يعتمد خوارزمية التطور التفاضلي differential evolution (DE) لشبكة CNN من أجل تعزيز الأداء. تم استخدام خوارزمية DE للعثور على التكوين الأمثل لكل من بنية CNN وبارمترات الشبكة. تم ضبط خمس بارمترات بواسطة خوارزمية DE، بما في ذلك أحجام مرشح الالتفاف convolution filter ، وعدد الخلايا العصبية في الطبقة المتصلة بالكامل، ووضع التهيئة، ومعدل التسرب dropout. تم تقييم إطار عمل DE-CNN على خمس مجموعات بيانات للمشاعر العربية وحقق دقة أعلى واستهلاكاً أقل للوقت مقارنةً بأحدث الخوارزميات.

اقترح الباحثون [24] نظام ABSA الذي يستخدم LSTM ثنائي الاتجاه Bi-LSTM والحقل العشوائي المشروط (CRF) . أدى دمج Bi-LSTM مع ميزات التضمين على مستوى الحرف إلى تحسين أداء التعبير عن وجهة نظر الجانب بشكل كبير. كانت الدقة الإجمالية التي تم تحقيقها لتحديد قطبية المشاعر 82.7%.

-استخدام القواميس فقط:

قدم الباحثون [25] دراسة استخدمت فقط النهج القائم على المعجم. قاموا بإنشاء معجم من 120 ألف مصطلح باللغة العربية من خلال توسيع معجم سابق[9] ثم قاموا بتقييم جودة المعجم فاستخدموا مجموعة بيانات من 300 تغريدة لكل فئة إيجابية وسلبية ومحايدة. بعد ذلك أزالوا تكرار حروف العلة، وصححوا الأخطاء الإملائية. حققوا دقة بلغت 87% وذلك اعتماداً على القاموس الخاص بهم.

-استخدام طرق هجينة:

اقترح الباحثون [26] طريقة هجينة لتحليل المشاعر باللهجة السعودية فقط. استخدموا مجموعة من الميزات التي تم تصميمها لتكون مستقلة عن اللهجة وتقييمها باستخدام طريقة اختيار ميزة رجعية feature-backward. ثم تم تطوير ثلاث نماذج تصنيف لتحليل المشاعر السعودية على تويتر ومقارنتها على النحو التالي: طريقتان(إيجابية وسلبية)، ثلاث طرق(إيجابية، وسلبية، ومحايدة)، وأربع طرق(إيجابية، وسلبية، ومحايدة، ومختلطة). حقق الباحثون في تأثير مجموعات الميزات المقترحة على جميع نماذج التصنيف المطورة، فوجدوا أن الميزات المستخرجة من معجم AraSenTi كانت موجودة في أفضل مجموعات الميزات لجميع طرق التصنيف المجربة.

قدم الباحثون [27] HILATSA وهو نهج التعلم الإضافي الهجين ل ASA، حيث كانت الفكرة الرئيسية هي تقديم أداة لتحليل المشاعر للتغريدات العربية القادرة على التأقلم مع التغيير السريع للكلمات واستخداماتها. قام الباحثون ببناء بعض القواميس الأساسية مثل قاموس الكلمات، قاموس الاصطلاحات، قاموس الرموز، وقاموس الكلمات المكثفة للمعنى. استخدم الباحثون خوارزمية المسافة levenshtein للتعامل مع أشكال الكلمات المختلفة والأخطاء الإملائية. أما بالنسبة للتصنيف، استخدموا خوارزميتين للتعلم الالي (SVM,LR) ونموذج التعلم العميق (RNN). أظهرت التجارب نتائج واعدة من خلال إظهار الدقة العالية والأداء الموثوق في البيئة الديناميكية.

7. القسم العملي:

هناك العديد من المشاكل التي واجهت الباحثين عند تحليل المشاعر باللغة العربية، لذلك سنوضح في هذا القسم كيف يمكن للمعالجة الأولية تحسين نتائج تحليل المشاعر.

1- مجموعة البيانات: سنستخدم خلال هذا البحث مجموعتي بيانات:

- المجموعة الأولى:[28] وهي مجموعة بيانات كبيرة لمراجعات باللغة العربية(LABR)، تتضمن حوالي 63000 مراجعة لكتب جُمعت من موقع goodreads.com . يتم تقييم كل كتاب على هذا الموقع ضمن المجال من 1 إلى 5 ومن ثم يتم تحويلها إلى قطبية مشاعر محددة. يتم استخدام مجموعة البيانات هذه لعملية تصنيف المشاعر. تتألف كل مراجعة review من: تقييم المستخدم rating، رقم المراجعة review id، رقم المستخدم user id، رقم الكتاب book id ، ونص المراجعة review.

-المجموعة الثانية: [29] وهي مجموعة بيانات باللغة العربية مأخوذة من مواقع التواصل الاجتماعي تتضمن حوالي 1200 سجل، ومقسمة إلى ثلاث فئات (إيجابي، سلبي، أو محايد).

2- عمليات المعالجة الأولية: سيتم تطبيق عمليات المعالجة الأولية على الدخل وذلك على عدة مراحل: التقطيع tokenization، التقييس Normalization، إزالة كلمات التوقف، إعادة الكلمات إلى أصلها Stemming.

بدايةً ستتم عملية التقييس في بحثنا هذا تبعاً لعدة قواعد وهي: حذف إشارة "_" ، الصوتيات، علامات الترقيم، الأعداد، المحارف الخاصة، والأحرف المكررة. مثلاً سنستبدل كل المحارف (أ، إ، آ) بالحرف (ا) والأحرف (ئ، ؤ) بالحرف (ء).

أما فيما يخص عملية إعادة الكلمات لأصلها Stemming فهي عملية إزالة اللواحق من الكلمة لينتج لدينا الجذر. سنستخدم في هذا البحث ثلاث طرق لإعادة الكلمات إلى

أصلها: الأولى للتجذيع الخفيف(تاشفين tashaphyne) والثانية التجذيع للوصول إلى الجذر(Khoja stemmer)، أما الثالثة فهي للتجذيع الخفيف (Assem's Arabic) .

يعيد Khoja الجذر المؤلف من ثلاث محارف للكلمة وأحياناً أربعة أو أكثر، بينما يقوم تاشفين بتجذيع الكلمة، استخراج الجذر، تقطيع الكلمة إلى جميع الحالات الممكنة، تنميط الكلمة، يعطي قائمة مسبقة للزوائد وحروف الزيادة مع إمكانية تعديل هذه القوائم [30] . وبذلك نلاحظ أن تاشفين يقوم بنوعي التجذير معاً. أما فيما يخص المجذر الثالث فهو للتجذيع الخفيف حيث يعمل على إزالة اللاحقة والبادئة من الكلمات، ويتميز بدعمه للعديد من اللغات وبسرعته مقارنة بغيره.

3- استخراج الميزات: سنعتمد على ميزة TF-idf التي تم ذكرها سابقاً.

4- القواميس: الهدف من استخدام القواميس في عملية تصنيف القطبية هو المساعدة في البحث عن كلمة معينة وتحديد فيما إذا كانت سلبية أم إيجابية أم محايدة، سنستخدم في هذه الدراسة قاموسين:

- ArSenL: تم بناءه من قبل Badaro وأخرين [31] ، وذلك من خلال دمج SentiWordNet و Arabic WordNet و المحلل SAMA .

- MPQA [32] : يعتبر هذا القاموس ترجمة لقاموس MPQ للذاتية subjectivity والمطور في جامعة Pittsburgh إلى اللغة العربية. يحوي هذا القاموس 1406 كلمة عربية إيجابية و 2361 كلمة عربية سلبية.

5- سنقوم في هذا البحث بحساب قيمة الشعور الكلي للمستند بشكل كامل أي تصنيف الشعور فيه إلى سلبي، إيجابي، محايد. يمكن تقسيم عملية تحليل المشاعر على مستوى العبارة إلى مهمتين الأولى هي التركيز على عبارات المستند من خلال علامات الترقيم والأحرف الكبيرة لمعرفة من أين تبدأ العبارة وبما أن اللغة العربية لا تحوي أحرف كبيرة إذاً سيتم التركيز على علامات الترقيم ("." ، "،" ، "!" ، "؟"). أما المهمة الثانية فهي إيجاد الشعور لكل عبارة بهدف حساب الشعور الكلي للمستند.

هناك العديد من الطرق لحساب قيمة الشعور الكلي للمستند من خلال حساب الشعور لكل عبارة، سنقدم في هذا البحث عدة طرق ونقارن نتائج كل طريقة على حدى:

الطريقة الأولى: وهي الطريقة التقليدية التي تقوم بإعطاء +1 من أجل كل كلمة إيجابية و -1 من أجل كل كلمة سلبية، وفي النهاية يتم جمع هذه القيم بهدف الحصول على قيمة الشعور الكلي.

الطريقة الثانية: سنأخذ بعين الاعتبار فقط العبارة الأولى والأخيرة في المستند لحساب القطبية وسيتم تجاهل باقي العبارات.

الطريقة الثالثة: سيتم تجاهل العبارة الأولى والأخيرة واستخدام العبارات الوسطى فقط.

6- النفي والتركيز:

لدينا قائمة من الكلمات المستخدمة لعملية النفي، فعند اكتشاف أي منها يتم فحص الكلمة التي تليها فيما إذا كانت سلبية أم إيجابية المشاعر يتم عكس قطبيتها أما إذا كانت حيادية يتم فحص الكلمة التي تليها، وبذلك في حالة النفي يتم فحص كلمتين متاليتين على الأكثر. أما في حالة اكتشاف كلمة من قائمة الكلمات التي تستخدم بهدف التركيز يتم فحص الكلمتين السابقتين وإعطائهما وزناً مضاعفاً.

7- استخدام نماذج التصنيف:

في هذه المرحلة سيتم بناء نماذج التصنيف اعتماداً على الخوارزميات الأكثر انتشاراً وهي المصنف البايزياني NB، SVM، خوارزمية الجار الأقرب KNN . تم تدريب واختبار النموذج باستخدام بايثون، وتقسيم كل مجموعة بيانات إلى مجموعات تدريب واختبار بنسبة 70:30 ، ثم ضبط المصنفات على قيمها الافتراضية.

8- التجربة والنتائج:

الهدف الأساسي من بحثنا هذا هو قياس مدى تأثير استخدام أنواع التجذيع لحساب قطبية المشاعر، حيث تم تطبيق التجارب على مجموعتي بيانات وباستخدام قاموسين كما ذكر سابقاً.

سنستخدم أربعة معايير شائعة لقياس دقة التصنيف وهي:

(1) Accuracy: يقيس هذا المعيار كم عينة تم توقعها بشكل صحيح نسبةً إلى عدد التوقعات الكلي ويُعطى وفق الصيغة التالية:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

(2) Precision: يقيس هذا المعيار نسبة العينات التي تم تصنيفها بشكل صحيح إلى العدد الكلي للعينات التي تم تصنيفها بشكل صحيح والعينات التي تم تصنيفها بشكل خاطئ misclassified، ويُعطى وفق الصيغة التالية:

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

(3) Recall: يقيس هذا المعيار نسبة العينات التي تم تصنيفها بشكل صحيح إلى العينات التي تم تصنيفها بشكل صحيح والعينات التي لم يتم تصنيفها، ويُعطى وفق الصيغة التالية:

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

(4) F–Measure: دمج للمعيارين السابقين Recall و Precision، حيث أنه كلما كانت قيمته أكبر كان الأداء أفضل، ويُعطى وفق الصيغة التالية:

$$F - Measure = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right)$$

الجدول(1): الدقة باستخدام مختلف الStemmer .

القاموس	الطريقة	LABR			Dataset2		
		Tashaphyne	Assem	Khoja	Tashaphyne	Assem	Khoja
ArSenL	1	58.1%	58%	57.3%	46.6%	43.9%	42.5%
	2	63%	59.4%	60.2%	77.2%	70.9%	64.1%
	3	59.1%	57.2%	56%	64.3%	62.4%	63%
MPQA	1	55.1%	54%	55%	59.4%	52.5%	49.3%
	2	61.7%	59.1%	58.4%	55%	53.2%	50.1%
	3	56.5%	55.2%	57.2%	61.2%	68.1%	63.2%

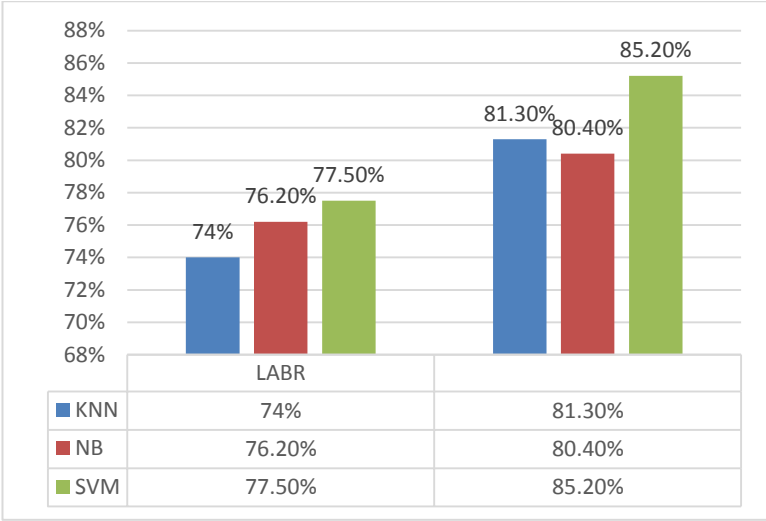
من الجدول(1) نلاحظ أن دقة المجذع الخفيف (تاشفن tashaphyne) أعلى من دقة طريقة التجذيع للوصول إلى الجذر. وبالتالي فإن تاشفن يحسن النتائج ، ويعزى ذلك كون معظم الكلمات في اللغة العربية تملك نفس الجذر مع اختلاف كبير بالمعنى أو الشعور.

اعتماداً على ماسبق سنستخدم تاشفين فقط في عملية التجذيع أثناء معالجة البيانات، وسنقارن دقة خوارزميات التصنيف الثلاثة التي تم ذكرها سابقاً كما نلاحظ في الجدول(2).

الجدول(2): مقارنة الدقة للخوارزميات الثلاثة.

DataSet	SVM	NB	KNN
LABR	77.5%	76.2%	74%
Dataset2	85.2%	80.4%	81.3%

نلاحظ من الشكل(6) أن خوارزمية SVM الأفضل من حيث الدقة بين الخوارزميات.



الشكل–6– : مقارنة الدقة للخوارزميات الثلاثة.

8. الاستنتاجات والتوصيات:

قدمنا في هذا البحث دراسة مرجعية عن تحليل المشاعر باللغة العربية، إضافةً إلى تطبيق عدد من المجذعات stemmer على مجموعتي البيانات المستخدمة لنجد أن دقة المجذع الخفيف (تاشفن tashaphyne) أعلى من دقة Khoja والمجذع الخفيف (Assem's Arabic).

فيما بعد اعتمدنا على نتيجة ماسبق وطبقنا ثلاث خوارزميات وهي المصنف البايزياني و SVM و KNN لنجد أن خوارزمية SVM هي الأفضل بالنسبة لمجموعة البيانات الأولى والثانية.

هناك مجموعة من الرؤى المستقبلية نتمنى تطويرها وهي:

– تم التصنيف في بحثنا هذا إلى ثلاث فئات سلبي، إيجابي، أو محايد. ولكن هناك الكثير من الأراء التي تحوي مشاعراً متناقضةً. فمن الممكن أن يعبر الشخص عن إعجابه بمنتج ما ولكنه غير معجب بإحدى ميزاته، أو أن تكون المشاعر مقسمة لأكثر من ثلاث صفوف (غضب، كره، حب، تفاؤل، الخ). لذلك يجب تطوير البحث ليتماشى مع الحالات المختلفة.

– اعتمدنا في بحثنا هذا على طريقة واحدة لاستخراج الميزات، لذلك يمكن لنا في البحث القادم أن نقارن بين طرق كثيرة لاستخراج الميزات للوصول للطريقة ذات الدقة الأفضل.

– اتجهت الأبحاث الحديثة إلى معالجة مشكلة السخرية الكامنة في أراء المستخدمين، وهذا ما سنتوجه إليه في أبحاثنا القادمة.

9. المراجع:

[1] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan &Claypool Publishers, May 2012.

[2] Kaur, H. and Mangat, V., 2017, February. A survey of sentiment analysis techniques. In 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 921-925). IEEE.

[3] Zhang, L., Wang, S. and Liu, B., 2018. Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), p.e1253.

[4] Joshi, M.,Prajapati, P., Shaikh, A. and Vala, V., 2017. A survey on Sentiment Analysis. International Journal of Computer Applications, 163(6), pp.34-38.

[5] Joshi A, Tripathi V, Patel K, Bhattacharyya P, and Carman M. Are word embedding-based features useful for sarcasm detection? In Proceedings of the Conference on Empirical Methods on Natural Language Processing , 2016.

[6] Rui Xia, Chengqing Zong, and Shoushan Li. Ensemble of feature sets and classification algorithms for sentiment classification.

[7] Ramy Baly, Alaa Khaddaj, Hazem Hajj, Wassim El-Hajj, and Khaled Bashir Shaban. Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in arabic Levantine tweets. arXiv preprint arXiv:1906.01830, 2019.

[8] Ashraf Elnagar and Omar Einea. Brad 1.0: Book reviews in arabic dataset. In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), pages 1–8. IEEE, 2016.

[9] Ashraf Elnagar, Yasmin S Khalifa, and Anas Einea. Hotel arabic-reviews dataset construction for sentiment analysis applications. In Intelligent Natural Language Processing: Trends and Applications, pages 35–52. Springer, 2018.

[10] Samah Alhazmi, William Black, and John McNaught. Arabic sentiwordnet in relation to sentiwordnet 3.0. 2180, 1266(4):1, 2013.
[11] Khalid Al-Rowaily, Muhammad Abulaish, Nur Al-Hasan Haldar, and Majed Al-Rubaian. Bisal–a bilingual sentiment analysis lexicon to analyze dark web forums for cyber security.Digital Investigation, 14:53–62, 2015.
[12] Bansal, P. and Kaur, R., 2018. Twitter sentiment analysis using machine learning and optimization techniques. International Journal of Computer Applications, 179(19), pp.5-8.
[13] Ortigosa, A., Martín, J.M. and Carro, R.M., 2014. Sentiment analysis in Facebook and its application to e-learning. Computers in human behavior, 31, pp.527-541.
[14] Gull, K., Padhye, S. and Jain, D.S., 2017. A comparative analysis of lexical/NLP method with WEKA's bayes classifier. International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), 5(2), pp.221-227.
[15] Lei, X., Qian, X. and Zhao, G., 2016. Rating prediction based on social sentiment from textual reviews. IEEE transactions on multimedia, 18(9).
[16] Dhande, L.L. and Patnaik, G.K., 2014. Analyzing sentiment of movie review data using Naive Bayes neural classifier. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 3(4), pp.313-320.
[17] Akhtar, M.S., Kumar, A., Ghosal, D., Ekbal, A. and Bhattacharyya, P., 2017, September. A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis.
[18] Mittal, N., Agarwal, B., Agarwal, S., Agarwal, S. and Gupta, P., 2013. A hybrid approach for twitter sentiment analysis.
[19] Muhammad, A., Wiratunga, N., Lothian, R. and Glassey, R., 2013. Domain-Based Lexicon Enhancement for Sentiment Analysis.
[2] Soni, R. and Mathai, K.J., 2015. Improved Twitter sentiment prediction through cluster-then-predict model. arXiv preprint arXiv:1509.02437.
[21] Mohcine Maghfour and Abdeljalil Elouardighi. Standard and dialectal arabic text classification for sentiment analysis. In

International Conference on Model and Data Engineering, pages 282–291. Springer, 2018.
[22] Khaled Mohammad Alomari, Hatem M ElSherif, and Khaled Shaalan. Arabic tweets sentimental analysis using machine learning. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pages 602–610. Springer,2017.
[23] Abdelghani Dahou, Mohamed Abd Elaziz, Junwei Zhou, and Shengwu Xiong. Arabic sentiment classification using convolutional neural network and differential evolution algorithm, 2019.
[24] Mohammad Al-Smadi, Bashar Talafha, Mahmoud Al-Ayyoub, and Yaser Jararweh. 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):2163–2175, 2019.
[25] Mahmoud Al-Ayyoub, Safa Bani Essa, and Izzat Alsmadi. Lexicon-based sentiment analysis of arabic tweets. International Journal of Social Network Mining, 2(2):101–114, 2015.
[26] Nora Al-Twairesh, Hend Al-Khalifa, AbdulMalik Alsalman, and Yousef Al-Ohali. Sentiment analysis of arabic tweets: Feature engineering and a hybrid approach. arXiv preprint arXiv:1805.08533, 2018.
[27] Kariman Elshakankery and Mona F Ahmed. Hilatsa: A hybrid incremental learning approach for arabic tweets sentiment analysis, 2019.
[28] http://www.mohamedaly.info/datasets/labr
[29http://saifmohammad.com/WebDocs/Arabic-Sentimnt Corpora/bbn_shared-2.xls
[30] T. Zerrouki, Tashaphyne,Arabic light stemmer, https://pypi.python.org/pypi/ Tashaphyne/0.2
[31] Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. A large scale arabic sentiment lexicon for arabic opinion mining,2014.
[32]http://mpqa.cs.pitt.edu/lexicons/
http://nlp4arabic.blogspot.com/2012/05/arabic-mpqa-subjective-lexicon-arabic.html