COMPARISON STUDY OF AUTOMATIC CLASSIFIERS PERFORMANCE IN EMOTION RECOGNITION OF ARABIC SOCIAL MEDIA USERS

Abdullah Adnan Daood
MSc.Thesis - abdullah_64730@svuonine.org

MWS_ATH_C3_S15

supervisor:
Dr. Nada Mohammad Rifai Ghneim
Ph. D (Syrian Virtual University)
T_nghneim@svuonline.org

Co-supervisor:
Eng. Issam Mohamad Salman
MSc (Czech Technical University in Prague)
issam.salman@fjfi.cvut.cz

Syrian Arab Republic
Ministry of Higher Education
Syrian Virtual University

الجمهورية العربية السورية
وزارة التعليـم العـالـــــي
الجامعة الافتراضية السورية

رقم المشروع: MWS-S15

اسم المشرف:	Dr. Nada Ghneim
اسم المشرف المساعد	MSc (Czech Technical University in Prague)،Ing. Issam Salman
بريد المشرف:	nada.ghneim@gmail.com & t_nghneim@svuonline.org issam.salman@fjfi.cvut.cz
اسم الطالب:	Ing. Abdullah Adnan Daood
بريد الطالب:	Abdullah_64730@svuonline.org
	S15

عنوان المشروع باللغة العربية:	دراسة مقارنة لأداء المصنفات الآلية في تحديد مشاعر مستخدمي وسائل التواصل الاجتماعي باللغة العربية.
عنوان المشروع بالإنكليزية:	Comparison study of automatic classifiers performance in Emotion Recognition of Arabic Social Media Users.

Contents

الملخص

مع الانتشار الواسع لشبكات التواصل الاجتماعي وتأثيرها المتزايد في حياتنا اليومية يوماً بعد اخر، أحدثت هذه الشبكات تحولاً جذرياً في توجه الويب ليصبح التركيز فيه على الإنفتاح وحرية المعلومات والمحتوى المقدم من قبل المستخدمين user-generated content. ومما ساعد على ذلك تنوع وسائل ولوج المستخدمين إلى شبكة الانترنت مع انتشار أجهزة الموبايل والأجهزة اللوحية المختلفة. فأصبحت منصات شبكات التواصل الإجتماعية هي المصدر الأساسي للأخبار والمعلومات بما فيها من أفكار وآراء ومشاعر ومواقف وردود أفعال يبديها المستخدمون، كل ذلك أدى إلى تحفيز الجهود البحثية لمحاولة تحقيق مختلف أشكال الاستفادة والإستثمار لهذه البيانات، وهذا ما عرف لاحقاً بـ "علم دراسة وتحليل البيانات الإجتماعية".

تعتبر طرائق التنقيب في النصوص الـ Text Miningمن أهم الطرق لدراسة وتحليل البيانات الإجتماعية وذلك بما تقدمهُ من أدوات مختلفة وخاصة ما يتعلق بمجال "تحليل المشاعر والآراء Emotional Analysis" وذلك من خلال استخدام الطرق التحليلية والحسابية ومعالجة اللغات الطبيعية للتعرف على ميول وآراء وعواطف المستخدمين.

ولكن وعلى الرغم من الاستخدام الكبير للغة العربية ضمن شبكات التواصل الاجتماعية، إلا أن الجهود البحثية تفتقر لأي نتاج نوعي يحقق استفادة حقيقة من البيانات المكتوبة باللغة العربية. وعليه فإن هذه الأطروحة تقدم دراسة لتحليل العواطف من النصوص المكتوبة باللغة العربية – اللهجة المحكية السورية، أجريت هذه الدراسة على تعليقات المستخدمين في موقع التواصل الإجتماعي تويتر لتخلص إلى تطوير نظام يتيح طرائق فعالة لتحليل العاطفة والتعرف عليها، يمكن أن تستخدم في الأبحاث والتطبيقات على حد سواء.

استخدم في هذه الدراسة ما يزيد عن 1320 تغريده وتعليق باللغة العربية – اللهجة المحكية السورية. تم جمعها بشكل آلي وتصنيفها يدويا وذلك لبناء نموذج تصنيف للمشاعر والذي أحرز دقة 66.9% عند اختيار أسلوب Cross-Validation وذلك باستخدام المصنف Random Forest.

تم تقسيم هذه البحث إلى ستة فصول:

1- الفصل الأول يتضمن مقدمة البحث وتفصيلاً عن مفهومي تحليل الشعور وتحليل العاطفة.
2- الفصل الثاني يقدم شرحاً لخوارزميات التصنيف.
3- الفصل الثالث يقدم دراسة مرجعية لمنهجيات التعرف على المشاعر من النصوص.
4- الفصل الرابع ينفرد بتقديم النظام المقترح والصعوبات التحديات ضمن البحث وخطة العمل.
5- في الفصل الخامس نقدم الدراسة العملية وصولاً إلى النتائج.
6- أما في الفصل السادس نقوم بدراسة وتقييم النتائج.

الفصل الأول – مقدمة البحث

يعتبر مجال تحليل العاطفة أو الشعور والآراء من المجالات الواعدة في ميدان التنقيب عن البيانات لما يعود به من خدمات وفوائد جمّة لمتخذي القرار في المؤسسات السياسية والإجتماعية والمالية والتسويقية على حد سواء. وقد تضاعف دور هذا المجال مع الأهمية المتزايدة التي تكتسبها شبكات التواصل الإجتماعي وما تعرضه هذه الشبكات من تعليقات لمستخدميها يعبرون من خلالها عن آرائهم وأفكارهم وتقييمهم للمنتجات التي يشترونها، أوالكتب التي يقرؤونها، والخدمات التي يحصلون عليها، حتى أصبح أي حدث أو موضوع ينتشر على هذه الشبكة عرضة لإصدار الأحكام والقطبية في الآراء. وأصبحت محاولة استقطاب الجمهور الإلكتروني العريض من المهام الأولى للشركات التسويقية والمؤسسات المختلفة التي تحاول بناء سمعة حسنة لكسب مزيد من الزبائن أو العملاء، وتحقيق مزيد من الأرباح والفوائد وهذا ما يصعب تحقيقه دون وجود أدوات ووسائل فعالة لفهم آراء ومواقف الزبائن تجاه أي خدمة أو منتج أو أداء تقوم به هذه المؤسسات. ولهذا فقد حاز مجال تحليل العاطفة والآراء أهمية كبيرة من قبل الباحثين والدارسين وذلك لبناء أدوات وتقنيات تساعد منظمات الأعمال لبناء سمعتها على أساس فهم عميق لرغبات وميول زبائنها. ومما ساعد ايضاً على نشاط الأبحاث في هذه المجال، توفر الأدوات المختلفة لمعالجة النصوص في اللغة الإنكليزية من أدوات تقطيع النص إلى كلمات، مروراً بفلترتها، ثم تجريدها من أحرف الزيادة تمهيداً لمعالجتها. وذاد على ذلك توفر أدوات ذكية في معالجة اللغة الإنكليزية، هذه الأدوات قادرة على تحديد بداية ونهاية كل جملة في النص، وتحليل الجملة إلى مكوناتها من فعل وفاعل ومفعول به، بالإضافة إلى توفر القواميس والبرامج والتطبيقات والتي غالبا ما تكون متوفرة جميعها باللغة الإنكليزية.

أما فيا يخص اللغة العربية وخصوصاً اللهجات المحكية، فإن هكذا أبحاث ما تزال تقتصر على بعض الدراسات الخجولة في مواضيع محددة والتي غالبا ما تبقى في نطاق التوصيف والمقارنات والاقتراحات دون أن تشكل أساساً عملياً لبناء تطبيقات حقيقية وفاعلة في معالجة هذه اللغة، والسبب في ذلك هو عدم توفر الأدوات المناسبة التي تساعد الباحثين بالإضافة إلى صعوبة قواعد اللغة العربية وعدم إلمام أغلب الباحثين بهذه القواعد.

بشكل عام، نستطيع القول بأن المهمة الأساسية في تحليل الشعور(1) من النص هي تصنيف الدلالة الشعورية لنص ما في وثيقة أو جملة إذا كان الراي المعبَّر عنه إيجابي أو سلبي أو محايد. حيث يتم العمل على مستوى الوثيقة Document level أو على مستوى الجملة sentence level. يجدر الإشارة إلى أن تحديد الشعور على مستوى الجملة يعاني من صعوبة اعتماد دلالة الكلمات على السياق بشكل كبير، أما بالنسبة للعمل على مستوى الوثيقة فصعوبته تكمن بأن الوثيقة الواحدة قد تحوي على مجموعة آراء متناقضة عن نفس الهدف.

أما بشكل خاص فأن المهمة المتقدمة في مجال التعرف إلى الشعور العام التي يهدف اليها النص، هي التخصص في عملية تصنيف الحالات العاطفية Emotion Analysis للكشف عن دلالات عاطفية بشرية أعمق مثل "غضب، قرف، ثقة، حزن، فرح، تفاجؤ .. "، وهذا ما سنقدمه في هذه الاطروحة وذلك من خلال دراسة نصوص مكتوبة باللغة العربية – اللهجة المحكية السورية ومحاولة التقاط أو تحديد الصنف العاطفي ضمن هكذا نصوص وذلك بالاستفادة من أدوات معالجة اللغات الطبيعية والعلوم الرياضية وطرائق ومنهجيات التصنيف.

1-1- تحليل الشعور Sentiment Analysis من النص

تحليل الشعور (1) Sentiment Analysis أو التنقيب عن الرأي opinion mining من النص، هو استخدام كل من تقنيات معالجة اللغة الطبيعة واللغويات الحاسوبية لتحديد واستخراج المعلومات الموضوعية من المصادر النصية المختلفة. يهدف تحليل الشعور بشكل عام لتحديد موقف الكاتب بالنسبة لفكرة ما أو تحديد القطبية العامة لوثيقة ما (إيجاباً أو سلباً). هذا الموقف قد يكون التقييم الشخصي أو الحالة الشعورية (للمؤلف)

1-1-1- تطبيقات تحليل الشعور

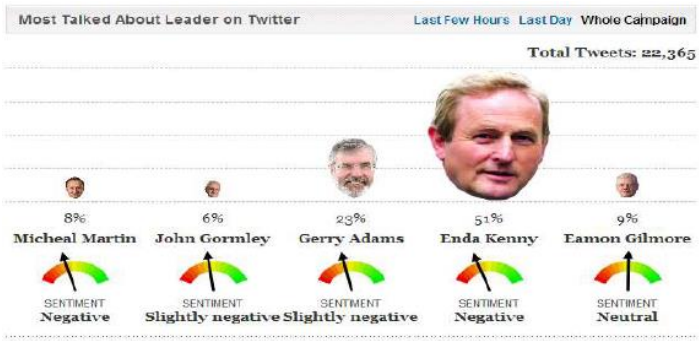
يوما بعد يوم يزداد الإهتمام بشكل كبير بهذا المجال بسبب العديد من التطبيقات المهمة فمثلا:

a- ترغب الشركات عادة بعد إصدار منتجاتها بمعرفة رأي الناس في هذه المنتجات،

وتستخدم لأجل ذلك Sentiment Analysis tools، فبدلاً من أن تضطر هذه الشركات للغوص في آلاف الآراء فإنها تقوم باستخدام هذه الأدوات التي تلخص مدى رضى الزبائن بهذه المنتجات وما هي ملاحظاتهم عليها، وبالتالي إمكانية التفاعل مع الزبائن والحصول على معلومات "Feedback" منهم بفاعلية كبيرة. الشكل 1.

الشكل 1: إحصائية لآراء مستخدمي كاميرا HP Officejet

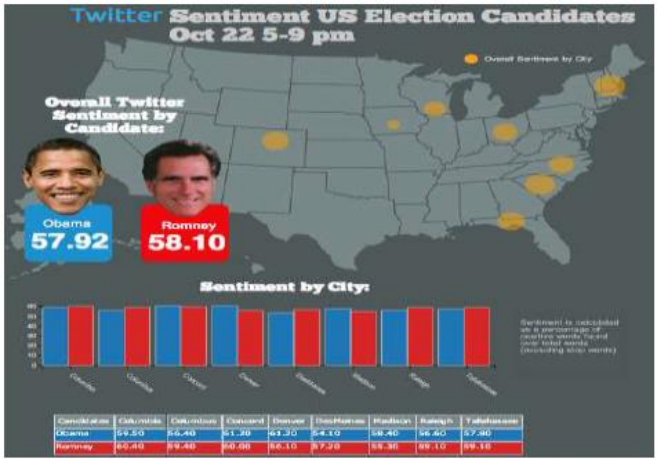
b- إن الإنتشار الكبير لوسائل التواصل الإجتماعي social media واستخدام الناس لها كوسيلة للتعبير عن آرائهم، أدى لانتشار دراسات كثيرة تحدد مما يكتبه الشخص أو يشاركه على صفحته الشخصية فيما إذا كان هذا الشخص ذو شخصية مائلة إلى الإيجابية أو السلبية أو الحيادية، ويمكن من خلالها معرفة إذا كان شخص ما ذو تأثير في محيطة من خلال آرائه وانطباعاته الشخصية وبالتالي التأثير في آرائهم. يمكن أيضا الإستفادة من ال sentiment analysis in social media للقيام بدراسات من أجل قياس معدل السعادة لمجتمع ما، منطقة أوبلد معين. الشكل 2.



الشكل 2: إحصائية على شبكة تويتر لمدى رضى الناس عن بعض القادة السياسيين

c- استخدام تحليل الآراء sentiment analysis في المجال السياسي:

إذ يمكن من خلال تحليل المنشورات على مواقع التواصل الإجتماعي معرفة رأي الناس بمرشح معين للإنتخابات، أو تحديد المناطق التي تدعم مرشح ما أكثر من الآخر، وحتى التنبؤ بنتائج الانتخابات. الشكل 3.



الشكل 3: يوضح شعبية كلا من مرشحي الرئاسة الامريكية وفقا لما يتفاعل به الناس على تويتر

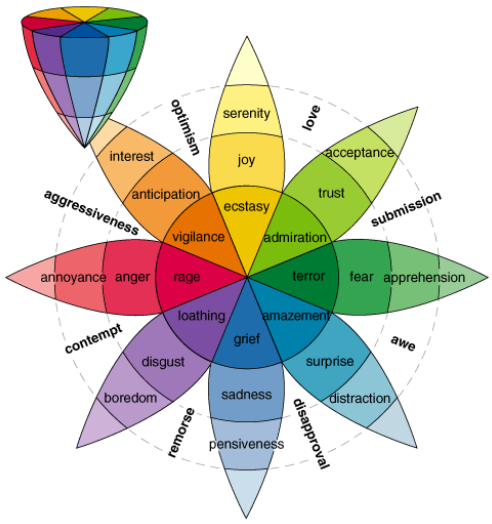
2-1- تحليل العاطفة Emotion analysis من النص

إن مفهوم الكشف عن العاطفة "حزن، فرح، غضب،.. " التي تشير إليها النصوص قد اجتذب اهتماماً متزايداً وذلك نظراً لتطبيقاته المفيدة. على سبيل المثال، يمكن تطوير علاقة قوية بين الإنسان والحاسوب من خلال التفاعل النصي المتبادل، كما يمكن كشف رؤى مثيرة للإهتمام من خلال تحليل التغريدات والمدونات التي يحررها المستخدمون.

قام العالم Ekman ضمن دراسة (2) بعنوان "Facial expression and emotion" بتمثيل المشاعر التي يبديها الإنسان ضمن ستة فئات عاطفية "السعادة، الحزن، الغضب، الخوف، التفاجؤ، القرف" تماماً كما أشار اليها العالم Parrott في عام 2001 في كتاب (3) بعنوان "Emotions in Social Psychology" يشرح فيه العاطفة البشرية ويقوم بتصنيف العواطف البشرية من خلال التسلسل الهرمي العاطفي ضمن ست فئات على المستوى العام.

ظهرت فيما بعد دراسات توسعت بهذه العواطف الأساسية إلى مجموعات "مفردات عاطفية" جزئية أكبر - مثلما قام العالم Robert Plutchik بشرحه (4) ضمن Robert Plutchik's wheel of emotions إذ قام بافتراض ثمانية قطاعات رئيسية للإشارة إلى أن هناك ثمانية أبعاد للعاطفة الأولية وهم: الغضب، الترقب، الفرح، الثقة، الخوف، المفاجأة، الحزن والاشمئزاز. ثم قام بعد ذلك بإضافة بُعد أخر يمثل كثافة العاطفة، يتم ضمن هذا البُعد تكثيف الحالة العاطفية إلى واحد من القطاعات الرئيسية. على سبيل المثال، الشعور بالملل قد يكثّف إلى الكذب إذا تُرك دون رادع.

بعد ذلك تم إضافة بُعد خاص بالعلاقات العكسية، وفيه يتم إرفاق كل قطاع من القطاعات الرئيسية بعاطفة عكسية، "عكس الحزن هو الفرح، وعكس الثقة هو الاشمئزاز". وهكذا فإن العواطف من أي لون تمثل العاطفة التي هي مزيج من العواطف الأساسية. على سبيل المثال، التوقع والفرح يتحدان إلى التفاؤل. الفرح والثقة يتحدان إلى الحب. الشكل 4



Robert Plutchik's wheel of emotions:4 الشكل

2-1-2- تطبيقات تحليل العاطفة

بالنسبة للأسواق التجارية المختصة بمجال الكشف عن المشاعر والتعرف عليها، فمن المتوقع(30) أن ينمو حجم ايرادات الكشف عن المشاعر والإعتراف بها بمعدل سنوي يبلغ 39.9% في الفترة من 2016 الى 2021 ضمن الأسواق العالمية، ومن المتوقع أن تشهد منطقة آسيا والمحيط الهادئ فرص نمو واسعة خلا السنوات القليلة القادمة. تم تقسيم أسواق الكشف عن العواطف على أساس التكنولوجيا المستخدمة، البرمجيات، الخدمات المقدمة، مجالات التطبيق، المستخدم النهائي، والمنطقة الجغرافية للمستخدمين.

فيما يلي بعض أهم الشركات العالمية التي تقدم خدمات تجارية للكشف عن المشاعر والتعرف اليها:

- Affective (U.S.)
- Beyond Verbal (Israel)
- Crowd Emotion Ltd. (U.K.)
- Emotient، an Apple company (U.S.)
- Eyeris (U.S.)
- Kairos Ar. Inc. (U.S.)
- Noldus (Netherlands)
- nViso (Switzerland)
- Realeyes (U.K.)
- Sentiance (Belgium)
- Sightcorp (Netherland)
- SkyBiometry (Lithuania)

إن مسألة تعرف العواطف من النصوص تعتبر مسألة تصنيف للنص إلى إحدى أصناف العواطف المدروسة، لذا سنقدم في الفصل التالي دراسة لأهم خوارزميات التصنيف المستخدمة في هذا المجال.

الفصل الثاني – خوارزميات التصنيف

2-1- مقدمة الفصل

في حال دراسة أساليب وطرق تحقيق تحليل العاطفة من النص Emotional analysis from text فإنه لابد من التطرق لاستعراض منهجيات التصنيف والتنقيب عن البيانات. والتي تنقسم (5) إلى:

1. Supervised learning (Classification)

تكون أمثلة التدريب مصنفة حسب الصنف التي تنتمي إليه (الدلالة العاطفية مثلاً) وممثلة بشعاع من السمات Features. ثم يتم استخدام هذه الأشعة من قبل المصنف كبيانات تدريب حيث أن مجموعة معينة من الـ Features تؤدي إلى ظهور صف معين "عاطفة" كخرج للمصنِف وذلك بتطبيق أحد خوارزميات التصنيف مثل:

- Naive Bayesian Classifier
- Random Forest
- Support vector machine (SVM)
- Entropy

بشكل عام تقوم خوارزميات Supervised learning بتحليل أمثلة التدريب. ومن ثم تطبيق مجموعة التوابع التي ستقوم الخوارزميات باستخدامها لتوليد أمثلة جديدة وبالتالي الكشف عن معارف جديدة.

المشكلة هنا أن خوارزميات هذه الطريقة تشترط أن تكون أمثلة التدريب تحتوي على مجموعة كبيرة من النصوص ذات الدلالات العاطفية "كما في حالتنا في هذه الأطروحة" والتي بدورها يجب أن تحتوي على الكم الكافي من المؤشرات العاطفية الخاصة باللغة.

يندرج تحت هذا النوع أنواع فرعية أخرى وذلك بحسب الخرج المطلوب من نظام تعلم الآلة، من أهم هذه الأنواع:

التصنيف (classification): وهو النوع الأكثر استخداماً في تعلم الآلة، في هذا النوع يكون الدخل مصنفاً إلى نوعين أو أكثر وهدف عملية التعلم هذه هو إنتاج نموذج يستطيع تصنيف أي دخل جديد إلى نوع أو أكثر من الأنواع المعروفة سابقا. أهم أمثلة هذا النوع هو عملية تصنيف البريد الإلكتروني وعملية التعرف على الوجوه.

الإنحدار (Regression): هذا النوع شبيه بالتصنيف إلا أنه يتنبأ بقيم مستمرة بدلاً من التبوء بأصناف منفصلة، هناك العديد من التطبيقات لهذا النوع أيضاً كالتنبؤ بأسعار البورصة والتنبؤ بعمر شخص يشاهد مقطع فيديو والتنبؤ بدرجة الحرارة في داخل مبنى ما اعتماداً على معلومات الطقس والوقت والحساسات الموجودة.

2. Unsupervised learning (Clustering)

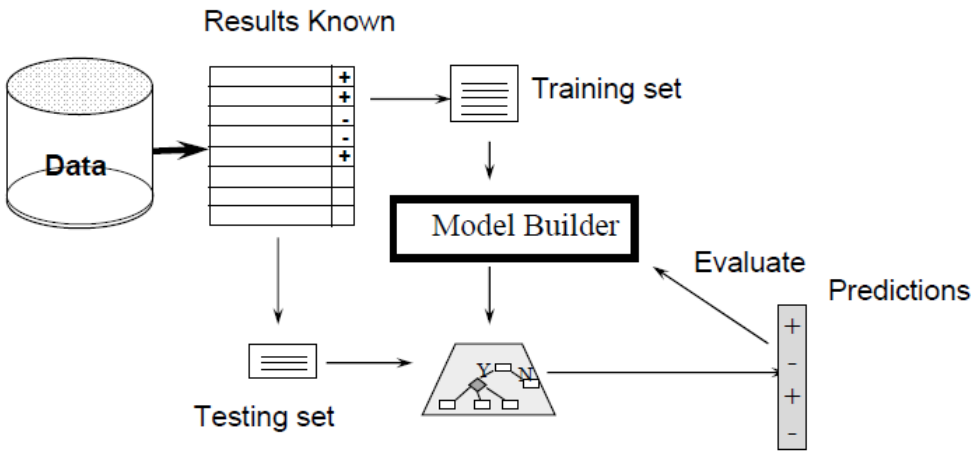
هذه الخوارزميات تقوم بتحليل النص والعمل على استخراج دلالات العاطفة منه وجمع هذه الدلالات ضمن عناقيد. يُستخدم لذلك معاجم للشعور Emotional lexicons وفيه توضع كل كلمة يقابلها semantic intensity كرقم يدل على الصف الخاص بها، ثم يتم استخدام هذا المعجم لاستخراج كلمات الشعور من الجملة واستخلاص دلالتها العاطفية من مجاميع ال semantic intensity لكلمات الجملة. والهدف هنا هو استنباط نماذج جديدة وعلاقات خفية بين البيانات، من أهم فروع هذا النوع:

التجميع أو العنقدة (clustering): من أهم تطبيقاته، تعلم حركات الشخص الواقف أمام كاميرا فيتم تسجيل تحركاته بحيث يستطيع النظام لاحقاً التعرف على هذه الحركات وربطها بردود فعل مناسبة، ومن التطبيقات الأخرى في مجال التجارة الإلكترونية عملية تجميع المستخدمين في مجموعات بناء على عمليات الشراء التي قاموا بها وسلوك التصفح الخاص بهم، ومن ثم استخدامها لإرسال رسائل إعلانية موجهة بحسب كل مجموعة.

3. Reinforcement Learning

التعلم التعزيزي (6) (Reinforcement learning): في هذا النوع يتم تعلم كيفية التصرف عند حدث معين من خلال إعطاء إشارات ترمز إلى مكافئة أو عقاب بناءاً على السلوك الحالي.

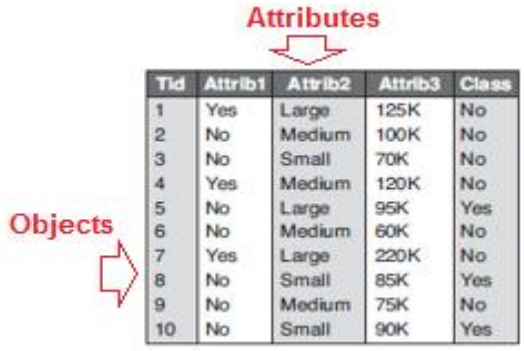
إن عملية التصنيف تتم من خلال خطوتين (بناء النموذج، استخدام النموذج)، الشكل 5:



الشكل 5 :خطوات تدريب المصنف واستخدام النموذج

- الخطوة الأولى: بناء النموذج model construction

"أمثلة التدريب – الشكل 6": وهي عبارة عن مجموعة من البيانات "الأغراض" لكل غرض من هذه الأغراض مجموعة من الواصفات Attributes



Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

الشكل 6: مثال عن أمثلة التدريب

يوجد عدد مختلف من أنماط الواصفات:

- اسمي، Nominal:
  الرقم المعرف، لون العيون، zip codes
- قيم ترتيبية، Ordinal:
  علامات، طول (طويل، متوسط، قصير)
- مستمرة، Interval:
  تاريخ، الحرارة بالدرجة المؤية أوالفهرنهايت
- نسبية، Ratio:
  الوقت، عدادات، طول بالسنتي متر

انطلاقاً من أمثلة التدريب training set نسعى لتدريب خوارزمية تصنيف بهدف الوصول إلى خرج نسميه نموذج model هذا النموذج عبارة عن مجموعة من القواعد التي نسميها knowledge base or Model والتي بدورها تعبر عن المعرفة. بناء هذا النموذج يتم من خلال خوارزميات خاصة بالتصنيف "سنقوم باستعراضها تباعا". وهكذا، ستكون المعرفة الناتجة عن النموذج أوالـ Model واحداً من الأشكال التالية:

- Classification rules

- Decision trees
- Mathematical formulae

- الخطوة الثانية: استخدام النموذج model usage

بالنسبة لاستخدام المعرفة التي تَوصلنا إليها "Model" يجب أولاً احتساب درجة الثقة أو الدقة Accuracy[1] التي سيتم وفقها العمل مع هذه المعرفة والتي تمثل مدى وثوقيه العمل مع النموذج.
يتم احتساب هذه الدقة من خلال تخصيص مجموعة من أمثلة التدريب training set وعادة تكون بنسبة 20-30% من مجمل الأمثلة المستخدمة كدخل للخوارزمية. نسمي هذه المجموعة بـ Test set، ونستخدمها ضمن عملية احتساب الدقة وذلك من خلال تطبيق المعرفة في عملية التعرف على هذه الأمثلة "كما سنشرح لاحقا".

2-2- خوارزميات التصنيف

2-2-1- Bayesian

هذه الخوارزميات تنمذج(8) مصنف احتمالي بسيط قائم على نظرية بايز مع افتراضات استقلال بين السمات Features. وتُعتبر إحدى طرق التعلم بإشراف. وهي واحدة من التقنيات الأساسية في تصنيف النصوص في تطبيقات عديدة مثل الكشف عن البريد المزعج وفرز البريد الإلكتروني الشخصي وتصنيف الوثائق والكشف عن اللغة والكشف عن المشاعر.
تستخدم هذه الخوارزميات لحل مسائل التشخيص والتنبؤ. وتمتاز بأنها لا تحتاج إلى كمية كبيرة من بيانات التدريب كما أنها تزودنا برؤية مفيدة لفهم وتقييم الكثير من خوارزميات التعليم الأخرى. كما أن هذه المصنفات مناسبة للمدخلات المتعددة الأبعاد وغالبا ما يكون أدائها أفضل في الحالات الواقعية الكثيرة التعقيد.
على الرغم من الإفتراضات المبسطة التي تفترضها هذه التقنية فإن Naïve Bayes تؤدي غرضها بشكل جيد في العديد من مشاكل العالم الحقيقي، وهي تستخدم كأساس في العديد من الأبحاث. إذ أنه وعلى الرغم من أن كثيراً من التقنيات مثل support Vector Machines، random forests، Max Entropy، Boosted trees تفوقت على Naïve Bayes إلا أن الأخير فعال جدا لأنه أقل كثافة حسابية (في كل من وحدة المعالجة المركزية والذاكرة) ويحتاج كمية صغيرة من بيانات التدريب إضافة إلى أن وقت التدريب الذي يتطلبه أقل بكثير مقارنة مع الطرف البديلة.
هناك العديد من أشكال Naïve Bayes ومن أهمها ثلاثة اشكال:

- Multinomial Naïve Bayes
- The Binarized Multinomial Naïve Bayes
- The Bernoulli Naïve Bayes

وكل منها يمكن أن يحقق نتائج مختلفة تماما لأنها تستخدم نماذج مختلفة عن بعضها البعض.
عادةً نستخدم Multinomial Naïve Bayes عندما تكون تكرارات الكلمات ذات أهمية كبرى في التصنيف. ويتم استخدام The Binarized Multinomial Naïve Bayes عندما لا تلعب تكرارات الكلمات دوراً رئيسياً في التصنيف.
وأخيرا يمكن استخدام The Bernoulli Naïve Bayes عندما يكون غياب كلمة معينة مؤثراً.

آلية عمل خوارزمية Naïve Bayes:

ليكن لدينا المجموعة Dالتي تدل على عدد من الأسطر في جدول بحيث كل سطر يمثل شعاع ب n بعد بالشكل التالي:

$$X = (X_1، X_2، X_3، .. ، X_n)$$

وليكن لدينا مجموعة الأصناف C المؤلفة من m صنف:

$$C = (C_1، C_2، C_3، .. ، C_n)$$

يتوقع مصنِف بيز أن الأسطر X تنتمي إلى الصنف Ci إذا كان $P(C_i|X) > P(C_j|X): i \neq j، 1 \leq j \leq m$
أي إذا كان احتمال وقوع Ci، أكبر من احتمال وقوع أي من الأصناف الأخرى علماً أن X قد وقع.
يمكن حساب تلك الإحتمالات باستخدام العلاقة:

$$p(C_i|X) = {p(X|C_i)p(C_i)}/{p(X)} \qquad ; i \in \{1، 2، \ldots .. ، m\}$$

[1] معادلات احتساب هذه الدقة تشرح في الفصل الثاني ضمن فقرة "تقييم اداء المصنفات"

وتكون أكبر تلك القيم من أجل الصنف Ci:

$$p(C_i|X) = max\ \{{p(X|C_i)p(C_i)}/{p(X)}\}$$

وفي حال كانت X متعددة القيم وتلك القيم مستقلة شرطياً عن بعضها البعض يكون حساب P (X |Ci) بالشكل التالي:

$$p(C_i|X) = \prod_{k=1}^{n} p(x_k|C_i)$$

والمكافئة للعلاقة:

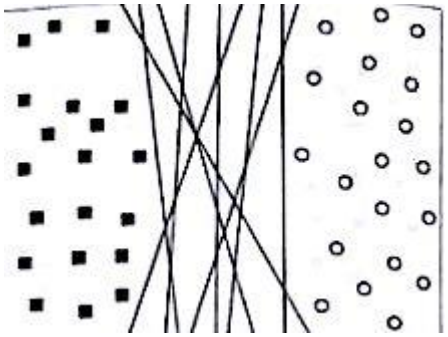
$$p(X|C) = p(x_1|C_i) * p(x_2|C_i) * \ldots\ldots * p(x_n|C_i)$$

Support Vector Machine -2-2-2

نالت خوارزمية التصنيف SVM(8) اهتماماً كبيراً ولهذه التقنية جذورها في نظرية التعليم الإحصائي وقد أظهرت نتائج تجريبية واعدة في العديد من التطبيقات العملية بدءاً بالتمييز الرقمي للكتابة بخط اليد وانتهاءً بتصنيف النص. كما تعمل هذه الخوارزمية بشكل جيد مع البيانات التي لها عدد كبير من الأبعاد. هناك جانب آخر لهذه الطريقة هو أنها تمثل حد القرار باستخدام مجموعة جزئية من أمثلة التدريب تُعرف بمتجهات الدعم support vector.

قبل دراسة أنواع SVM سنتحدث عن مفهوم المستوي الفائق ذي الهوامش القصوى maximum margin hyperplane.

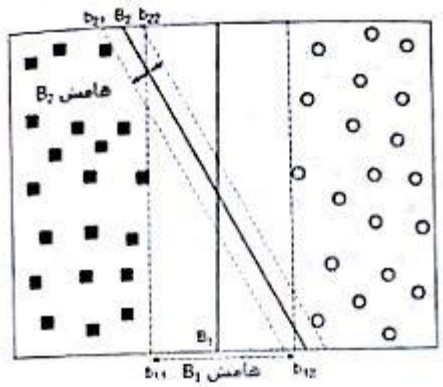
يعرض الشكل 7 مخططاً لمجموعة بيانات تحوي أمثلة تنتمي إلى صنفين مختلفين، ممثلة بالمربعات والدوائر، إن مجموعة البيانات هذه قابلة للفصل خطياً، أي أنه يمكننا إيجاد مستوي فائق بحيث تستقر كافة المربعات في جهة من المستوي الفائق فيما تستقر كافة الدوائر في الجهة الأخرى.



الشكل 7: مجموعة بيانات مفصولة خطيا تنتمي إلى صنفين مختلفين

هناك عدد غير منته من هذه المستويات الفائقة. وعلى الرغم من أن خطأ التدريب لها هو صفر، فإنه لا توجد أية ضمانات بأن المستويات الفائقة ستؤدي الغرض المطلوب بنفس مستوى الجودة على الأمثلة التي لم تُر من قبل "التعميم".

لتوضيح الصورة حول آلية تأثير الخيارات المختلفة على أخطاء التعميم، سنأخذ حدي القرار B_2،B_1 المبينان في الشكل 8، يمكن لكلا حدي القرار فصل أمثلة التدريب إلى الأصناف المقابلة لها بدون ارتكاب أية أخطاء في التصنيف. يترافق كل حد قرار Bi مع زوج من المستويات الفائقة، يرمز إليها ب bi_1،bi_2 على الترتيب. يتم الحصول على bi_1 بتحريك مستوي فائق متوازي بعيداً عن حد القرار إلى أن يلمس أقرب مربع (مربعات)، في حين يتم الحصول على bi_2بتحريك المستوي الفائق إلى أن يلمس أقرب دائرة (دوائر). تُعرَف المسافة بين هذين المستويين الفائقين بهامش (margin) المصنِف. نلاحظ من الرسم البياني في الشكل 8 أن الهامش من أجل B_1 أكبر إلى حد بعيد منه بالنسبة ل B_2. وفي هذا المثال يتحول B_1 ليصبح المستوي الفائق ذا الهامش الأقصى لأمثلة التدريب.



الشكل 8: رسم يوضح حدود القرار لمجموعة بيانات منفصلة خطياً

تنحو حدود القرار التي لها هوامش كبيرة لأن تكون لها أخطاء تعميم أفضل مقارنة بتلك التي لها هوامش أقل. وبشكل حدسي فإذا كان الهامش صغير، فإن أية تغيرات صغيرة على حد القرار يمكن أن يكون لها تأثير هام على تصنيفه. ولذلك فإن المصنفات التي ينتج عنها حدود قرار بهوامش صغيرة يكون تعميمها رديئاً على الأمثلة التي لم تُر سابقاً.

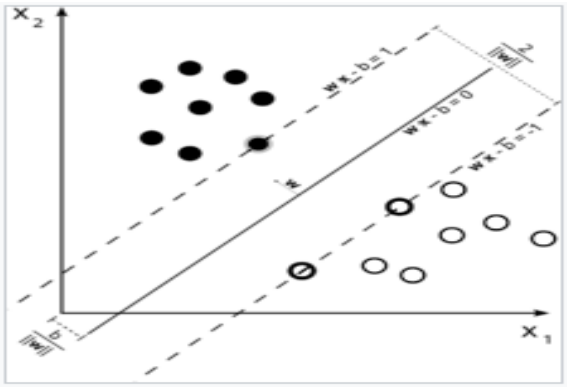
فالنماذج التي لها هوامش أصغر يكون استيعابها أعلى لأنها أكثر مرونة ويمكن أن تلائم الكثير من مجموعات التدريب، بخلاف النماذج التي لها هوامش كبيرة. على أية حال، فإنه بزيادة الإستيعاب سيزداد خطأ التعميم أيضاً. لذلك من المفضل تصميم مصنفات خطية تجعل هوامش حدود القرار الخاصة بها أعظمية بهدف ضمان أن تكون أخطاء تعميمها الأسوأ أصغرية، مثال على هذه المصنفات SVM الخطية.

- **SVM الخطية**

لتكن مجموعة بيانات تدريب تتألف من n نقطة يتم التعبير عنها بالصيغة $i = 1 \to n : (\overrightarrow{x_i}, y_i)$ حيث y_i تأخذ قيمة 1 أو1- بحيث تعبر عن الصف الذي تنتمي إليه $\overrightarrow{x_i}$.

كل $\overrightarrow{x_i}$ عبارة عن شعاع حقيقي ذو بُعد .p نريد إيجاد المستوي الفائق ذي الهامش الأقصى الذي يقسم النقاط $\overrightarrow{x_i}$ إلى مجموعة نقاط تنتمي إلى الصف $y_i = 1$ ومجموعة نقاط تنتمي للصف $y_i = -1$ بحيث يكون الهامش أكبر ما يمكن:

أي مستوي فائق يمكن أن يكتب كمجموعة من النقاط $\overrightarrow{x_i}$ بالشكل $\overrightarrow{w}.\overrightarrow{x_i} - b = 0$ حيث $\overrightarrow{w}$ هو الشعاع القياسي للمستوي الفائق والبارامتر $b/||\overrightarrow{w}||$ يحدد مقدار إزاحة المستوي الفائق.



الشكل 9. SVMالخطية

هناك حالتان للـ SVM الخطية:

1- الهامش الصلب (hard margin) :وفيه تكون البيانات قابلة للفصل خطياً ولدينا مستويين فائقين يفصلان البيانات إلى صفين لذلك تكون المسافة بينهما أكبر ما يمكن ويكون المستوي الفائق ذي الهامش الأقصى هو المستوي الفائق التموضع في منتصف المسافة بينهما. هذان المستويان يعطيان بالمعادلات:

$$w^{\rightarrow}.x_i^{\rightarrow} - b = -1 \;\&\; w^{\rightarrow}.x_i^{\rightarrow} - b = 1$$

هندسياً، إن المسافة بين المستويين هي $2/||w^{\rightarrow}||$ لذلك حتى نزيدها يجب أن ننقص $||w^{\rightarrow}||$. وأيضاً يجب أن نمنع نقاط البيانات من التواجد ضمن الهامش لذلك نضيف القيد التالي لكل نقطة

$$w^{\rightarrow}.x_i^{\rightarrow} - b \geq 1 \; if \; y_i = 1$$

أو

$$w^{\rightarrow}.x_i^{\rightarrow} - b \leq -1 \; if \; y_i = -1$$

هذه القيود تضمن أن تتواجد كل نقطة على الجانب الصحيح من الهامش. وهكذا نستطيع كتابة الشرطين كما يلي:

$$y_i(w^{\rightarrow}.x_i^{\rightarrow} - b) \geq 1, \; for \; al \; 1 \leq i \leq n$$

وللحصول على حل أمثلي نكتب الصيغة التالية

تصغير $||w^{\rightarrow}||$ تبعاً لـ $y_i(w^{\rightarrow}.x_i^{\rightarrow} - b) \geq 1, i = 1,2 \dots \dots, n$

إن $w^{\rightarrow}$ و b التي تحل هذه المشكلة هي محددات المصنف $x^{\rightarrow} \rightarrow sgn(w^{\rightarrow}.x_i^{\rightarrow} - b)$

نتيجة هذا الوصف الهندسي يتم تحديد المستوي الفائق ذو الهامش الأقصى بشكل كامل عن طريق المتجهات $x_i^{\rightarrow}$ المتوضعة بالقرب منه، وتدعى بمتجهات الدعم support vectors.

2- الهامش المرن:(soft margin) في هذه الحالة البيانات تكون غير قابلة للفصل خطياً لذلك نستخدم التابع

$$\max(0,1 - y_i(w^{\rightarrow}.x_i^{\rightarrow} - b))$$

يأخذ هذا التابع قيمة صفر إذا كان $x_i^{\rightarrow}$ تتوضع على الجانب الصحيح وإلا يأخذ قيمة تتناسب مع البعد عن الهامش. ثم نقوم بتصغير المقدار

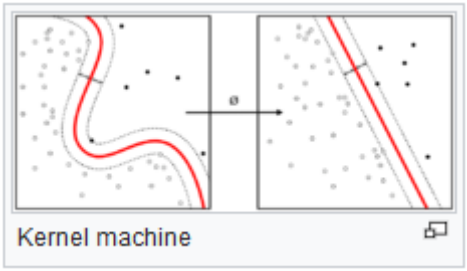
$$\left[\frac{1}{n}\sum_{i=1}^{n}\max(0,1 - y_i(w^{\rightarrow}.x_i^{\rightarrow} - b))\right] + \gamma||w^{\rightarrow}||^2$$

حيث أن البارامتر "غاما" يتخذ قرار المبادلة بين زيادة حجم الهامش والتأكد من أن تقع $x_i^{\rightarrow}$ على الجانب الصحيح.

- **SVM غير الخطية:**

تم إنشائها بتطبيق خدعة النواة للحصول على المستوي الفائق ذو الهامش الأقصى، هذه الخوارزمية مشابهة للخوارزمية الخطية ماعدا أن كل منتج نقطة dot product يستبدل بتابع نواة غير خطي مما يسمح للخوارزمية بملائمة المستوي الفائق ذي الهامش الأقصى مع تغيرات فضاء الميزة. هذا التغيير قد يكون غير خطي.

من الجدير بالملاحظة أن العمل في فضاء ميزة ذو بعد أعلى يزيد من خطأ التعميم في SVM بالرغم من أنه يعطي عينات كافية لتعمل الخوارزمية بشكل جيد.



الشكل 10: SVM غير الخطية

تحقيق SVM باستخدام خوارزمية SMO:

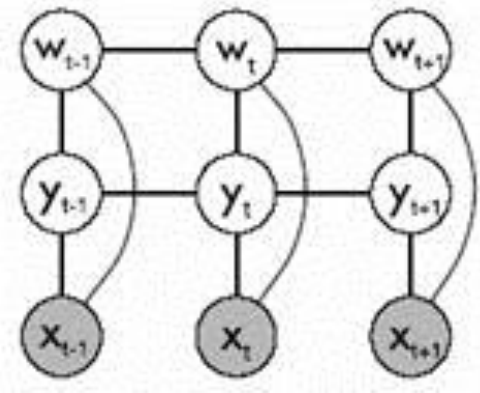
البارامترات الموجودة في المستوي الفائق ذو الهامش الأقصى يتم استنتاجها عن طريق الحل بطريقة أمثليه وخوارزمية SMO (sequential minimal optimization) هي إحدى الخوارزميات المستخدمة لإيجاد حل سريع للمشاكل، حيث تقسم المشكلة إلى مشاكل جزئية ثنائية البعد بحيث تكون قابلة للحل بشكل تحليلي، وبذلك لم نعد بحاجة لاستخدام خوارزميات أمثليه عددية وتخزين المصفوفات.

SMO خوارزمية ذات مفهوم بسيط وسهلة التحقيق وأسرع في التوليد ولديها خصائص تقييس أفضل من أجل مشاكل SVM الصعبة.

Conditional Random Field -3-2-2

هو نوع من أكثر النماذج الاحتمالية المميزة التي تستخدم من أجل تعريف معطيات متسلسلة، مثل نص لغة طبيعية أو سلاسل حيوية.

تعتمد CRF على نموذج ماركوف المخفي HMM (Hidden Markov Model) الشكل 11. لكنه أكثر قوة منه ويعتبر هذا المصنف من أحدث تقنيات تعريف السلسلة.

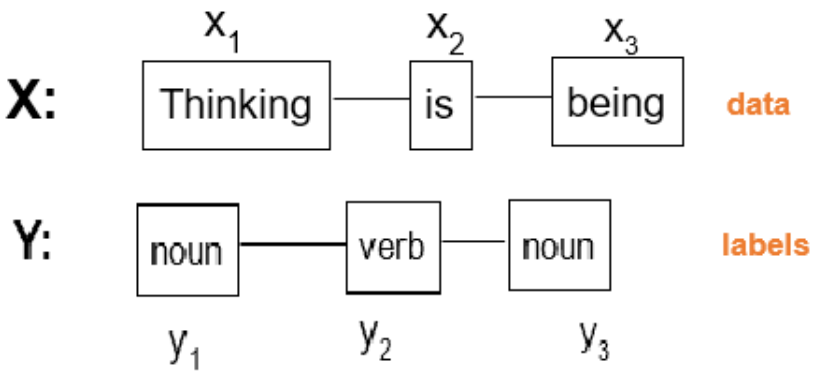


الشكل 11: مثال عن سلاسل ماركون المخفية

لنفترض X متحول عشوائي يدل على سلاسل المعطيات التي يجب تعريفها. ولنفترض Y متحول عشوائي يدل على سلاسل التعاريف الافتراضية. Y_i تفترض مجالاً يدل على مجموعة أبجدية منتهية من التعاريف:

$$\sum = \{noun، verb، adjective، and\ so\ on\}$$

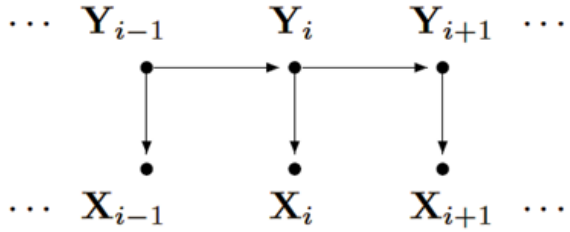
المشكلة هنا هي: تعلم كيفية إعطاء تعاريف من مجموعة مغلقة Y إلى سلسلة معطيات X.



الشكل 12: اعطاء بطاقات تعريف من مجموعة مغلقة Y إلى سلسلة معطيات X

في حال استخدام HMM - Standard tool vs the hidden Markov Model

- يتم تخصيص وصلة احتمالية من أجل كل ثنائية "سلسلة ملاحظة" بالإضافة إلى تخصيص سلاسل تعاريف خاصة بهذه الوصلات.
- يتم تدريب البارامترات بشكل نموذجي للحصول على أعظم احتمال للوصلة في أمثلة التدريب.



$$P(\mathbf{X}, \mathbf{Y}) = \prod_i P(\mathbf{X}_i | \mathbf{Y}_i) P(\mathbf{Y}_i | \mathbf{Y}_{i-1})$$

الشكل: 13 HMM Model

محاسن HMM:

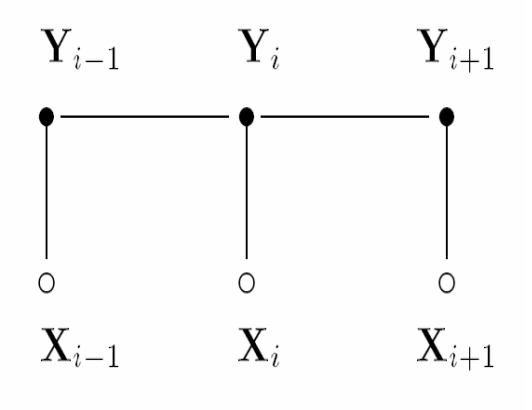
- التخمين عن طريقها سهل جداً.
- البارامترات يمكن أن يتم تخمينها بثقة عالية نسبياً من عينات صغيرة.

الصعوبات والمساوئ في HMM:

- نحتاج إلى إحصاء كل سلاسل الملاحظات الممكنة.
- ليس عملياً في تمثيل الميزات التفاعلية المتعددة أو التبعيات طويلة المدى للملاحظات.
- الفرضيات مستقلة بشكل صارم عن الملاحظات.

مقارنة بين CRFوHMM:

- ✓ CRF تستخدم الإحتمال الشرطي $P(y|x)$ حيث **y** سلسلة تعريف و **x** سلسلة ملاحظة ، أكثر من استخدام الوصلة الاحتمالية $P(y،x)$ المعتمدة من قبل HMM.
- ✓ احتمالية الإنتقال بين التعاريف في CRF ربما تعتمد على الملاحظات السابقة والمستقبلية.
- ✓ CRF تخفف الاستقلالية القوية للتخمينات الموجودة في HMM.



الشكل 14: CRF undirected and acyclic

Random Forest -4-2-2

تنمو في الغابات العشوائية Random Forest عدة أشجار تصنيف. لتصنيف غرض جديد من شعاع دخل نضع شعاع الدخل أسفل كل شجرة في الغابة. كل شجرة في هذه الغابة تعطي تصنيف ونستطيع أن نقول إن الأشجار تصوت من أجل التصنيف والغابة تختار الصنف ذو التصويت الأعلى.

كل شجرة تنمو كما يلي:

1) إذا كان عدد الحالات في مجموعة التدريب يساوي N يتم إيجاد عينة من N حالة بشكل عشوائي من المعطيات الأصلية، هذه العينة ستصبح مجموعة التدريب من أجل نمو الشجرة.
2) إذا كان لدينا Mمتحول وعدد m<<M يحدد كل عقدة، المتحولات m يتم تحديدها بشكل عشوائي، أفضل تجزئ على هذه المتحولات يستخدم لتجزئ العقد. قيمة m تبقى ثابتة خلال نمو الغابة.
3) كل شجرة تنمو إلى أقصى حد ممكن، ولايوجد اقتطاع.

ومعدل خطأ تعميم الغابة يعتمد على شيئين:

1) الارتباطات بين شجرتين في الغابة، اذ أن ازدياد الارتباطات يؤدي إلى زيادة معدل الخطأ عند التعميم.
2) طول كل شجرة في الغابة، فشجرة مع معدل خطأ منخفض هي مصنف قوي كما أن زيادة طول الشجرة يؤدي لزيادة معدل الخطأ.

إن تقليل عدد المتحولات m يؤدي إلى تناقص كل من الارتباطات والطول وزيادته تؤدي إلى ازدياد الاثنان. في بعض الأحيان يكون المجال المثالي للمتحولات m واسع جداً.

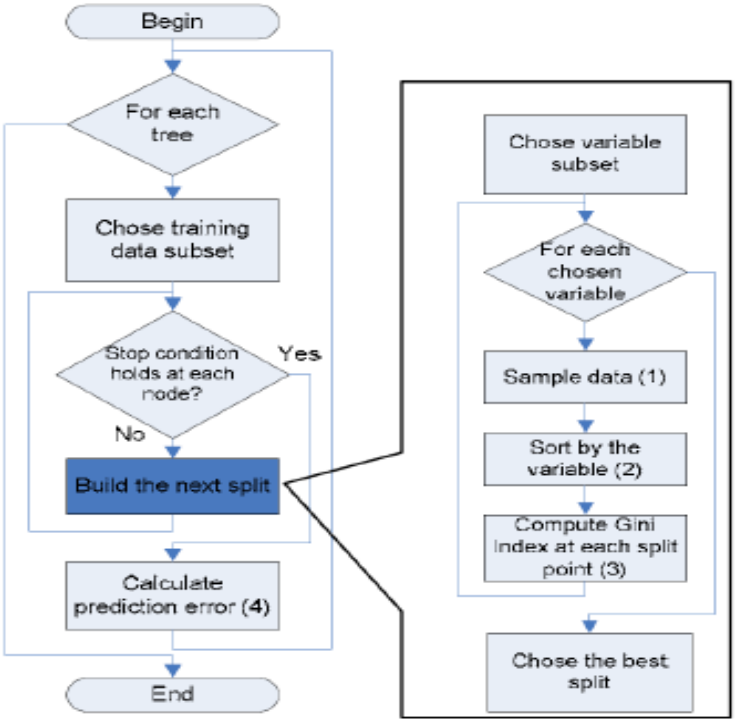
ميزات الأشجار العشوائية:

- تنفذ بكفاءة على قواعد البيانات الضخمة.
- تستطيع المحافظة على آلاف متحولات الدخل من دون حذف أي متحول.
- تعطي تخمينات حول المتحولات المهمة في عملية التصنيف.

- تولد تخمين داخلي غير متحيز لخطأ التعميم أثناء بناء الغابة.
- لديها طريقة فعالة لتخمين البيانات المفقودة عندما يتم فقدان نسبة كبيرة من البيانات.
- لديها طريقة لموازنة الخطأ في صف مجموعات المعطيات غير المتوازنة.
- الغابات المولدة يمكن أن تحفظ لتستخدم في المستقبل على معطيات أخرى.
- النماذج تُحسب بحيث تعطي معلومات عن العلاقة بين المتحولات والتصنيف.

الغابات العشوائية تنفذ على الذي نريده من الأشجار وهي سريعة، تنفذ على قاعدة بيانات تحوي 50.000 حالة و100 متحول وتنتج 100 شجرة في 11 دقيقة، من أجل مجموعات المعطيات الضخمة الذاكرة الرئيسية المطلوبة هي سعة تخزين المعطيات بحد ذاتها بالإضافة إلى ثلاث مصفوفات لها نفس بعد المعطيات.

كيف تعمل الغابات العشوائية: عندما ترسم مجموعة التدريب عن طريق أخذ العينات مع التبديل، ثلث الحالات تكون خارج العينة، وهذه البيانات تدعى out_of_bag(oob) تستخدم للحصول على تنفيذ تخمين غير متحيز لخطأ التصنيف كلما تم إضافة شجرة للغابة، وتستخدم لتخمين أهمية المتحول. بعد أن يتم بناء كل شجرة، كل المعطيات تنزل إلى أسفل الشجرة ويتم حساب التقارب لكل زوج من الحالات إذا كان لحالتان نفس العقدة النهائية سوف يزداد تقاربهما بمقدار واحد في نهاية التنفيذ التقارب يحسب بالتقسيم حسب عدد الأشجار. يعرض الشكل 15 المخطط التدفقي لهذه الخوارزمية:



الشكل 15: المخطط التدفقي لخوارزمية Random Forest

2-3- تقييم خوارزميات التصنيف

هناك مجموعة من المناظير(9) التي يتم من خلالها مقارنة وتقييم نماذج التصنيف:

- السرعة Speed

وذلك يشير إلى تكاليف عمليات الحساب والمعالجة اللازمة لتوليد واستخدام النموذج.

- المتانة Robustness

وهي درجة تأثر الخوارزمية بالبيانات غير الواضحة noise data أو البيانات غير المكتملة missing values.

- قابلية التوسع Scalability

إمكانية بناء نموذج فعال وذلك عند التعامل مع بيانات كبيرة الحجم.

- التفسير Interpretability

تعتمد مدى وضوح وفهم النتائج التي يعيدها نموذج التصنيف.

في هذه الأطروحة سنقوم بتقييم نماذج خوارزميات التصنيف وفقا لاحتساب كل من Precision(P)، Recall(R) & F measure(F) كالتالي(10) :

P = #correct guesses / #total guesses

R = #correct guesses / #total

F = 2PR / (P+R)

:

#correct guesses:

The number of Statement marked correctly as expressing an emotion X by the classifier.

#total guesses:

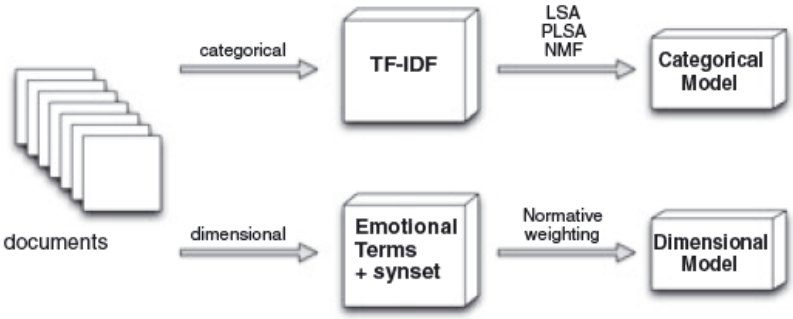
The total number of Statement that are marked by the classifier as expressing the emotion X (including correct and wrong guesses) .

#total:

The number of Statement expressing the emotion X in the dataset.

3-1- مقدمة الفصل

لدراسة وتحليل طرق الكشف عن العاطفة يجب التركيز على كيفية تصنيف الكلمات ذات الدلالات العاطفية:
وفقا لبحوث علم النفس، نجد عدد من النظريات حول كيفية تصنيف الكلمات ذات الدلالات العاطفية أهمها (11) :



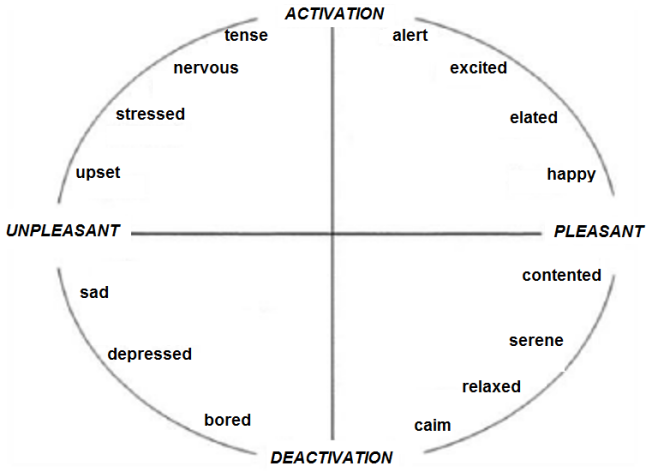
الشكل 16: طرق تصنيف الكلمات ذات الدلالات العاطفية

الفئات العاطفية (11) Emotional categories

قام مجموعة من الباحثين وبالاستعانة مع علماء نفس بتحديد ودراسة الكلمات اللغوية ذات الدلالات العاطفية، ومن ثم قاموا بتمثيلها كنقاط ضمن فضاء متعدد الأبعاد، المسافة بين نقطتين "كلمتين" تدل على درجة تشابه الدلالة العاطفية بين هاتين الكلمتين. ومن خلال دراسة هذا التمثيل البياني لوحظ عدة مجموعات "كثافات بيانية"، كل مجموعة منها تحتوي جملة نقاط "كلمات" ذات دلالات عاطفية متقاربة وبناء عليه تم الافتراض بوجود فئات عاطفة منفصلة عن بعضها البعض.
في بعض المناهج (2) تم الإتفاق على ستة فئات أساسية: "غضب، اشمئزاز، خوف، سعادة، حزن، مفاجأة"، وفي مناهج أخرى (3) تم إضافة كل من "الثقة، والترقب" إلى الفئات السابقة. كل من الفئات السابقة نظمت فيما بعد ضمن أربع مجموعات قطبية: "الفرح مقابل الحزن، الغضب مقابل الخوف، الثقة مقابل الاشمئزاز، المفاجأة مقابل الترقب"

الأبعاد العاطفية (11) Emotional dimensional

تم هنا التطوير على النموذج السابق، فمن أجل كل كلمة تم الأخذ بعين الإعتبار كل من الشدة العاطفية ومدى فاعلية التعبير التي تتمتع به هذه الكلمة وذلك من خلال دراسة المسافة البيانية بين النقطة التي تمثل هذه الكلمة عن مركز المجموعة التي تنتمي اليها. وهكذا أصبحت العواطف تتوزع ضمن مساحة دائرية ثنائية البعد – الشكل 17.



الشكل 17: شكل يوضح توزيع الابعاد العاطفية بشكل دائري ثنائي البعد

يشير البعد الأول إلى درجة العاطفة الإيجابية "Unpleasant،Pleasant" . ويشير البعد الثاني إلى مدى فاعلية العاطفة " Deactivation،Activation"

3-2- منهجيات وطرائق الكشف عن العاطفة في النص

3-2-1- تعاريف

جذب مفهوم الحوسبة الوجدانية في عام 1997 العديد من الباحثين في مجال علوم الكمبيوتر والتكنولوجيا الحيوية وعلم النفس والعلوم المعرفية وباقي العلوم المشابهة. فبدأت في الظهور أبحاث في مجال الكشف عن الدلالات العاطفية من البيانات النصية وذلك لتحديد العاطفة الإنسانية التي تشير اليها هذه البيانات. وعليه فقد تم صياغة مفهوم كشف العواطف من النصوص بالشكل التالي:

بافتراض E: مجموعة كلمات اللغة ذات الدلالات العاطفية.

A: مجموعة أسماء الكتاب أصحاب النصوص.

T: مجموعة الدلالات العاطفية المحتملة للنصوص.

R: التابع المعبر عن العاطفة E المرتبطة بالكاتب A انطلاقاً من النص T

فإن

$$R: A\ x\ T\ \rightarrow E$$

وهكذا فإن الهدف العام من منهجيات الكشف عن العاطفة من النص هي البحث عن العلاقة التي تحقق التابع R.

المشكلة الحقيقية في أنظمة التعرف على المشاعر من النص تكمن في حقيقة أنه على الرغم من أن تعريفات E وT هي تعريفات صريحة إلا أنه سيكون هناك الكثير من الحالات الفرعية لكل منهما! فبالنسبة لـ T هناك عناصر جديدة تضاف إلى اللغة بشكل دائم بسبب أن اللغات البشرية آخذه بالتطور والتوسع، وكذلك الأمر بالنسبة لمجموعة التصنيفات العاطفية، فليس هناك تصنيفات موحدة تشمل جميع المشاعر الإنسانية بسب الطبيعة المعقدة للعقول البشرية.

المهام الحالية للكشف عن العاطفة من النص تختلف وفقا لـ:

- أنواع الصفوف المتوقعة
  - عاطفة (حزن، فرح، تفاجئ، اشمئزاز، خوف، غضب)..
  - شعور (ايجابي، سلبي، محايد).
- مستويات التصنيف المستخدمة (الجملة، العبارة، النص).
- أنواع الخصائص التي يجب اعتبارها للتصنيف وتقنية التصنيف المستخدمة.
- دلالية، حيث يتم اخذ دلالة الكلمة بعين الإعتبار. لأجل ذلك يمكن الاستعانة بقواميس تشير إلى درجة دلالة كلمة اللغة ضمن كل عاطفة.
- نحوية، حيث يتم أخذ بنية الكلمة والجملة بعين الإعتبار.
  - تردد جذور الكلمات.
  - تردد الكلمات N-gram
  - وجود علامات الترقيم
- اسلوبية، حيث يتم أخذ الرموز وأوزان الكلمات "مجموع الكلمات وطولها وتوزع المحارف .... "بعين الإعتبار.

3-2-2- المنهجيات العامة

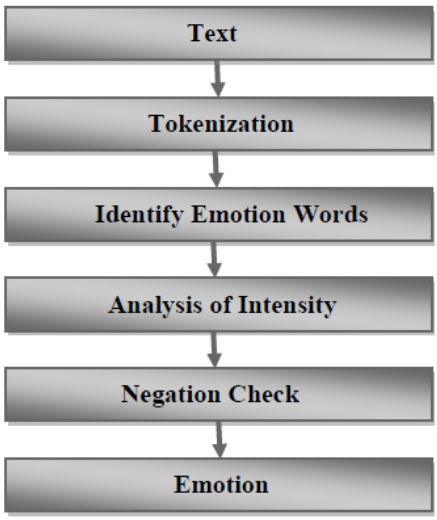
بناء على ما سبق، نستعرض الطرق المستخدمة للتعرف على المشاعر من النصوص المكتوبة:

a) Keyword Based Method (21) (20)

(b Learning –based Methods (23).
(c Hybrid Methods (21).

Keyword Based Method

في المنهجيات السابقة لعمليات الكشف عن العاطفة في النصوص، كان يتم احتساب تكرار كلمات رئيسية من مجموعة عاطفية معينة ضمن النص المدروس وذلك بالإستعانة بـ موارد معجمية Lexical resources تحتوي الكلمات اللغوية ذات الدلالات العاطفية. وكان يطلق على هذه المنهجية Keyword pattern matching. تم التحسين على هذه المنهجية فأطلق عليها فيما بعد بـ Keyword Spotting Technique وأصبح يتم الإعتماد في هذه المنهجية على الموارد المعجمية الخاصة باللغات الطبيعية لأجل الكشف عن المشاعر في النص. وأصبحت تعمل وفقا للسياق التالي – الشكل 18



الشكل 18: Keyword Spotting Technique

هذه الطريقة تتكون من خمس خطوات رئيسية:

- يتم التعامل مع الوثيقة النصية كدخل للنظام، والخرج هو فئة العاطفة الذي سيتم الكشف عنها.
- يتم تحويل كلمات النص إلى رموز Token، ومن هذه الرموز سيتم تحديد كلمات العاطفة وكشفها.
- بعد ذلك تأتي عملية تحليل كثافة "توزين" الكلمات العاطفية في النص.
- ومن ثم مرحلة تفحص حالة النفي negation check.
- أخيرا ارجاع فئة العاطفة كخرج للنظام.

يتميز الكشف عن العواطف وفقا لهذه المنهجية بأنه سهل الإستخدام، إذ يتم استخدام محلل صرفي "مجذع" Stemmer لاسترجاع كلمات النص إلى جذعها ومن ثم استخدام مجاميع "قواميس" لغوية خاصة بالمصطلحات العاطفية إلى جانب مجاميع "قواميس" لغوية خاصة بمفردات اللغة نفسها. يحتوي القاموس الأول على مجموعة الكلمات ذات الدلالة العاطفية مع وزن كل منها، بينما القاموس الثاني يحتوي كل كلمات اللغة بغض النظر عن احتوائها على دلالة عاطفية أم لا. بعد القيام بتحليل النص إلى مفرداته الرئيسية، يتم إيجاد التقارب بين المفردات والدلالات العاطفية لكل منها ومن ثم إيجاد صنف العواطف المرتبط بالنص بأفضل احتمال ممكن وذلك بشكل مستقل تماماً عن سياق الحديث.

من عيوب هذه الطريقة غياب الدلالة العاطفية الأكثر عمقا من مستوى الكلمة نفسها، فعلى سبيل المثال:

كلمة "*accident*" تشير إلى مشاعر سلبية باحتمال كبير على الرغم من أنها قد تحمل معاني مختلفة وذلك وفقا لسياق النص:

"*I have an accident*" or "*I met my friend by accident*"

تم بناء هذه الخوارزمية لتستخدم WordNet-Affect(22)وهو معجم مفرداتي ضخم خاص بالكلمات العاطفية مقسم لمجموعات متعددة، كل مجموعة تحتوي كلمات ذات معاني مترابطة بشكل مفاهيمي ومفرداتي. هذا المعجم موسع عن WordNet Domains والذي يحتوي على قرابة 200 لصاقه a-labels، كل واحدة منها تمثل صنف من أصناف المعاني الدلالية للكلمات.

A-Labels
EMOTION
MOOD
TRAIT
COGNITIVE STATE
PHYSICAL STATE
EDONIC SIGNAL
EMOTION-ELICITING SITUATION
EMOTIONAL RESPONSE
BEHAVIOUR
ATTITUDE
SENSATION

جدول 1 جدول يستعرض بعض لصاقات المعجم WordNet Domains

هذه التوسعة "*WordNet-Affect*" تقوم بالربط بين الأصناف العاطفية "*a-labels*" والكلمات اللغوية التي تشير اليها مع الأخذ بعين الاعتبار الحالات المزاجية وحالات كشف المشاعر أوالاستجابات العاطفية.

Category	Example Term
emotion	anger
cognitive state	doubt
trait	competitive
behavior	cry
attitude	skepticism
feeling	pleasure

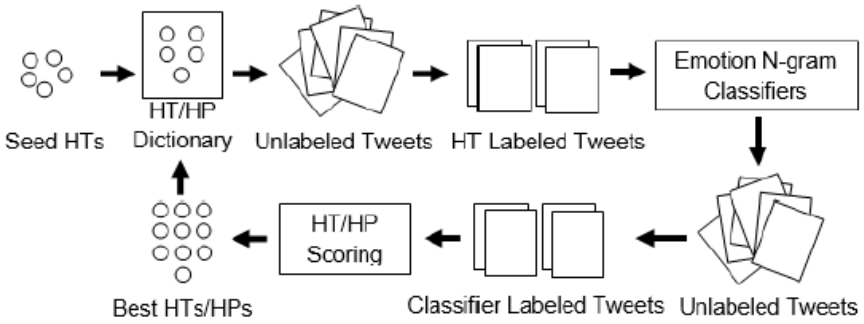
جدول 2 : توسعة WordNet-Affect

Learning –based Methods.

في هذه المنهجيات يتم الإستفادة من خوارزميات التعلم الآلي، إذ يتم العمل على بناء نموذج يستند على المدخلات "النص" وباستخدام هذا النموذج يقوم النظام بالكشف عن العاطفة. وجدنا في الفصل الثاني أنه يتم تقسيم منهجيات التعلم إلى:

- Supervised learning approaches
- Unsupervised learning approaches
- Reinforcement learning

في الطرق السابقة كانت المشكلة تكمن في تحديد العاطفة من النص المدخل، بينما المشكلة هنا تتجلى في كيفية تصنيف النص المدخل إلى مشاعر مختلفة. وعلى عكس طرق الكشف القائمة على الكلمة بشكل مجرد Keyword-based detection methods، فإن هذه الطريقة تحاول الكشف عن المشاعر من النص انطلاقاً من تدريب مصنِفات وفقا لخوارزميات التعلم الالي الشكل 19:



الشكل 19: تدفق خوارزميات التعلم الالي

إذ يتم تقسيم العواطف إلى فئات مشاعر أساسية ومن أجل كل واحده يتم التقسيم إلى خمسة وسوم Hashtags شائعة ترتبط بقوة مع العواطف الرئيسية والتي سيتم استخدامهم كـ بذور "كلمات ذات دلالات عاطفية". الجدول 3

صفوف المشاعر	Seed Hashtags، بذور الوسوم
AFFECTION	#loveyou، #sweetheart، #bff #soulmate،#romantic
ANGER	#angry، #mad، #hateyou #furious،#pissedoff
FEAR	#afraid، #petrified، #scared #worried،#anxious
JOY	#happy، #excited، #yay #thrilled،#blessed
SADNESS	#sad، #depressed #disappointed، #unhappy #foreveralone

جدول 3 الوسوم الشائعه الخاصة بكل فئة عاطفية

ثم تبدأ هذه المنهجية بعملية جمع النصوص التي تحتوي البذور، بالإضافة إلى تحديد العواطف المقابلة لهم، ثم استخدام ذلك في عملية تدريب مجموعة من المصنِفات.

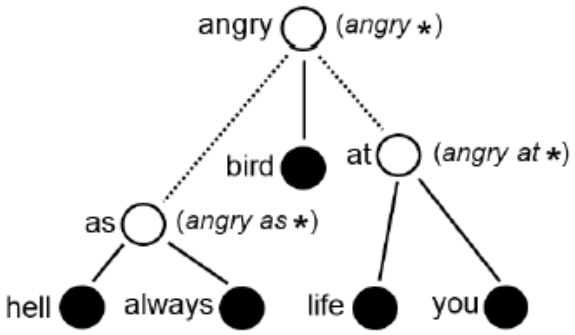
لأجل كل صف من صفوف المشاعر Emotion's class سيتم تدريب Logistic regression classifierواحد. يتميز تحليل الإنحدار الخطي بانة يساعد بإنتاج الإحتمالات مع توقعاتها مما يسمح بتوزين الوسوم ضمن النص الواحد وهذا ما يمكن المصنِف من عملية انتخاب الوسوم المناسبة للتوقع. وفي نهاية هذه المرحلة سيقوم المصنِف بعملية تصنيف الوسوم الواردة في النص ضمن صفوف المشاعر– مثال على ذلك، الجدول 4.

AFFECTION	ANGER	FEAR	JOY	SADNESS

Yourthebest#	Donttalktome#	Haunted#	Tgfad#	Foreverugly
Bestfriendforever#	Pieceofshit#	Shittingmyself#	Greatmood#	Singleprobs#
Loveyoulots#	#	Worstfear#	Thankful#	Lonerlyfe#
Flyhigh#	Fuming#	Scaresme#	Atlast#	Teamlonely#
Comehomesoon#	Hateliars#	Nightmares#	Feelinggood#	Unloved#
Wuvyou#	Heated#	Paranoid#	happygirl#	friendless#
alwaysandforever#				

جدول 4 : جدول يوضح تصنيف الوسوم في نص ضمن صفوف المشاعر العاطفية

ثم تأتي مرحلة Learning Hashtag Patterns وفي هذه المرحلة يتم إعادة الوسوم إلى كلمات متتالية ذات معنى باستخدام خوارزميات N-Gram[2] ولأجل ذلك يتم استخدام البنية Prefix Tree (Tire)لتمييز بادئات الكلمات في العبارة الواحدة – الشكل 20.



الشكل 20: Prefix Tree

بعد عملية البحث عن الوسوم وتوزينها وتصنيفها وفقا لأصناف المشاعر الخمسة، يتم احتساب متوسط احتمالات الوسوم ضمن كل مجموعة "Emotion's class" ومن ثم يتم انتخاب الوسوم الـ 10 الأكثر وزناً. وهكذا نكون قد حصلنا على أفضل الوسوم ضمن النص - الجدول 5، لتقوم الخوارزمية بعد ذلك بتكرير نفسها من جديد.

Emotion	Hashtag Pattern	Examples of Matching Hashtags
AFFECTION	#bestie*	#bestiefolyfe, #bestienight, #bestielove
	#missedyou*	#missedyoutoomuch, #missedyouguys, #missedyoubabies
ANGER & RAGE	#godie*	#godieoldman, #godieyou, #godieinahole
	#pissedoff*	#pissedofffather, #pissedoffnow, #pissedoffmood
FEAR & ANXIETY	#tooscared*	#tooscaredtogoalone, #tooscaredformama, #tooscaredtomove
	#nightmares*	#nightmaresfordays, #nightmaresforlife, #nightmarestonight
JOY	#feelinggood*	#feelinggoodnow, #feelinggoodforme, #feelinggoodabout
	#goodmood*	#goodmooditsgameday, #goodmoodmode, #goodmoodnight
SADNESS & DISAPPOINTMENT	#bummed*	#bummedout, #bummedaf, #bummednow
	#singlelife*	#singlelifeblows, #singlelifeforme, #singlelifesucks

جدول 5: جدول يوضح تصنيف الوسوم في نص ضمن صفوف المشاعر العاطفية

[2] 1-gram sequence (to be or not to be > to، be، or، not، to، be)، 2-gram sequence (to be or not to be > to be، be or، or not، not to، to be)، 3-gram sequence (to be or not to be > to be or، be or not، or not to، not to be)

في المرحلة النهائية Creating Phrase-based Classifiers يتم تطبيق خوارزميات الـ word segmentation على كل وسوم التعلم بهدف تحويلها إلى العبارات التي تكونت منها الوسوم، مثلاً: (*#lovemylive >> love my live*)

بناء على ذلك نكون قد قمنا بتدريب logistic regression classifier على كل وسم من وسوم المشاعر الأساسية في النص الواحد بالإضافة إلى البذور seeds الخاصة بكل منها.

بالنسبة لأهم تجارب الكشف عن المشاعر من النص بالإعتماد على منهجيات التعلم غير المشرف عليها Unsupervised learning method قام كل من (Bak & Kim 2012)(24) باستخدام قاعدة بيانات معيارية تسمى "ANEW" وذلك للوصول إلى ما يسمى بـ Tree-dimensional vector (تكافؤ، اثارة، هيمنة) وذلك لكل وثيقة. قاعدة البيانات المعيارية هذه تعتمد مبدأ الأبعاد العاطفية motional Dimension في عملية تمثيل الكلمات ذات الدلالات العاطفية إذ يتم الأخذ بعين الإعتبار كل من الشدة العاطفية ومدى فاعلية التعبير التي تتمتع به هذه الكلمة.
ومن جهة أخرى قام كل من (25) (Strapparava & Mihalcea) بالإعتماد على Categorical Emotion كأسلوب لتمثيل العواطف في عملية كشفها من النصوص بالإعتماد على منهجيات التعلم غير المشرف عليها.

Hybrid Methods.

بهدف البحث عن الدقة في النتائج، فقد وجدت بعض المنهجيات التي تلخص الشكل الهجين لكل من منهجيتي keyword-based methods وlearning-based methods.
تستخدم هذه المنهجيات طرق قائمة على قواعد لاستخراج كلمات دلالية من النص تتعلق بعواطف محددة بالإضافة إلى استخدام انطولوجي خاصة باللغة فيتم استخراج الصفات العاطفية منها. هذه الدلالات semantic والصفات attributes ترتبط مع العواطف Emotion على شكل "قواعد عاطفية". وهكذا فإن هذه القواعد تحل محل كلمات العاطفة الأصلية فيتم تدريب المصنِفات عليها.

3-2-3- الأبحاث المجراة في هذا المجال

- بالنسبة لاعتماد الخصائص النحوية والإسلوبية في عملية تحليل أو التقاط الشعور من النص:

نجد دراسة(12) بعنوان "Sentiment Analysis in Multiple Languages" قام فيها كل من A. Salem & H. Chen & A. Abbase باستخدام مزيج من الخصائص النحوية والإسلوبية لتصنيف الشعور "ايجابي أو سلبي" في الويب الرسمي (لغات رسمية وليست محكية) باللغتين العربية والإنكليزية، فكانت أمثلة التدريب تقارب 1300 مثال استخدمت لتدريب خوارزمية التصنيف SVM بطريقة 10 Fold Cross validation. طبعا لم يتم الأخذ بعين الإعتبار أي خصائص شعورية للكلمات واستخدموا Entropy وحققوا دقة عالية 90% نتيجة الجمع بين الخصائص النحوية والإسلوبية، قاموا باختبار دقة المصنفات باستخدام Recall(R) & F measure(F)،Precision(P) .

أما R.Obeidat & RAl-Shalabi في دراسة(13) تهدف الى كشف الفئة الاخبارية التي تنتمي اليها وثيقة انطلاقاً من النص المكتوب باللغة العربية، قاموا بجمع ما يقارب 1445 خبر من مواقع اعلامية ناطقة باللغة العربية الرسمية مثل الجزيرة والنهار، ثم قاموا بتدريب مصنفات KNN "KNN K-Nearest Neighbor" مرة بالإعتماد على N-grams ومرة اخرى بالاعتماد على الطرق التقليدية لفهرسة الكلمات ضمن الأصناف التي تنتمي اليها، وخلصت هذه الدراسة الى ان استخدام N-grams تعطي نتائج افضل 6-7% مما هي عليه عند استخدام طرق فهرسة الكلمات ضمن الاصناف، انتهت الدراسة الى الاشارة الى ان دقة النظام قاربت 64.5% اذ تم اختبار دقة المصنفات باستخدام Recall(R) & F measure(F) (10)،Precision(P).

بالنسبة لاستخراج المعلومات في اللغة العربية فإن Zreik & Hajjar قدما(14) مخططات مختلفة للتحليل الصرفي للكلمات العربية، وقارنوا نهج فهرسة الكلمات بالشكل المجذر مع نهج N-gram واقترحوا منهج هجين يستخدم كل منهما.

- أما بالنسبة لاستخراج العاطفة من النص وفقا للخصائص النحوية والأسلوبية:

عند الحديث عن الإعتماد على الخصائص الأسلوبية في تحديد العاطفة، فلا بد من التطرق الى أساليب طرق توزين الكلمات لما لها من أهمية في دقة تحديد العاطفة والتي تنوعت نتائجها مع تنوع أساليب التوزين، فمثلا قام كل من Taner Danisman & Adil Alpkocak (15) و كل من Hyo Jin Do & Ho-Jin Choi (16) بتقديم دراستين تختصان بكشف العواطف من النص، تم تمثيل بيانات التدريب ضمن فضاء شعاعي vector space model بالإعتماد على قواميس يتم بناءها من بيانات التدريب بشكل مباشر. في الدراسة الأولى (15) تم تدريب مجموعة من المصنفات "SVM"،"LibSVM & Naïve Bayes" باستخدام تقنية 10-Folds cross validation وذلك على مدونة مكونة من 801 مثال مكتوب اللغة الانكليزية الرسمية وذلك بهدف بناء نموذج للتصنيف الى واحد من خمسة سمات للمشاعر "غضب، خوف، فرح، حزن وقرف".
أما في الدراسة الثانية (16) فقد تم تدريب المصنف SVM على مدونة تدريب مؤلفة من 4447 تغريده مكتوبة باللغة الكورية، تم جمعها من موقع التواصل الاجتماعي تويتر، عملية التدريب تمت باستخدام النهج 10-Folds cross validation بهدف الوصول لنموذج تصنيف الى واحد من ستة عواطف "غضب، خوف، فرح، حزن، تفاجئ وقرف". والذي حقق دقة f- 70% measure.
في كل دراسة تم استخدام طريقة لتوزين كلمات بيانات التدريب بطريقة تختلف عن الأخرى. ففي (15) كانت عملية احتساب الأوزان تتم باستخدام معادلة رياضية TF-IDF فكانت تأخذ بعين الاعتبار تردد كل كلمة ضمن الوثيقة وترددها ضمن كل صنف، بالإضافة الى عدد الوثائق ضمن كل صنف عاطفي. اما معادلة احتساب الاوزان في (16) كانت تشابه سابقتها باستثناء انها لم تأخذ تردد الكلمة ضمن كل صنف بعين الإعتبار.

في عمل أخر بعنوان (33) Emotion Recognition from text using knowledge-based ANN، قام الباحثون بالبحث عن الصنف العاطفي الذي تشير إليه جمل مكتوبة بالإنكليزية الرسمية، قاموا ببناء مدونة تدريب بحجم 3200 جملة وتعاملوا مع ثمانية أصناف عاطفية (الغضب، الخوف، التمني، الحزن، السعادة، الحب، الشكر، والصنف المحايد). واستخدموا في ذلك منهجية Keyword-Based بالإضافة الى منهجيات التعلم الآلي (خوارزمية Knowledge Based Artificial Neural Network KBANN) وحققوا دقة تصنيف 58.8%.

وكذلك الأمر، قام باحثون ضمن دراسة بعنوان (34) Using YouTube comments for text-based emotion recognition باستخدام منهجية Keyword-Based وذلك لتصنيف الدلالة العاطفية لتعليقات مستخدمي اليوتيوب باللغة الإنكليزية الرسمية فتعاملوا مع خوارزميات التعلم الآلي غير المشرف عليها، وتعاملوا مع ستة عواطف رئيسية (2)، تم العمل في هذه الدراسة على تمثيل كل فئة عاطفية من قبل قائمة من الكلمات التعبيرية فبدأوا أولا باستخراج الأسماء والأفعال والصفات من نصوص مدونة التدريب واستبعدوا الضمائر وحروف الجر والعطف، ثم احتساب احتمالية انتماء كل كلمة إلى كل صنف عاطفي، وهكذا افترضوا أن احتمالية الجملة بأكملها ضمن كل صنف هي متوسط الاحتمالات التي تم الحصول عليها من كلمات هذه الجملة. بعد اخبار الدقة على تعليقات تختلف عن مدونة التدريب حقق دقة f-measure 68.82%.

أما في دراسة بعنوان (35) Emotion Recognition from Text Based on Automatically Generated Rules، قام الباحثون ببناء مدونة تدريب بحجم 18000 تغريده تم جمعهم من موقع التواصل الإجتماعي "تويتر" وقاموا بتصنيفهم الى ستة عواطف وفقا لإيكمان (2). وحققوا دقة تصنيف 80% باستخدام k-nearest neighbors (KNN) وذلك عند اختبار التصنيف على أمثلة تختلف عن أمثلة التدريب. في هذه الدراسة تم استخدام أنطولوجي مثل WordNet and ConceptNet وذلك بهدف التقاط الكلمات العاطفية "البذور" التي يحتويها النص ومن ثم العمل على توليد قاعدة التحليل العاطفية emotion recognition rule (ERR) الموافقة للجملة المدروسة وذلك من خلال تحليل الجملة إلى مكوناتها باستخدام (36) Stanford pos tagger ثم وبعد استبعاد الكلمات التي لا تحمل دلالة عاطفية يتم البحث عن كيفية الترابط بين كلمات هذه الجمل باستخدام (37) Stanford dependency parser . بعد ذلك تبدأ عملية تدريب المصنفات.

- اما بالنسبة لمهام اكتشاف العاطفة وفقا لأنواع الصفوف المتوقعة:

نجد ان كل من (17) A. Rapport & O. Tsur & D. Davidiv في دراسة بعنوان Enhanced Sentiment Learning Using Twitter Hashtags and Smileys قد استخدموا KNN K-Nearest Neighbor خوارزمية الجار الأقرب لتصنيف عاطفة التغريدات باللغة الإنكليزية واستخدموا smileys & hash-tags "50 علامة تويتر و15 وجه تعبيري" بشكل

رئيسي كخصائص ضمن المصنفات. تم التدريب على مدونة بحجم 3852 تغريده مكتوبة باللغة الانكليزية الرسمية بهدف التقاط الدلالة الشعورية للتغريدات "حزن، سعادة، غضب وملل"
من ناحية أخرى، J. Feng & L. Barbosa (18) استخدموا المصنف SVM لتصنيف المشاعر "ايجابية سلبية او محايدة" في التغريدات. واعتمدوا بشكل رئيسي على استخراج الكلمات المميزة للشعور بشكل رئيسي من بيانات التدريب وبناء معجم مفرداتي يحتوي هذه الكلمات ودرجة كل منها ضمن الشعور. تم مقارنة اداء ثلاث نماذج تم تدريبها باستخدام ثلاث مدونات تحتوي تغريدات "اقل من " حجم تغريدات هذا المدونات الثلاث كانت 24508 للمدونة الأولى و7969 للثانية و1312 للثالثة. ودقة التصنيف كانت 77% و82 %و89% على الترتيب.

عمل اخر يستحق الذكر هو المشروع الذي قام به Elfeky & Elhawary (19) في العام 2016 بعنوان "Mining Arabic Business Reviews" استخدموا فيه ويب الأعمال الرسمي ليكون مصدر البيانات المدخلة واستخرجوا نشرات اعمال مكتوبة باللغة العربية، هذه النشرات تم تحليلها ودراسة المشاعر فيها. لا يوجد تقييم تم اجراؤه من قبل هذا العمل.

في هذا البحث:
قمنا باكتشاف العاطفة من النصوص العربية المكتوبة باللهجة المحكية السورية، مع الأخذ بعين الإعتبار "ستة" عواطف رئيسية (حزن، فرح، تفاجئ، اشمئزاز، خوف، غضب). حاولنا الإستفادة من البنية النحوية للكلمة مثل محاولة العمل على الكلمة مرةً بشكلها الكامل full form ومرةً اخرى بشكلها المجذع ISRI-Stem form وتهجين أفضل الشكلين مع النهج N-grams في عملية تمثيل بيانات التدريب. بالإضافة إلى ذلك عملنا على الإستفادة من الخصائص الإسلوبية للكلمات ضمن بيانات التدريب حيث يتم أخذ الرموز وأوزان الكلمات بعين الاعتبار. بالإضافة الى ما سبق، عملنا على البحث عن أفضل نموذج رياضي لالتقاط أوزان الكلمات ودلالاتها ضمن الأصناف العاطفية بهدف تمثيل كل من هذه البيانات "التغريدات" كـ vector space (15) model ذي بنية رقمية تشير بمدلولها إلى نفس الصنف العاطفي الذي كان يشار اليه ضمن النموذج النصي لبيانات التدريب "التغريدات".

الفصل الرابع - النظام المقترح

4-1- الصعوبات وتحديات البحث

يوجد الكثير من دراسات تحليل المشاعر من النصوص Emotional Analysis From text وقد كانت غالبيتها تعنى باللغة الإنكليزية، حيث تتوفر الأدوات اللازمة لتحليل اللغة إضافة إلى الأبحاث الكثيرة في هذا المجال والتي تسهل على الباحثين الجدد عملهم. بالمقابل نعاني من قلة الدراسات التي قامت على تحليل الشعور من نصوص مكتوبة باستخدام اللغة العربية.
أن التحدي الأكبر لتحليل الشعور من نصوص مكتوبة باللغة العربية هو التعقيد الكبير لهذه اللغة سواء من حيث:

- البنية الصرفية حيث أن قواعد اللغة العربية معقدة للغاية.
- أومن حيث اختلاف الجمل وفقا للبنية القواعدية، إذ لدينا الجمل الفعلية التي تبدأ بفعل، والجمل الإسمية التي تبدأ باسم ولدينا العديد من أجزاء الكلام المختلفة، كما أن كل كلمة يمكن أن تمتلك اشتقاقات مختلفة، إضافةً إلى تأثير تشكيل الكلمات على معانيها.

إضافةً لتحديات أخرى تتعلق باستخدام جذر الكلمة، حيث أن الجذر الثلاثي للكلمة قد يعطي كلمات مختلفة ذات معاني مختلفة. ونفس الجذر بعد إضافة بادئات ولاحقات مختلفة بإمكانه أن ينتج كلمات جديدة ومختلفة. نضيف إلى ذلك تحديات استخدام التشكيل والحركات والذي يختلف حسب مكان الكلمة من الجملة وإعرابها.
نعاني أيضا من مشكلة قلة الأدوات المتوفرة لتحليل الشعور وكذلك قلة المعاجم والقواميس الدلالية للغة العربية والحاجة الماسة لمحللات صرفية تقوم باستخراج جذور الكلمات وإزالة البادئات واللاحقات من الكلمات، والحاجة لمحللات قواعدية تستطيع تحديد أجزاء الكلام كالفعل والفاعل والمفعول به... الخ
تم تطوير مجموعة من المحللات الصرفية للغة العربية ومحللات أجزاء الكلام وتسمى Grammatical analyzers or part of speech taggers (POS) مثل:

- Buckwalter Arabic Morphological analyzer (BAMA) (26)
- Morphological Analysis and Disambiguation for Arabic (MADA) (27)

ولكن وعلى الرغم من ذلك لا يوجد محددات أقسام الكلام POS taggers معقدة تستطيع تحديد جميع أجزاء الكلام والتمييز بين الأنواع المختلفة للجمل العربية.
كل هذه المشكلات تشكل تحدي للتنقيب عن الشعور أو العاطفة في اللغة العربية والتي تتطلب تحليلاً دلالياً للكلمات إضافة لتحليلاً قواعدياً للنص.

4-2- تحليل المشاعر من النص

كما ذكرنا سابقا، يوجد بشكل رئيسي (20) نهجين لتصنيف العاطفة من النص
Emotion Orientation (EO) & Emotion Analysis with Machine Learning (EAML)

EAML Approach

هو نهج مشرف عليه Supervised approach وفيه تكون مجموعة البيانات مصنفة يدوياً وفقا للصنف الذي تنتمي اليه. وممثلة بشعاع من السمات features. تستخدم هذه الأشعة من قبل خوارزميات تصنيف كبيانات تدريب حيث أن مجموعة معينة من السمات تؤدي لظهور صنف معين كخرج للمصنِف وذلك بتطبيق أحد خوارزميات التصنيف من النوع Supervised مثل:

- Naïve Bayesian classifier
- Maximum support vector machine SVM
- Entropy

EO Approach

نهج غير مشرف عليه Unsupervised approach يُستخدم فيه معاجم العاطفة Emotion lexicon وفيه توضع كل كلمة يقابلها semantic intensity كرقم يدل على الصف الخاص بها، ثم يتم استخدام هذا المعجم لاستخراج الكلمات ذات الدلالة العاطفية من الجملة وجمع مقدار كل منها بهدف تحديد العاطفة التي تنتمي اليه الجملة.

حاولنا في هذا البحث العمل على تهجين المنهجين السابقية من خلال استخدام EAML ولكن بصيغة محسنة وذلك من خلال الأخذ بعين الإعتبار أهمية كلمات امثلة التدريب، إذ سنقوم باحتساب وزن كل كلمة بالنسبة للدلالة العاطفية للجملة التي تنتمي إليها الكلمة "وذلك على مستوى كامل أمثلة التدريب"، ومن ثم أخذ أوزان الكلمات ضمن عملية بناء أشعة السمات وبالتالي ضمن عملية التدريب وبناء نموذج التصنيف.

3-4- خطة العمل

كما ذكرنا سابقا، تهدف خطة العمل إلى تقديم دراسة لتحليل العاطفة من النصوص العربية المكتوبة باللهجة المحكية السورية والتي أجريت على تعليقات المستخدمين ضمن أحد مواقع التواصل الاجتماعي "تويتر". تم الأخذ بعين الاعتبار "ستة" عواطف رئيسية (حزن، فرح، تفاجئ، اشمئزاز، خوف، غضب).
استخدمنا في هذه الدراسة ما يزيد عن 1320 تغريده تم جمعها بشكل آلي من موقع التواصل الاجتماعي تويتر، وتم تصنيفها يدويا وذلك لبناء نموذج تصنيف للمشاعر "في مراحل متقدمة قمنا باختبار هذه النماذج بأسلوب Precision and Recall and F-measure (10)" حاولنا الإستفادة من البنية النحوية للكلمة مثل محاولة العمل على الكلمة مرةً بشكلها الكامل full form ومرةً أخرى بشكلها المجذع Stem form وتهجين أفضل الشكلين مع النهج N-grams في عملية تمثيل بيانات التدريب. بالإضافة إلى ذلك عملنا على الإستفادة من الخصائص الإسلوبية للكلمات ضمن بيانات التدريب بيانات حيث يتم أخذ الرموز وأوزان الكلمات بعين الاعتبار والبحث عن أفضل نموذج رياضي لالتقاط أوزان الكلمات ودلالاتها ضمن الأصناف العاطفية بهدف تمثيل كل من هذه البيانات "التغريدات" كـ Vector Space Model ذي بنية رقمية تشير بمدلولها إلى نفس الصنف العاطفي الذي كان يشار إليه ضمن النموذج النصي لبيانات التدريب "التغريدات".

نلخص العملية كالتالي:

- بناء المدونة المنمّطة بالمشاعر.
- بناء القواميس.
- المعالجة النصية للتغريدات.
- توليد أشعة واصفات البيانات.
- تدريب المصنفات.

4-3-a- بناء المدونة المنمّطة بالمشاعر

1- جمع مجموعة بيانات التدريب

في جميع الأبحاث التي تعتمد على التنقيب في البيانات وتعلم الآلة يتم الحصول على مجموعات البيانات من مراكز الأبحاث على الإنترنت ومن مستودعات الكترونية مخصصة لهذا الغرض. أحد أبرز التحديات التي واجهتنا في هذا الصدد، عدم وجود مجموعات بيانات مناسبة باللغة العربية تخدم موضوع بحثنا، علما أننا وجدنا أكثر من مجموعة بيانات باللغة العربية لكنها كانت تستخدم لهجات تتبع إلى لهجة بلد الباحث، خصوصا أن كل البيانات المجمّعة كانت لأغراض تحليل الشعور "Sentiment analysis" اي التقاط الدلالة الشعورية للنص "إيجابي أو سلبي " وليست لأغراض تحديد العاطفة، كانت هذه البيانات مولدة من قبل المستخدمين على مواقع التواصل الاجتماعي، والمتوفر منها باللغة العربية يفتقر للجودة ومكتوب بلهجات سعودية(31) أو أردنية(32) . لذلك قمنا بالعمل على بناء تطبيق ويب Web Application يستخدم المكتبة LINQ To TwitterCP.DLL بهدف الإقتران مع حساب على موقع التواصل الإجتماعي "تويتر" باستخدام Twitter Search Application Programming Interface(API)[3] ومن ثم جلب التغريدات.

2- تصنيف يدوي للعينات "تسمية العينات باسم الصنف المناسب"

نظرا لشعبية وسائل الاعلام الاجتماعية، استخدمنا الجمل التي تم جمعها من تويتر باستخدام عملية جمع الي للتغريدات "كما ذكرنا آنفاً". إن ناتج عملية الجمع الآلي للتغريدات كانت كمية كبيرة من البيانات في شتى المواضيع والكثير منها لا يحتوي على مواقف ذات دلالات عاطفية، مما حملنا عبئاً إضافياً للفحص اليدوي لكل حالة واختيار الحالات التي تحتوي على دلالات شعورية فقط، بعد عملية الفحص والفلترة اليدوية للتأكد من احتواء التغريدات على دلالات شعورية، قمنا بفرز هذه التغريدات وتنظيمها ضمن مجموعات تتناسب مع الصفوف العاطفية المدروسة.

العينات المجموعة تشمل 1320 عينة -تغريده، كل منها يشير ضمنياً إلى إحدى الأصناف العاطفية التالية: حزن، فرح، تفاجئ، اشمئزاز، خوف، غضب. مجموعات العينات كانت وفقا للأعداد التالية – جدول 8:

اسم المجموعة "الصنف العاطفي "	رمز المجموعة	عدد العينات "التغريدات"
الحزن	sad	220 تغريدة
الفرح	joy	220 تغريده
التفاجؤ	sur	220 تغريده
الإشمئزاز	loa	220 تغريده
الخوف	fea	220 تغريده
الغضب	ang	220 تغريده

جدول 6 بيانات التدريب

ملاحظة **1**: في مرحلة التطوير يمكن العمل أيضاً على صنفين عاطفيين إضافيين هما الـ "ثقة والتوقع" وهنا يجدر الإشارة إلى أن مدونة التدريب التي قمنا بجمعها تحتوي طيف واسع من العينات "التغريدات" المنمّطة بهذين الصنفين.

ملاحظة **2**: يمكن الإشارة إلى ان بعض الأصناف العاطفية مثل الـ "حب" لا يمكن اعتبارها أصناف رئيسية، إذ يمكن الإشارة عليها من حاصل اجتماع صنفين عاطفيين رئيسيين مثل الـ "فرح" والـ "ثقة".

4-3-b-بناء القواميس

من الجدير بالذكر أن هذه القواميس تعتبر أول مجموعة قواميس في مجال اللغة العربية باللهجة السورية حيث تفتقر جميع المصادر العربية على الانترنت – حتى تاريخ إعداد هذه الدراسة – لأي قاموس عربي لكلمات تحليل المشاعر باللهجة السورية، وهذا ما يمكن ملاحظته في جميع الأوراق البحثية العربية التي تعرض أساليب التحليل وتذكر أن أفضل الطرق هو استخدام القواميس ثم تتأسف لعدم وجود مثل هذه القواميس في اللغة العربية مما يدفع الباحث إلى اللجوء إلى أساليب أخرى.

[3] https://dev.twitter.com/

تم تجميع هذه القواميس طيلة فترة العمل على الأطروحة، أُخِذَت كلمات هذه القواميس من البيانات النصية التي تم جمعها "بيانات التدريب – المدونة المنمّطة بالشاعر"، بالإضافة إلى الكلمات الشائع استخدامها بين مستخدمي شبكات التواصل الاجتماعية.

1- قاموس تكرار كلمات المدونة المنمّطة بالمشاعر.

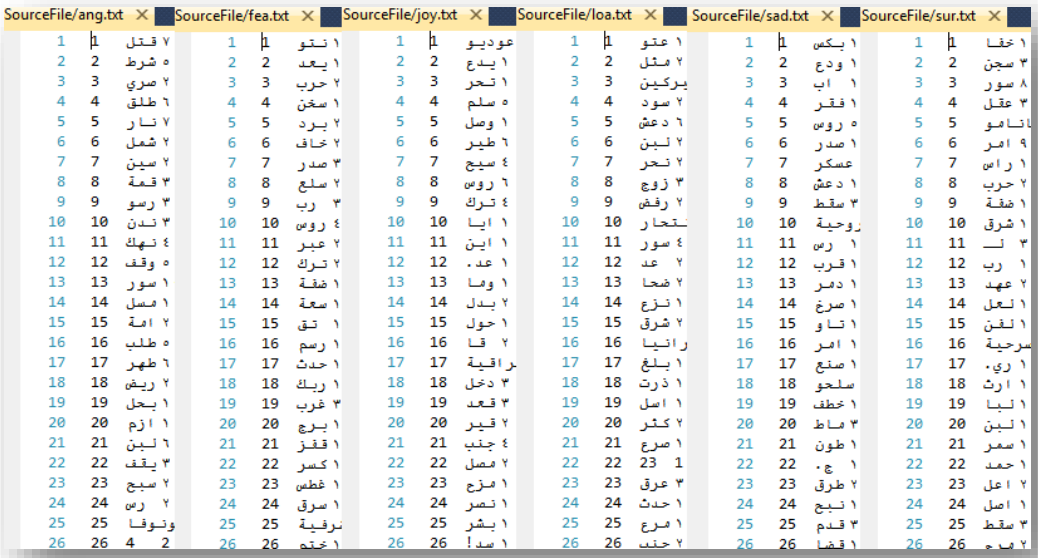
يتضمن هذا القاموس كل الكلمات الواردة في المدونة المنمّطة بالمشاعر مع تكرار كل منها ضمن المدونة. الشكل 21 يبين جزء من هذا القاموس.

الشكل 21: قاموس تكرار الكلمات

2- قاموس اوزان الكلمات وفقا لما تشير اليها من دلالة عاطفية COR Emotional lexicons:

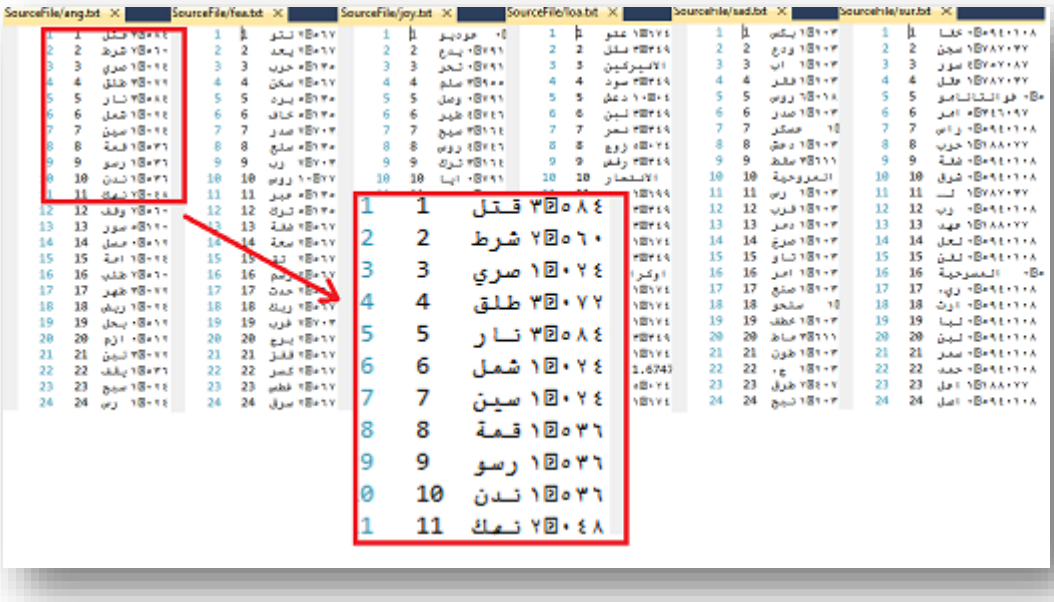
وهي مجموعة من القواميس "قاموس لكل صنف عاطفي"، كل من هذه القواميس تتضمن مجموعة الكلمات الواردة ضمن جمل تشير إلى العاطفة الموافقة للقاموس مع وزن كل من هذه الكلمات ضمن الصنف العاطفي "نستعرض طريقة احتساب هذا الوزن في الفقرة 4-3-d". تم استخلاص هذه القواميس من المدونة المنمّطة بالمشاعر. الشكل 22 يبين جزء من بعض هذه القواميس.

يتوفر عالمياً مجموعة من القواميس الجاهزة الخاصة باللغة العربية مثل قاموس "سيف (28)" ولكنها لا تستوعب كل كلمات اللهجات العامية العربية بما فيها السورية، "وهي موضوع بحثنا".



الشكل 22: قاموس اوزان الكلمات وفقا لما تشير اليه من دلالة عاطفية

القواميس بعد احتساب أوزان الكلمات "سنشرح طريقة ذلك لاحقاً" ستكون كالتالي – الشكل 23:



الشكل 23: قاموس تردد الكلمات وفقا لما تشير اليه من دلالة عاطفية

3- قاموس الكلمات المشددة على المعنى Emphasizing words dictionary:

الكلمات المشددة على المعنى هي مجموعة الكلمات التي تزيد من التوكيد الدلالي للكلمة التي تليها أو التي تسبقها في سياقات متعددة، نذكر منها على سبيل المثال لا الحصر: *كثير، جداً، أكيد، خصوصا، أبداً، نهائياً*.

هذه الكلمات تم إفراد قاموس خاص بها (حوالي 30 كلمة) ويتم تحديثه باستمرار مع تقدم العمل.

4- قاموس الكلمات النافية للمعنى Negation words dictionary:

يقسم أسلوب النفي في اللغة العربية إلى قسمين، وهما:

- التفي الصريح أو الظاهر.
- النفي الضمني أو غير الظاهر.

والنفي الصريح هو النفي الذي تستخدم فيه احدى أدوات النفي، وأشهر هذه الأدوات: (لا) و (ما) و (لم) و (لمَّا) و (لن) و (ليس) و (لات) و (إنْ) و (غير) و (لا النافية للجنس) و (لام الجحود).

أما النفي الضمني فهو النفي الذي يكون بغير أدوات النفي التي تحمل معنى النفي الصريح وإنما يكون بأدوات أخرى شرطية أو استفهامية تتضمن معنى النفي، ومن تلك الأدوات: (لمَّا) و (لو) و (لولا).

قمنا ببناء قاموس يحتوي قائمة بالأدوات النافية للمعنى (حوالي 70 كلمة).

5- قاموس كلمات الشتم

هذه الكلمات تم إفراد قاموس خاص بها (حوالي 100 كلمة) وتم تحديثه باستمرار مع تقدم العمل.

6- قاموس الكلمات المترافقة

وهي مجموعة الكلمات التي جرت العادة أن تأتي مع بعضها ليتم الإشارة من خلالها إلى معنى واحد مثل "صباح الخير، ياسلام، أستغفر الله، ...".

7- قاموس الوجوه التعبيرية Emoticons Emotion Dictionaries:

من مميزات النصوص المتوافرة على شبكات التواصل الاجتماعي هي وجود الوجوه التعبيرية فيها، حيث درج المستخدمون على التعبير عن مشاعرهم بوضع حروف واشكال ترسم معاني الابتسامات، كأن نكتب "(:" لنعبر عن ابتسامة سعيدة من الشكل ☺ أو")":" نعبر عن حزن أو"D:" لنعبر عن ابتسامة عريضة.

ما ضاعف من استخدام الوجوه التعبيرية على الشبكات الاجتماعية هو إتاحة هذه الشبكات للكثير من رموز الوجوه الجديدة بشكل دائم حتى وصلنا إلى الحد الذي باتت فيه بعضاً من الوجوه المرسلة من أحد التطبيقات على منصة ما مثل هواتف Android غير مقروءة بالنسبة لمستخدم الكمبيوتر الشخصي فتظهر له بشكل مربعات مفرغة.

هذه الوجوه التعبيرية توفر علينا الكثير من العناء في مهمتنا المنشودة للتعرف على مشاعر المستخدمين. لذلك فقد أعطيناها اهتماماً كبيراً من خلال إعداد قاموس خاص وتزويده بالوجوه التعبيرية المتعارف عليها في جميع المنصات لكي نتجنب المشكلة سابقة الذكر، قمنا في هذا القاموس بعملية إسناد قيمة لكل وجه بما يتناسب مع نوع وشدة الشعور المعبّر عنه وذلك بعد دراستنا للدلالات العاطفية لهذه الرموز وتأثير كل منها على كل صنف عاطفي. الشكل 24 يعرض عينة من هذا القاموس (حجم القاموس الإجمالي حوالي 100 رمز):

	حزن	فرح	تفاجؤ	اشمئزاز	خوف	غضب	ثقة	توقع
:-(	1	0	0	0	0	0	0	0
:(	1	0	0	0	0	0	0	0
):	1	0	0	0	0	0	0	0
>:O	0	0	0.25	0	0	0.75	0	0
:O	0	0	1	0	0	0	0	0
>:(	0.5	0	0	0	0	1	0	0
:/	0	0	0	1	0	0	0	0
:'(	1	0	0	0	0	0	0	0
3:)	0	1	0	0	0	0	0	0
O:)	0	1	0	0	0	0	0	0
-_-	0.25	0.75	0	0	0	0	0	0
o.O	0	0	0.5	0	0.5	0	0	0
:-)	0	1	0	0	0	0	0	0
:)	0	1	0	0	0	0	0	0
;-)	0	0.75	0	0	0	0	0.25	0
:P	0	1	0	0	0	0	0	0
:D	0	0.75	0	0	0	0	0.25	0
=D	0	0.75	0	0	0	0	0.25	0
(:	0	1	0	0	0	0	0	0
:V	0	1	0	0	0	0	0	0
^_^	0	1	0	0	0	0	0	0
8-)	0	1	0	0	0	0	0	0
<3	0	0.5	0	0	0	0	0.5	0
👍	0	0	0	0	0	0	1	0
🙄	0	0.5	0	0	0	0	0.5	0
😂	0	1	0	0	0	0	0	0
😜	0	0	1	0	0	0	0	0
😊	0	1	0	0	0	0	0	0

الشكل 24: قاموس الوجوه التعبيرية

بعض الملاحظات المهمة:

- سنقوم بمحاولة الإستعانة بقاموس سيف-NRC-Emotion-Lexicon-v0.92-InManyLanguages-web (28) إلى جانب القواميس السابقة.
- يلجأ بعض الباحثين ممن يتعاملون مع اللهجات المحكية بتحويل الكلمات من هذه اللهجات إلى اللغة الفصحى لسهولة التعامل معها! أنا لم أستخدم هذا الأسلوب لقناعتي بعدم جدوى هذه الطريقة خصوصا وأن بناء قاموس للتحويل من اللهجة المحكية إلى اللغة الفصحى يتطلب جهدا أكبر بكثير من بناء قواميس أصلية تتضمن مزيجا من الكلمات الفصحى والمحكية كما في حالتنا.

4-3-c-المعالجة النصية للتغريدات

يمكن أن نسرد عملية المعالجة المسبقة للبيانات التي قمنا بها في مجموعة من الخطوات:

- التقطيع إلى كلمات Tokenization
- تقييس الكلمات Normalization
  - إزالة حركات التشكيل.
  - إزالة الأحرف المكررة.
  - إزالة الروابط والاشارات والهاشتاغ.
  - تقييس الألف.
- التعامل مرة مع ازالة كلمات التوقف Remove Stop Words ومرةً اخرى الإحتفاظ بها وأخذها بعين الإعتبار
- تجذيع الكلمات Stemming
- التعامل مع أدوات النفي والتأكيد على المعنى ورموز وجوه التعبير.
- ثم تأتي مرحلة توليد أشعة واصفات البيانات بهدف نمذجة بيانات التدريب ضمن فضاء شعاعي فقرة "توليد أشعة واصفات البيانات Feature Vector Generation".

الخطوة الأولى: التقطيع إلى كلمات Tokenization:

نقوم هذه الخطوة بتقسيم النص في التغريدات إلى كلمات منفصلة ليتم التعامل مع كل كلمة بشكل مفرد. وهي خطوة ليست بالبساطة التي تبدو عليه، فهناك الكثير من الحالات التي لا يمكن فصل الكلمات فيها اعتماداً على الفراغات، مثل حالات الترقيم المختلفة أو الأرقام الملحقة بالكلمات بشكل مباشر. للقيام بهذه الخطوة بشكل متقن ومن خلال متابعة العمل ومراجعة التغريدات وجدنا أنه من المناسب القيام بهذه العملية برمجياً بشكل يراعي عدم تفضيل فصل بعض الكلمات عن الأخرى مثل "صباح الخير"، "يا سلام" أو "ما بدي" ... ولتحقيق ذلك كان لابد من بناء قاموس يحتوي هذه الكلمات "قاوس الكلمات المترافقة".

الخطوة الثانية: تقييس الكلمات Normalization:

في عالم الويب الاجتماعي والمحتوى الذي يكتبه المستخدمون، من المتوقع أن نجد مختلف أشكال النصوص فهناك الغالبية العظمى التي تفضل أن تكتب باللهجة المحكية، وهناك من قرأ خاطرة أو قصيدة نثرية ويرغب ان يشاركها مع أصدقائه والبعض قد يرغب في أن يستخدم العربية الفصحى للدلالة على تمسكه بلغتنا الجميلة، ويزيد عليه آخر بأن يضع بعض حركات التشكيل ليقصد معاني محدده، وهنا يكمن التحدي الكبير أمام أي مجهود لمحاولة فهم وتحليل هذا الكم الجارف من المعلومات الذي يتجدد في كل لحظة.

ومما يضاعف المشكلة هو عدم اتفاق مستخدمي الشبكة الإجتماعية على أساليب كتابة موحدة حتى في استخدامهم للهجة المحكية. فالبعض يضع الألف المهموزة أينما اتفق، والبعض الآخر لا يضع الهمزة اطلاقاً والبعض يكتب التاء المربوطة هاءً والبعض يعكس ذلك، هذا إذا لم نتحدث عن تكرار بعض الأحرف لإكساب الكلام معنى التأكيد كأن نقول: "أكيييييد" بدلاً من "أكيد".

أمام هذه الحقيقة قمنا بتطوير برمجية يتم من خلالها إزالة المحارف غير المرغوبة والمقاطع غير العربية من التغريدات وتعديل أشكال الكلمات لتصبح أكثر ملائمة.

تعتبر هذه المرحلة ذات أهمية بالغة من ناحية تأثيرها على دقة نتائج عملية التصنيف النهائية لذلك قمنا بتركيز الكثير من الجهد لإجراء عملية تقييس فعالة قمت من خلالها بتطوير الإجراءات التالية:

- معالجة قواعد الإملاء
- إزالة حركات التشكيل.
- إزالة الأحرف المكررة.
- إزالة الروابط والاشارات والهاشتاغ.
- تقييس الألف.

- **معالجة قواعد الاملاء**

عند التعامل مع النص الرسمي مثل المقالات الصحفية، من المستحسن معالجة الأخطاء الإملائية وحالات اختلاف أشكال الأحرف "تسوية الحرف" والتي تتعلق بأربع أحرف واشكالهم المختلفة وهي:

- أشكال مختلفة للحرف ألف وهي: أ – إ – ا
الخطأ في كتابة الأف قد يغير من زمن الفعل مثلاً "ادرس – فعل أمر" و "أدرس – فعل مضارع"
- حرف الياء "ي" والالف المقصورة "ى"
الأف المقصورة لا تأتي إلا في نهاية الكلمة وعادةً تتحول إلى ياء عند إضافة لواحق للكلمة مثلاً:
الكلمة "على" عند إضافة حرف الهاء كلاحقة تصبح "عليه".
وأحياناً تتحول الألف المقصورة إلى ألف، مثلا الفعل "يرى" عند إضافة اللاحقة "حرف الهاء" يصبح يراه.
- حرفي التاء المربوطة والهاء
التاء المربوطة تكتب فقط في نهاية الكلمة وتتحول إلى تاء مبسوطة عند إضافة اللواحق مثل "لعبة، لعبته"
- أشكال الهمزة المختلفة
قد تتحول الهمزة على السطر إلى همزة على واو أو على نبرة عند إضافة بعض اللواحق "سماء، سماؤه، سمائه"

- **إزالة حركات التشكيل Diacritics Removal.**

ما يميز اللغة العربية عن باقي اللغات هو استخدامها لحركات التشكيل. تفيد هذه الحركات في إكساب بعض الكلمات معانٍ مختلفة باختلاف الحركات على نفس مجموعة الأحرف. مثال: "*كَتَبَ*" هي فعل ماضٍ يدل على الإنتهاء من فعل الكتابة. أما "*كُتب*" فهي كلمة تدل على مجموعة الكتب.
النقطة الإيجابية في سياق هذا البحث هي أنه يمكننا إزالة جميع الحركات دون أن يؤثر ذلك على العاطفة الضمنية للكلمات ذات المعاني العاطفية وهي الكلمات التي تهمنا في الأساس.

- **إزالة الأحرف المكررة Repetition Removal.**

من العادات الشائعة لدى مستخدمي شبكات التواصل الاجتماعي تكرار أحد حروف الكلمة للدلالة على توكيد المعنى، حيث يتناسب مقدار التوكيد مع زيادة التكرار. على سبيل المثال: كلمة "*كتير*" التي تقابل باللغة الفصحى كلمة "*كثير*" يمكن أن تشدد لتصبح "*كتيير*" ويمكن ان تشدد لتصبح "*كتيييير*".

لا يمكن التعرف على الكلمات عندما تكون بهذا الشكل، لذلك يلزم حذف التكرار ويجب الأخذ بعين الإعتبار أنه لا يمكن التنبؤ بعدد مرات التكرار في الخوارزمية، وأحياناً يكون تكرار الحرف لمرتين من أصل الكلمة مثل كلمة "عدد". لحل هذه المشكلة قمنا بحذف أي تكرار لحرف تكرر وجوده لمرتين أو لأكثر، متغاضين بذلك عن الملاحظة الأخيرة وذلك بعد مراجعة الأمثلة في البيانات المتوفرة لدينا والتأكد من أن النسبة ضئيلة جدا لاحتمال ورود كلمات بتكرار حرفين من أصل الكلمة وتحمل شعور عاطفي ضمني في نفس الوقت، وفي حال وجدت يمكن التغاضي عن ذلك لصالح الكم الكبير من الكلمات التي يمكن التعرف عليها عند تطبيق هذا الأسلوب.

- **تقييس الألف**

الألف الممدودة يمكن أن تأخذ عدة أشكل ضمن المجموعة التالية (ا، أ، إ، آ) وبما أن ورود أحد هذه الأشكال الأربعة في نفس الكلمة سيفضي إلى أربع كلمات مختلفة (من وجهة نظر الحاسب) وكلمة واحدة من حيث المعنى. لذلك فان الحل هو بتقييس الأشكال المختلفة للألف لتصبح ألفا (ا) بدون همزة أو مَدة ثم تقييس كل الكلمات وفقا لذلك

- **إزالة الروابط والاشارات والهاشتاغ URL، Mention and Hashtag removal .**

من الشائع جدا وجود روابط لأخبار أو صور يشاركها المستخدمون أو مناداتهم لبعضهم البعض باستخدام الأسلوب المتعارف علية. مثلا كما في حالة استخدام الهاشتاغ Hashtag، اذ أصبح من النادر ان نرى منشورا لا يحوي سلسلة من المحارف مسبوقا بالرمز # وهي ميزة تجعل هذه السلسة من المحارف قابلة للنقر مما سيولد قائمة بالتغريدات التي تضمنت نفس السلسلة من المحارف المسبوقة بالرمز #. أي لائحة بكل التغريدات التي تضمنت نفس الهاشتاغ.

بالتأكيد كل ما سبق ذكره لن يفيد في تحديد الحالة العاطفية في التغريدة لذلك قمنا بكتابة خوارزمية لإزالة أي ورود للروابط والإيميل وال Mention وال Hashtag.

الخطوة الثالثة: التعامل كلمات التوقف Remove Stop Words:
في هذه الدراسة قمنا بمقارنة نتائج التصنيف مع كل من الحالات التالية:

- حذف وازالة كلمات التوقف
- الاحتفاظ بكلمات التوقف افتراضاً بأن هذه الكلمات قد تحمل دلالة عاطفية.

الخطوة الرابعة: تجذيع الكلمات Stemming:
تنقسم الكلمات العربية إلى ثلاث أنواع رئيسية هي: الأسماء، الأفعال، الحروف. والحروف عي عبارة عن المتصلات مثل حروف الجر والضمائر. تشتق الأسماء العربية والأفعال من الجذور بتطبيق قوالب على الجذور. تطبيق القوالب عادة يتضمن إضافة لواحق أو حذف أو استبدال حروف من الجذر. يبين الجدول 10 الكلمات التي يمكن توليدها من الجذر "*كتب*":

كتب ktb	He wrote	يكتب yhtb	He is writing	أكتب >ktb	I write
كاتب kAtb	Writer	كتَّاب ktAb	Book	كتَابه ktAbh	His book
وكتابة wktAbh	And his book	كتَابهم ktAbhm	Their book	كتب ktb	books

جدول 7 الكلمات التي يمكن توليدها من الجذر "كتب"

نظرا للتعقيد الصرفي في اللغة العربية فإن المعالجة الصرفية تساعد في استرجاع واحدات المعنى units of meaning. أحد الحلول هو استخدام light stemming ففي هذا النهج تتم إزالة البادئات واللواحق بشكل روتيني، وهناك نوعان لهذا النهج:

- Al-Stem

هذا النوع يطلق عليه "عدواني"، يقوم بإزالة البادئات واللاحقات التالية:
بادئات: (وال، قال، بت، يت، لت، مت، وت، ست، نت، بم، لم، وم، كم، قم، ال، ل، وي، لي، سي، في، وا، فا، لا، و، با)
لاحقات: (ات، وا، ون، وه، ان، تي، ته، تم، كم، هم، هن، ها، ية، تك، نا، ين، يه، ة، ه، ي، ا).

- Umass Light 10 stemmer

يقوم بإزالة البادئات واللاحقات التالية:
بادئات (ال، فال، بال، وال، كال، و).
لاحقات (ها، ان، ات، ون، ين، يه، ية، ة، ه، ي).

تفيد هذه المرحلة في تخفيض فضاء الكلمات، من خلال استبدال مجموعة الكلمات المشتقة من نفس الجذر بأصلها. يوجد نوعان من التجذيع:

- التجذيع الخفيف Light Stemming: يقوم بإزالة أدوات التعريف في بداية الكلمة والضمائر في نهايتها، ويترك الكلمة الأساسية كما هي.
- التجذيع الأساسي Stemming: يرد الكلمة إلى جذرها اللغوي.

في سياق هذا البحث قمنا باستخدام Light Stemmer ومن ثم استخدمنا أسلوب التجزيع الأساسي والذي يخفض بشكل واضح من حجم فضاء الكلمات.

فمثلا: الكلمات (تعلم، علم، يتعلم، معلم، سيتعلم، عليم، علامة، معلومة، معلوماتية) كلها سترد إلى جذر واحد فقط وهو "علم".
هذا يبين الفائدة الأساسية التي تقدمها عملية التجذيع في أي تطبيق يتعامل مع اللغات الطبيعية.
في هذا السياق قمت بعملية بحث مكثفة عن خوارزمية ناجعة لعمليات التجذيع الأساسي على اللغة العربية، حيث يوجد خوارزميتين شائعتين في هذا المجال:

1- خوارزمية ISRI
تعتمد على قواعد الصرف وعلم الأوزان.
2- خوارزمية Khoja
تعتمد على جداول مخزنة من الجذور الشائعة بالإضافة للقواعد الصرفية.

الخطوة الخامسة: التعامل مع ادوات النفي والتأكيد على المعنى ورموز المشاعر وعلامات التعجب والاستفهام، وهذا ما سنتطرق اليه لاحقا.

d-3-4-توليد أشعة واصفات البيانات Feature Vector Generation

يقصد بشعاع البيانات (15) التعبير عن كل مدخل من مدخلات مجموعة البيانات (تغريدة) بشعاع رياضي يفهمه الحاسوب ونقصد بأن يفهمه الحاسوب هنا أي أن يفهمه المصنِف Classifier المسؤول عن عملية التدريب.
ففي ميدان تعلم الآلة Machine Learning يتم تمرير بيانات التدريب مع الحالة العاطفية الصحيحة لكل منها إلى المصنف ليقوم بالتدرب عليها ومحاولة بناء نموذج رياضي Modelيعبر عن عملية الربط المنطقي بين هذه المدخلات وتلك المخرجات ليكون هذا النموذج المدرب قادراً مستقبلاً على التنبؤ بالحالة العاطفية بمجرد إدخال أشعة واصفات جديدة. تكمن أهمية هذه الخطوة في تأثيرها المباشر على الدقة النهائية لعملية التصنيف، حيث تتسابق الأوراق البحثية في هذا المجال على اقتراح الطرق المختلفة لنمذجة البيانات بهدف تحسين الدقة، وغالبا ما تكون ناقصة في مجال اللغة العربية. السبب في ذلك هو عدم وجود أدوات فعالة في مجال اللغة العربية للتعرف على أجزاء الكلام وهو ما يعرف بـ POS Tagging حيث يشير اختصار POS إلى Part Of Speech.
الإسلوب الشائع والمستخدم في العديد من دراسات تحليل المشاعر هو استخدام البيانات النصية مباشرة، حيث يتم اعتبار كل كلمة واردة في المدونة على أنها واصفة attribute في شعاع الواصفات، وقيمتها يمكن أن تختلف وفقا للمنهج المتبع:

- **المنهج الاول**: أبسط الأشكال هو أن يتم تمثيل الكلمة (الواصفة) في الشعاع بالرقم (1) ليعبر ذلك عن ورودها في النص الموافق أو الرمز (0) ليعبر ذلك عن غيابها.
- **المنهج الثاني**: هو أن نمثل كل كلمة بعدد يعبر عن تكرار ورودها Term Frequency في النص الموافق "التغريدة". تبدأ العملية باستخراج كل ال unigrams[4] وال bigrams[5] من المنشورات ثم الإحتفاظ بمن يتجاوز عتبة معينه من مرات التكرار ليتم اعتبارهم كسمات ضمن الشعاع.
ثم من أجل كل منشور نحسب تكرار كل feature مختار. من أجل كل منشور سينتج لدينا ال feature vector التالي:
(“word1: frequenc1، word2: frequenc 2، ”)

[4] 1-gram sequence (abcd > a، b، c، d)
[5] 2-gram sequence (abcd > ab، bc، cd)

N:	وثق	الله	غفر	رحم	راود	قلق	غريب	حيال	أزمة	سورية	..	..	..	Emotion
1	0.45	0.78	0.89	0.9	0	0	0	0	0	0	0	0	0	#ثقة
2	0	0	0	0	0.72	0.56	0.98	0.55	0.345	0.24	0	0	0	#قلق
3	..	..	..	..	..	..	..	..	..	..	..	..	..	..
4	..	..	..	..	..	..	..	..	..	..	..	..	..	..

جدول 25 : حالة اخذ كل كلمات التغريدة كواصفات ضمن شعاع الواصفات

- **المنهج الثالث**: وهو الأفضل وهو أسلوب مطور عن السابق، فمن اجل اعطاء أوزان للكلمات ضمن بيانات التدريب يتم الأخذ بعين الإعتبار كل من syntactic & stylistic features

Syntactic feature مثل تكرارات ال n-grams، تكرارات جذور الكلمات، وجود علامات الترقيم

Stylistic feature مثل العدد الكلي للكلمات، العدد الكلي للمحارف، تكرار المحارف الخاصة.

بالإضافة إلى الإستفادة من الدلالة العاطفية للكلمات وذلك عن طريق الإستعانة بقواميس تشير إلى الدلالة العاطفية لكلمات اللغة (الفقرة 2-b-3-4). وهنا قمنا بالمقارنة بين ثلاث اساليب لاحتساب هذا الوزن:

- TF-IDF (29) وهو اختصار لـ Term Frequency – inverse Document Frequency

TF-IDF:

"Q1" number of the term [x] in tweets which refer to emotion [e]. "Q2" number of the terms within tweets which refer to emotion[e].

$$fij = {}^{Q1}/_{Q2}$$

"N" number of all tweets with in training data. "dfi" number of all tweets which contain term [x] with in training data.

$$Freq\ of\ term_x\ in\ emotion_e =$$

$$(\frac{fij}{max_{fij}}) * \log(\frac{N}{dfi})/\log(2)$$

(16) weighted-TwF ○

weighted-TwF:

"D" number of all tweets with in training data. "ne" number of the tweets which refer to emotion [e]، and contain term [x]. "Q" Total number of terms in the d.

$$\text{Normalized\_Tweet\_Frequency} = {}^{ne}/_{Q}$$

$$\begin{cases} (ne < D) \ \ then\ weight = {}^{1}/_{ne} \\ else\ \ weiht = 0 \end{cases}$$

$$Freq\ of\ term_x\ in\ emotion_e = NTF * weight$$

Mod Tf-IDF ○

We modify TF-IDF to Be:

"Q" number of tweets which refer to the emotion [e].

"Z" number of the term [x] within tweets which refer to emotion [e]. "V" number of the terms within all tweets.

$$Freq\ of\ term_x\ in\ emotion_e = \frac{Z}{Q}/V$$

المشكلة في جميع الأشكال السابقة هي حجم شعاع الواصفات، والذي يمكن أن يصل إلى ألاف الواصفات وهكذا فإن هذا العدد يزداد مع ازدياد عدد الكلمات في مجموعة البيانات "المدونة المنمّطة بالمشاعر".

بمعنى آخر ستكون المهمة الملقاة على عاتق المصنِف في هذه الحالة هي أن يتعرف على الشكل البياني الذي سيتموج للحد الذي يستطيع فيه أن يفصل بين النقاط الواقعة في فضاء من 10 آلاف بعد ليتمكن بعد ذلك من التعرف على صنف نقطة جديدة من هذا الفضاء، وهي مهمة ليست بسيطة بالتأكيد.

الحل الذي فكرنا به هو تركيز الجهد على تصميم شعاع الواصفات بشكل ملخص بحيث نحمّله أكبر كمية ممكنة من المعلومات المميزة للنص "في حالتنا نحن نهتم بالمعلومات المعبرة عن نوع العاطفة في تغريدة معينة". لذلك قمنا باختزال شعاع الواصفات من الشكل الذي يحتوي كل كلمات التغريدة كواصفات إلى شكل آخر مكون من ثمانية واصفات فقط أو ستة "واصفة لكل صنف عاطفي". قيمة كل واصفة من هذه الواصفات ستكون عبارة عن عداد تراكمي يمثل مجموع أوزان كلمات الجملة وفقا للصنف العاطفي الذي يشير اليه هذا العداد.

مثال على هذا الشعاع، الشعاع التالي الذي يعبر عن المنشور "*انا متفاجئ انو الوقت رح يمر بشكل جيد*" والذي يشير بدرجة كبيرة إلى "التفاجؤ" – الشكل 26

		Sad	Joy	Surprise	Disgust	Fear	Anger
تفاجئ	surprise	0	0	0.8662509	0	0.2993424	0
وقت	time	0	0.3415608	0	0.2894182	0.2993424	0
يمر	pass	0	0	0.2887503	0	0	0
شكل	stat	0.6516933	0	0.5775006	0.5788364	0.2993424	0.8213903
جيد	well	0	0	0	0	0	0
F_Vector		0.6516933	0.3415608	1.7325018	0.8682546	0.8980272	0.8213903

جدول 26: شعاع يعبر عن تغريدة تشير على التفاجؤ

ملاحظة:

- سنقوم بتفصيل النتائج في فقرة – "تدريب المصنفات وتقييم الأداء".
- سنناقش حالات خاصة ضمن فقرة - "معالجة شعاع الواصفات 5-2-1".

4-3-e-تدريب المصنفات

قمنا بتدريب واختبار مجموعة من المصنفات وهي:

SMO

weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"

Naïve Bayes

weka.classifiers.bayes.NaiveBayes -output-debug-info

RandomForest

weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

وذلك باستخدام مجموعة البيانات التي قمنا بجمعها من موقع التواصل الاجتماعي تويتر 1320تغريده ذات دلالة عاطفية".
الشكل التالي يبين مجموعة التغريدات التي استخدمناها ضمن المدونة المنمّطة بالمشاعر بهدف تدريب المصنفات:

Examples of Training "مدونة التدريب":

Sad's Instances	220
Joy's Instances	220
Surprise's Instances	220
Disgust's Instances	220
Fear's Instances	220
Anger's Instances	220
	Total Instances = 1320

بالإضافة إلى أمثلة إضافية خاصة بعملية اختبار المصنفات، عددها يقارب 30% من عدد أمثلة التدريب.

Examples of Testing "مدونة الاختبار":

Sad's Instances	65
Joy's Instances	65
Surprise's Instances	65
Disgust's Instances	65
Fear's Instances	65
Anger's Instances	65
	Total Instances = 390

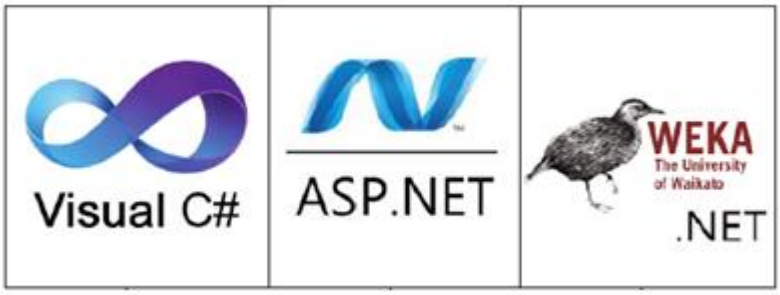
من أجل كل مصنف مما سبق، قمنا ببناء عملية التدريب في WEKA.Net مع أخذ المكون Validation بعين الإعتبار، والذي يقوم بتقسيم البيانات إلى عشر أقسام، ثم من خلال عشر دورات، يستخدم في كل دورة تسع أقسام من البيانات في عملية التدريب والقسم العاشر في عملية الإختبار، ثم في الدورة التي تليها يأخذ تسع أقسام لعملية التدريب وقسم مختلف في عملية الإختبار وهكذا حتى يتم استخدام جميع الأقسام في عملية الاختبار.
أما عملية تعميم المصنفات ستكون على مدونة منفصلة عن مدونة التدريب، تشكل مدونة الاختبار 30% من حجم مدونة التدريب.

سنأخذ بعين الإعتبار استخدام المكون Gain Ratio Validation بهدف تحديد الكلمات الأكثر تأثيراً على الدلالة العاطفية، كما سنأخذ بعين الاعتبار استخدام الطريقة Corss Validation 10-Folds[6]

[6]أظهرت تجارب واسعة النطاق أن 10-Folds يعتبر أفضل خيار للحصول على تقدير دقيق لعمل المصنف (since CART book by Breiman، Friedman، Stone، Olsen 1994)

5-1- الأدوات المستخدمة

لتحقيق هذا العمل، قمنا باستخدام التقنيات التالية:



الشكل 27: التقنيات المستخدمة في بناء التطبيق

1- Weka

WEKA هي أداة برمجية مفتوحة المصدر تم برمجتها باستخدام Java وهي توفر بيئة لاستخدام عدد من خوارزميات التنقيب عن البيانات.

من أجل استخدام هذه الأداة ضمنnet. نحتاج إلى وسيط لتفسير كل مما يلي للأخر .net $ Java class، ومن أجل ذلك قمنا باستخدام IKVM. NET والذي يتطلب ما يلي:

- تحميل Java SDK

عندما نريد عمل تفسير لكود جافا ضمن Dot Net فإننا بحاجة إلى مفسر جافا java compiler

وبما أن iknm.net لا يحتوي مفسر لذلك يجب الاستعانة بأي واحد أخر مثل Java SDK

- تحميل مكتبات IKVM.Dll إلى مشروع Dot Net

وبالتحديد IKVM.runtimedll بالإضافة إلى IKVM.openJDK

إن IKVM. NET يؤمن ما يلي:

- A Java Virtual Machine implemented in .NET
- A .NET Implementation of the Java class libraries
- Tools that enable Java and .NET interoperability

وهكذا فإن الهدف من استخدام IKVM.NET هو تحويل ملف WEKA.jar إلى WEKA.DLL وبالتالي إمكانية استخدام إمكانيات هذه الأداة ضمن DOT NET.

قمت بالعمل على الإصدارات التالية:

- WEKA 7.1.4532.2
- IKVM 7.1.4532.2

2- ASP.NET

ASP اختصار لـ Active Server Pages والتي تعني "صفحات الخادم النشط"، هو إطار لتطبيقات الويب تم تطويره وتسويقه من خلال شركة Microsoft، من أجل إعطاء القدرة للمبرمجين على بناء مواقع ويب ديناميكية، تطبيقات ويب وخدمات ويب.

تم إصدار النسخة الأولى منه عام 2002، تعتبر هذه التقنية خلفا لتقنية ASP، كما أن ASP.NET تم بناؤها لتستند على تقنية CLR (Common Language Runtime). مما يسمح للمبرمجين بكتابة أكوادهم الخاصة بإطار ASP.NET باستخدام أي لغة برمجية يفضلونها على أن تكون مدعومة بإطار عمل دوت نت .NET.

3- Visual C#

هو بيئة التطوير المتكاملة والرئيسية من Microsoft، يتيح برمجة واجهة المستخدم الرسومية والبرامج النصية إلى جانب Windows Forms ومواقع الويب وتطبيقات وخدمات الويب المدعومة من قبل Microsoft Windows وMobile Windows ذات إطار عمل .NET وMicrosoft Silver Light.

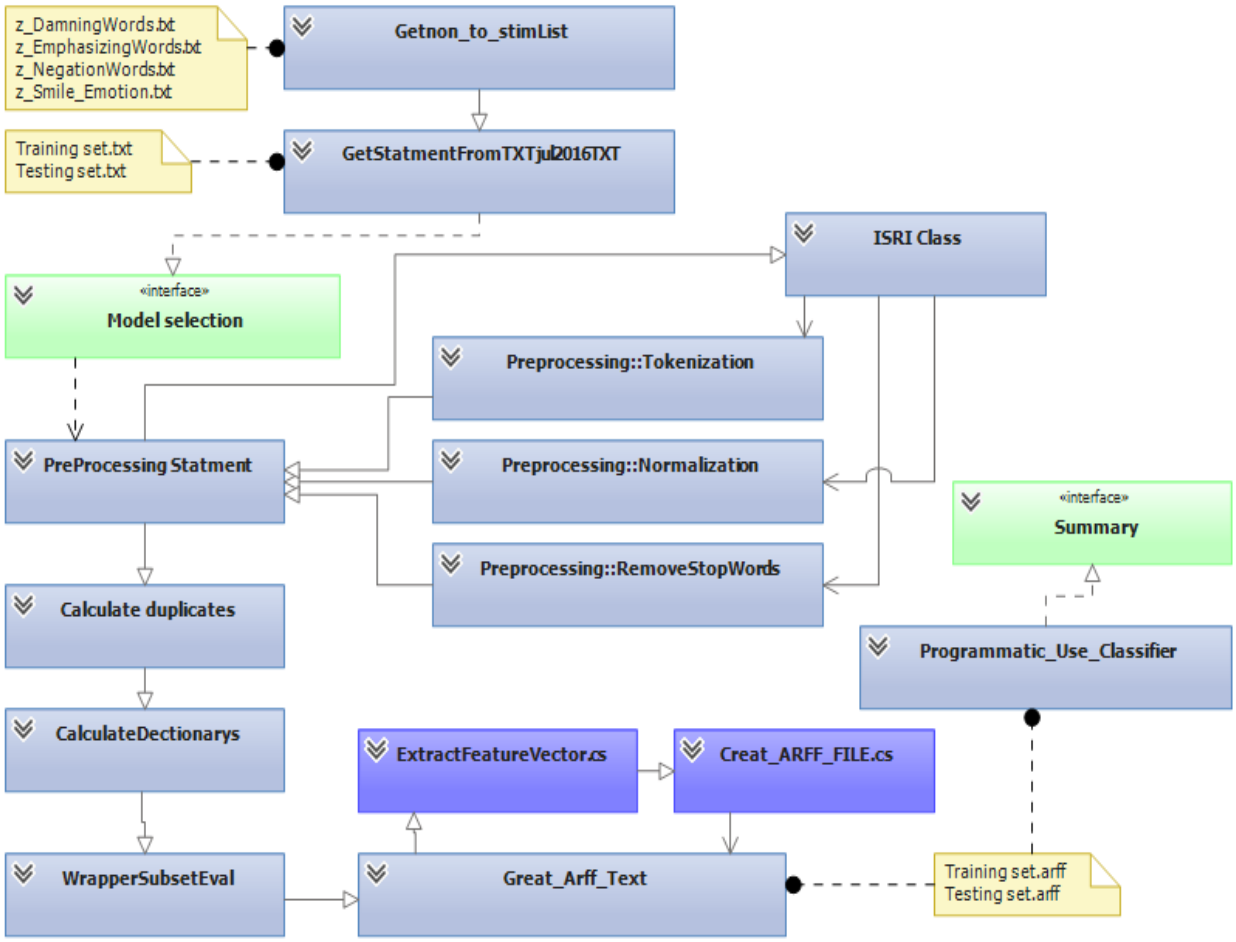
يحتوي Visual Studio 2010 على محرر أكواد يدعم تقنية IntelliSence وإعادة كتابة الكود، ويحتوي أيضا على مترجم يكشف الأخطاء الإملائية في الأكواد، ويحتوي أيضا على مصمم نماذج لبناء واجهات مستخدم رسومية ومصمم ويب ومصمم فئات ومصمم مخطط لقواعد البيانات ومصمم لتقارير الكريستال.

كما يدعم Visual Studio 2010 العديد من لغات البرمجة مثل Microsoft Visual C++ وMicrosoft Visual C# وMicrosoft Visual Basic وJavaScript والعديد أيضا من لغات الترميز مثل HTML، XML، XHTML وXSL.

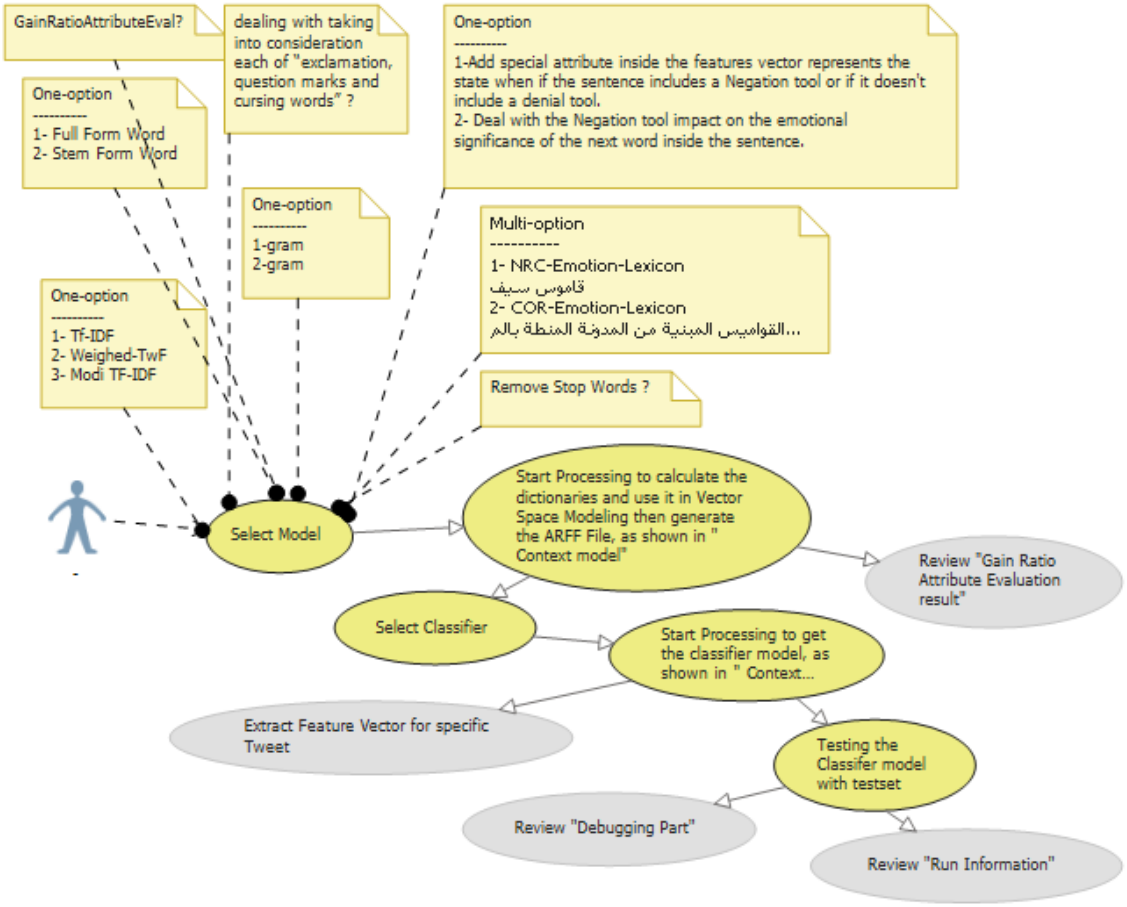
2-5- التطبيق البرمجي

قمنا بإنجاز كل العمل السابق الذكر برمجياً في واجهه خاصة للتعامل مع مجموعات البيانات والقواميس وأشعة الواصفات.

الشكل التالي يعرض النموذج التشغيلي للنظام ضمن السياق البرمجي ابتداء من مرحلة العمل على معالجة مدونة التدريب وصولاً إلى مرحلة تدريب المصنفات وبناء نموذج التصنيف واختباره على أمثلة تدريب جديدة وعرض النتائج.



Context model :28 الشكل



الشكل 29Use Case Diagram

استعرضنا سابقا طريقة استخراج شعاع الواصفات اذ لاحظنا شعاع الواصفات لإحدى التغريدات "*أنا متفاجئ انو الوقت رح يمر بشكل جيد*" – الفقرة 4-3-d. والتي تم تقييم هذه التغريدة يدويا بأنها تشير إلى الصنف العاطفي \التفاجؤ\ فذكرنا أنه وبعد عملية المواءمة النصية للتغريدة، تبدأ الخوارزمية بتمثيل التغريدة كشعاع ضمن فضاء شعاعي وذلك من خلال البحث عن وزن كل كلمة في التغريدة ضمن كل صنف عاطفي "بعد عملية المواءمة النصية"، فتكون البداية في التغريدة السابقة مع كلمة "أنا"، ليتم البحث عن وزن هذه الكلمة ضمن كل صف من صفوف المشاعر، بعد الحصول على هذه الأوزان يتم الإحتفاظ بها من أجل كل عداد من عدادات صفوف المشاعر المدروسة. وهكذا تتكرر العملية من أجل كل كلمة في التغريدة.

5-2-1- معالجة شعاع الواصفات

بعد الإنتهاء من مرحلة استخراج شعاع الواصفات الخاص بالتغريدة تقوم خوارزمية العمل بمعالجة هذا الشعاع ليتم الأخذ بعين الإعتبار كل من الحالات التالية:

5-2-1-1- تأثير الوجوه التعبيرية على شعاع الواصفات النهائي

نلاحظ في الشكل 30 تغير قيم شعاع الواصفات النهائي مع كل مرة نقوم فيها بإدخال أو تغيير وجه تعبيري جديد إلى التغريدة "*أنا اليوم قدمت أصعب امتحان بالسنة* ":

أنا اليوم قدمت أصعب امتحان بالنسبة كلا :)

	Sad	Joy	Surprise	Disgust	Fear	Anger
أنا	3.385707	4.391983	0.7800519	1.488157	1.482759	0.7328137
قدم	0.9233747	0.5855977	1.040069	0.4960524	0.7413794	0.4885425
صعب	0	0	0	0	0.3706897	0.2442712
امتحان	0	0	0	0	0	0
سنة	0	0.2927988	0	0	0	0.2442712
(:	0	2.63519	0	0	0	0
F.vector	4.3090817	15.81114	1.8201209	1.9842094	2.5948281	1.7098986

أنا اليوم قدمت أصعب امتحان بالنسبة كلا ☹

	Sad	Joy	Surprise	Disgust	Fear	Anger
أنا	3.385707	4.391983	0.7800519	1.488157	1.482759	0.7328137
قدم	0.9233747	0.5855977	1.040069	0.4960524	0.7413794	0.4885425
صعب	0	0	0	0	0.3706897	0.2442712
امتحان	0	0	0	0	0	0
سنة	0	0.2927988	0	0	0	0.2442712
☹	0	0	0	0	0	0
F.vector	6.263622	5.27038	1.820121	1.984209	2.594828	3.419797

الشكل 30: تأثير الوجوه التعبيرية على شعاع الواصفات

فكما سبق الذكر، قمنا بإنشاء قاموس الوجوه التعبيرية، كل واحدة من هذه الوجوه تمتلك القدرة للتأثير على المعنى "التأثير على العاطفة" بمقدار محدد "كما تم التعبير عنه في القاموس".
بعد أن يتم تحليل واستخراج شعاع الواصفات النهائي للتغريدة، تقوم خوارزمية التحليل بإعادة المرور على التغريدة للتأكد من وجود أي وجه تعبيري ضمنها. ففي حال ورود وجه S سيتم إضافة قيمة عددية إلى قيمة V العداد التراكمي C الخاص بكل صنف عاطفي i بمقدار يساوي إلى حاصل ضرب القيمة V لهذا العداد مع شدة تأثير هذا الوجه SM على هذا الصنف العاطفي i "كما ورد في القاموس الموافق".

$$V_{C_i} = V_{C_i} + (V_{C_i} * SM_i)$$

هنا يجدر بنا ملاحظة أن كل وجه تعبيري قد يحمل قيمة "وزن" ضمن كل صنف عاطفي بالإضافة الى القيمة المعبر عنها ضمن القاموس الخاص بالوجوه التعبيرية، يتم احتساب هذا الوزن انطلاقا من بيانات التدريب (الفقرة 1-b-3-4).

مثلا: الوجه **):** يحمل قيمة دلالية ضمن صنف الفرح بينما أن الوجه ☹ لا يحمل أي قيمة دلالية لأي صنف عاطفي، والسبب بذلك يعود إلى أن الوجه الثاني لم يرد ضمن أي تغريدة من تغريدات مدونة التدريب وبالتالي لم يتم احتساب أي وزن له على عكس الوجه الأول و على عكس كلمات التغريدة.

5-2-1-2- تأثير الكلمات المشددة للمعنى على شعاع الواصفات النهائي

إن ورود الكلمات المشددة للمعنى ضمن جملة في اللغة العربية قد يأخذ العديد من الأشكال التي تختلف عن بعضها البعض وفقا للمكان الذي تأتي به الكلمة المشددة للمعنى ضمن هذه الجملة.
فقد نكتب "أنا واثق من النجاح كتير" ونكتب أيضا "أنا كتير واثق من النجاح" أو "أنا واثق كتير من النجاح".
في جميع الأحول تقوم الكلمة المشددة على المعنى بالتأثير على العاطفة الضمنية التي تحتويها الكلمة التي تسبق كلمة التشديد أو التي تليها والتي "هذه العاطفة" تتطابق مع العاطفة الضمنية التي تشير اليها الجملة كاملة.

في حال ورود كلمة مشددة على المعنى ضمن التغريدة تقوم الخوارزمية بالبحث عن الصنف العاطفي E_{max} الذي تشير اليه التغريدة "الواصفة التي تمتلك أكبر قيمة ضمن شعاع الواصفات". ثم تقوم بمضاعفة قيمة العدادات التي تشير الى الصنف العاطفي E_{max} وذلك لكل من الكلمة السابقة واللاحقة لكلة التشديد على المعنى.
"الخوارزمية تراعي حالة ورود كلمة التشديد في الجملة". الشكل 31 يستعرض تأثير الكلمات المشددة للمعنى على نتائج تحليل شعاع واصفات يأخذ بعين الاعتبار ستة واصفات "غضب، خوف، اشمئزاز، تفاجئ، فرح والحزن".

يا اخي هدا الكلام مو صحيح

	Sad	Joy	Surprise	Disgust	Fear	Anger
اخي	0	0	1031515	0	0	0
هدا	0	0	873018.8	0	0	873018.8
كلم	640825.9	640825.9	640825.9	640825.9	640825.9	640825.9
مو	592283.2	592283.2	592283.2	592283.2	592283.2	592283.2
صح	0	873018.8	873018.8	0	0	0
F. vector	1233109	2106128	4010662	1233109	1233109	2106128

يا اخي هدا الكلام نهائيا مو صحيح

	Sad	Joy	Surprise	Disgust	Fear	Anger
اخي	0	0	1031515	0	0	0
هدا	0	0	873018.8	0	0	873018.8
كلم	640825.9	640825.9	1281652	640825.9	640825.9	640825.9
نهايا	0	0	0	0	0	0
مو	592283.2	592283.2	592283.2	592283.2	592283.2	592283.2
صح	0	873018.8	873018.8	0	0	0
F. vector	1233109	2106128	5933139	1233109	1233109	2106128

الشكل 31: تأثير الكلمات المشددة للمعنى على شعاع الواصفات

5-2-1-3- تأثير الكلمات النافية للمعنى على شعاع الواصفات النهائي

النفي في اللغة العربية مصدر من الجذر اللغوي (نفى) وضده: الإثبات. ويأتي النفي في اللغة بعدة معانٍ ومنها: الإنكار أو التكذيب، والإبعاد. فهو أسلوب من الأساليب العربية، يفيد الإنكار والإخبار بعدم وقوع شيء معين في الماضي أو الحاضر أو المستقبل.
إسلوب النفي يتم بإدخال إحدى أدوات النفي إلى الجملة فتقسم الجملة إلى جزئين "الجزء السابق لأداة النفي والجزء اللاحق لها".
للتعامل مع حالات النفي، قمنا بالمقارنة بين إسلوبين:

1- Negation_Feature إضافة واصفة خاصة ضمن شعاع الواصفات تعبر عن احتواء الجملة على أداة نفي.
2- Negation_Swap التعامل مع تأثير اداة النفي على الدلالة العاطفية للكلمة التالية للأداة ضمن الجملة.

ففي الاسلوب الثاني، عادة يقع غرض النفي على الجزء الثاني من الجملة المنفية "الاحق لأدوات النفي" وهكذا فإنه عند احتواء هذه الجملة "التغريدة" على إحدى هذه الكلمات فإن العاطفة الضمنية للجملة التالية لأدات النفي سوف تعكس.
فبعد التحقق من كل كلمات التغريدة وعند التأكد من ورود أداة نافية للمعنى "بالاستعانة بقاموس أدوات النفي سابق الذكر"، فإن العاطفة الضمنية لكل الكلمات التالية لهذه الأداة سوف تعكس فيتم تبادل قيم العدادات لكل كلمة كالتالي "إفتراضاً":

- عداد الحزن سيتبادل قيمته مع عداد الفرح.
- عداد التفاجؤ سيتبادل قيمته مع عداد التوقع.
- عداد الاشمئزاز سيتبادل قيمته مع عداد الفرح.
- عداد الغضب سيتبادل قيمته مع عداد الفرح.
- عداد الثقة سيتبادل قيمته مع عداد الخوف.
- عداد التوقع سيتبادل قيمته مع عداد الثقة.

مثال، لنأخذ أولاً تغريدة لا تحتوي أي كلمة نافية للمعنى

انا واثق انو الوقت رح يمر بشكل جيد

Term	حزن	فرح	تفاجئ	اشمئزاز	خوف	غضب	ثقة	توقع
انا	3.385707	4.391983	0.7800519	1.488157	1.482759	0.7328137	4.009072	1.07189
وثق	0	0	0.5200346	0	0	0	6.949059	0
انو	0.3077916	0.8783965	0.5200346	0.9921048	1.112069	0.2442712	2.138172	1.07189
وقت	0	0.2927988	0	0.2480262	0.3706897	0	0.534543	0.8039171
رح	0	0.2927988	0.2600173	0.7440786	0.3706897	0.2442712	2.672715	1.07189
يمر	0	0	0.2600173	0.2480262	0	0	0.534543	0
شكل	0.6155831	0.5855977	0.7800519	0.9921048	0.7413794	0.9770849	1.336357	0.2679724
جيد	0	0	0	0	0	0	0.8018144	0

حزن	فرح	تفاجئ	اشمئزاز	خوف	غضب	ثقة	توقع
4.309082	6.441575	3.120208	4.712498	4.077587	2.198441	18.97627	4.28756

الشكل 32: تحليل شعاع الواصفات لتغريدة تشير بدرجة كبيرة إلى الثقة

عند تعديل التغريدة من "*أنا واثق انو الوقت رح يمر بشكل جيد*" لتصبح "*أنا واثق انو الوقت مارح يمر بشكل جيد*" سنلاحظ تبادل قيم العدادات وذلك لكل كلمة تالية لأداة النفي التي تم الكشف عنها، "لاحظ عدادات الكلمات: يمر – شكل – جيد":

انا واثق انو الوقت مارح يمر بشكل جيد

Term	حزن	فرح	تفاجئ	اشمئزاز	خوف	غضب	ثقة	توقع
انا	3.385707	4.391983	0.7800519	1.488157	1.482759	0.7328137	4.009072	1.07189
وثق	0	0	0.5200346	0	0	0	6.949059	0
انو	0.3077916	0.8783965	0.5200346	0.9921048	1.112069	0.2442712	2.138172	1.07189
وقت	0	0.2927988	0	0.2480262	0.3706897	0	0.534543	0.8039171
مرح	0.2927988	0	0.2600173	0.8018144	0.6155831	0	0.3706897	0
يمر	0	0	0.2600173	0.534543	0	0.2480262	0	0
شكل	0.5855977	0.9770849	0.7800519	1.336357	0.6155831	0.9921048	0.7413794	0.2679724
جيد	0	0	0	0.8018144	0	0	0	0

حزن	فرح	تفاجئ	اشمئزاز	خوف	غضب	ثقة	توقع
4.670584	6.80331	3.120208	3.840357	3.843914	4.451614	14.86201	3.21567

الشكل 33: تحليل شعاع الواصفات لتغريدة تحتوي اداة نافية للمعنى وذلك عند اعتماد اسلوب مبادلة القيمة

كذلك الأمر مع هذه التغريدة "*أنا ماني واثق انو الوقت رح يمر بشكل جيد*":

انا ماني واثق انو الوقت رح يمر بشكل جيد

Term	حزن	فرح	تفاجئ	اشمئزاز	خوف	غضب	ثقة	توقع
انا	3.385707	4.391983	0.7800519	1.488157	1.482759	0.7328137	4.009072	1.07189
مني	0	0	0	0	0	0	0	0
واثق	0	0	0.5200346	6.949059	0	0	0	0
انو	0.8783965	0.2442712	0.5200346	2.138172	0.3077916	0.9921048	1.112069	1.07189
وقت	0.2927988	0	0	0.534543	0	0.2480262	0.3706897	0.8039171
رح	0.2927988	0.2442712	0.2600173	2.672715	0	0.7440786	0.3706897	1.07189
يمر	0	0	0.2600173	0.534543	0	0.2480262	0	0
بشكل	0.5855977	0.9770849	0.7800519	1.336357	0.6155831	0.9921048	0.7413794	0.2679724
جيد	0	0	0	0.8018144	0	0	0	0

حزن	فرح	تفاجئ	اشمئزاز	خوف	غضب	ثقة	توقع
4.851335	7.616324	3.120208	4.082984	3.532351	15.70002	4.932446	4.28756

الشكل 34: تحليل شعاع الواصفات لتغريدة تحتوي اداة نافية للمعنى وذلك عند اعتماد اسلوب مبادلة القيمة

5-2-2- تحويل مجموعة بيانات التدريب إلى ملف أشعة الواصفات

وهي المرحلة الأخيرة من التطبيق، فبعد اختبار العديد من النصوص والتأكد من سير خوارزمية المواءمة بشكل منطقي وفقاً للقواميس، ننتقل للخطوة الأخيرة وهي تحويل مجموعة البيانات Dataset إلى ملف يحوي جميع أشعة الواصفات.

وبما أن العمل مع خوارزميات التصنيف سيكون من خلال الأداة WEK Data mining tools فإن مجموعة البيانات يجب ان يتم تمثيلها وفقا لصيغة Syntax محددة تفهمها هذه الأداة وذلك ضمن ملف من النوع ARFF:

ملفات Attribute-Relation File Format "arff" وهي ملفات نصية تصف قائمة من الحالات والسمات Instances والتي تتقاسم مجموعة من الواصفات Attributes فيما بينها.

بدايةً يجب التصريح عن كل الواصفات Attributes وهي الأصناف العاطفية نفسها، سنتعامل مع هذه الأصناف من خلال العدادات التراكمية الخاصة بكل منها لذلك يجب أن تكون أنماط هذه الصفات من النوع numeric.

```
@relation Entries
@attribute sad numeric
@attribute joy numeric
@attribute sur numeric
@attribute loa numeric
@attribute fea numeric
@attribute ang numeric
@attribute con numeric
@attribute exp numeric
```

الشكل 35: طريقة التصريح عن انماط الواصفات التي تمثل شعاع البيانات ضمن ملف التدريب ARff

في قسم أخر من ملف ARFF يجب التصريح عن الحالات التي تحتويها المدونة، وهي مجموعة الأصناف العاطفية "المشاعر" التي يتم تصنيف التغريدات اليها.

```
@attribute Emotion {حزن،فرح،تفاجئ، اشمئزاز،خوف، غضب،ثقة،توقع}
```

الشكل 36: طريقة التصريح عن الأصناف العاطفية التي تصنف اليها التغريدات

وهكذا فإن كل تغريدة سيتم تمثيلها كـ اجتماع لواحدة من الحالات Instance مع مجموعة من الصفات Attribute وهذا ما سيحتوي عليه القسم الأخير من هذا الملف.

```
@data
14.43751, 5.53476, 37.99767, 5.021745, 5.133338, 40.94073, 0, 20.98873, توقع
0, 0.7906801, 0, 0, 0, 1.023518, 1.300001, 12.10888, توقع
4.812504, 0, 0, 0, 0, 0.5117592, 0, 0, حزن
12.03126, 7.9068, 7.718277, 21.7609, 5.133338, 15.86453, 1.300001, 3.229035, غضب
3.609379, 6.325441, 9.499417, 6.695659, 0, 6.65287, 2.600003, 4.843553, تفاجئ
2.406252, 5.53476, 10.68684, 5.021744, 17.96668, 9.211666, 9.100009, 12.10888, تفاجئ
0, 1.58136, 2.374855, 0, 2.566669, 4.094073, 1.300001, 9.687106, توقع
0, 0.7906801, 2.374854, 1.673915, 0, 3.582315, 11.70001, 7.265329, توقع
 •
 •
 •
```

الشكل 37: تمثيل بيانات التدريب ضمن ملف التدريب ARFF

قمنا بإجراء مجموعة من الإختبارات بهدف المقارنة بين طرائق مختلفة لتصنيف الدلالة العاطفية من التغريدات

6-1- "6 مشاعر" + "أشعة واصفات ذات سمات بعدد كلمات المدونة"

بداية قمنا باختبار مجموعة من المصنفات على مدونة التدريب التي تم تمثيلها شعاعياً بشكل يأخذ بعين الإعتبار كل كلمات المدونة كسمات ضمن شعاع الواصفات، وذلك بهدف استعراض دقة المصنفات عند التعامل مع أشعة ذات فضاء سمات واسع جداً وغير قابل للحصر.

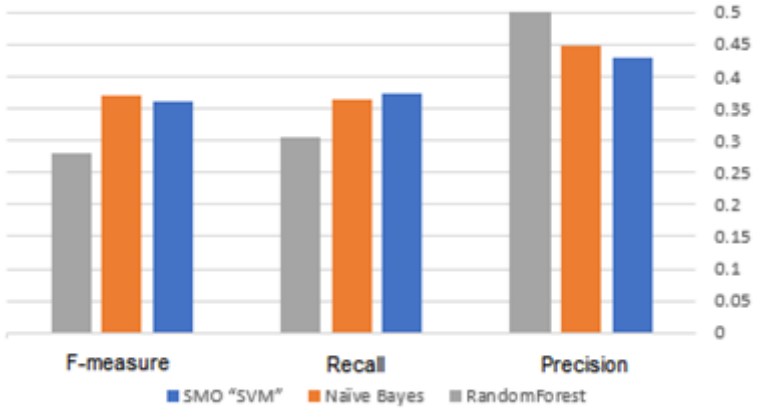
من أجل كل مصنف، قمنا ببناء عملية التدريب والاختبار في WEKA.Net مع أخذ الخيار Validation بعين الاعتبار وكانت النتائج كالتالي:

النتائج:

	AVG			سعادة			اشمئزاز			حزن		
ALG	**P**	**R**	**F**	P	R	F	P	R	F	P	R	F
SMO	**0.428**	**0.373**	**0.362**	0.281	0.677	0.397	0.442	0.224	0.297	0.418	0.155	0.227
NB	**0.449**	**0.365**	**0.369**	0.573	0.323	0.413	0.241	0.664	0.354	0.358	0.297	0.325
CRF	**0.558**	**0.304**	**0.279**	0.279	0.899	0.342	0.652	0.099	0.171	0.632	0.081	0.144

غضب			تفاجئ			خوف		
P	R	F	P	R	F	P	R	F
0.35	0.468	0.4	0.611	0.357	0.451	0.472	0.338	0.394
0.387	0.279	0.325	0.509	0.37	0.429	0.619	0.258	0.364
0.349	0.24	0.285	0.771	0.24	0.366	0.75	0.238	0.362

جدول 8: نتائج المصنفات عند أخذ كلمات المدونة كسمات ضمن شعاع الواصفات



الشكل 38: نتائج المصنفات عند أخذ كلمات المدونة كسمات ضمن شعاع الواصفات

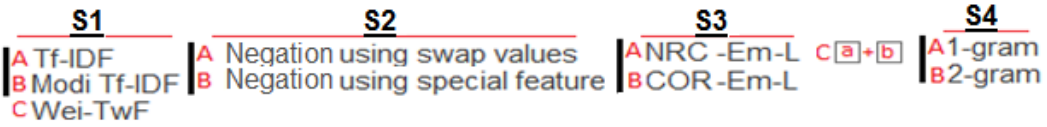
نلاحظ في هذه النتائج قيم متدنية لدقة التصنيف لذلك عملنا على تحسين هذه النتائج من خلال تحسين شكل أشعة بيانات التدريب وذلك من خلال اختصار شعاع الواصفات (الفقرة 4-3-d).

6-2- "6 مشاعر" + "أشعة واصفات ذات 6 سمات رئيسية"

هكذا، وبعد تحسين أشعة الواصفات (الفقرة 4-3-d) قمنا بتدريب المصنِفات وفقاً لمجموعة من الصيغ "النماذج" كل من هذه النماذج يختلف عن الأخر تبعا لمجموعة الخيارات التالية:

- طريقة توزين الكلمات.
- اسلوب التعامل مع حالات النفي.
- القواميس المستخدمة.
- الأساس N-gram المستخدم في بناء أشعة الواصفات.
- أسلوب التعامل مع الكلمة "بشكلها الكامل أو بعد التجذيع".

Symbols Explain:

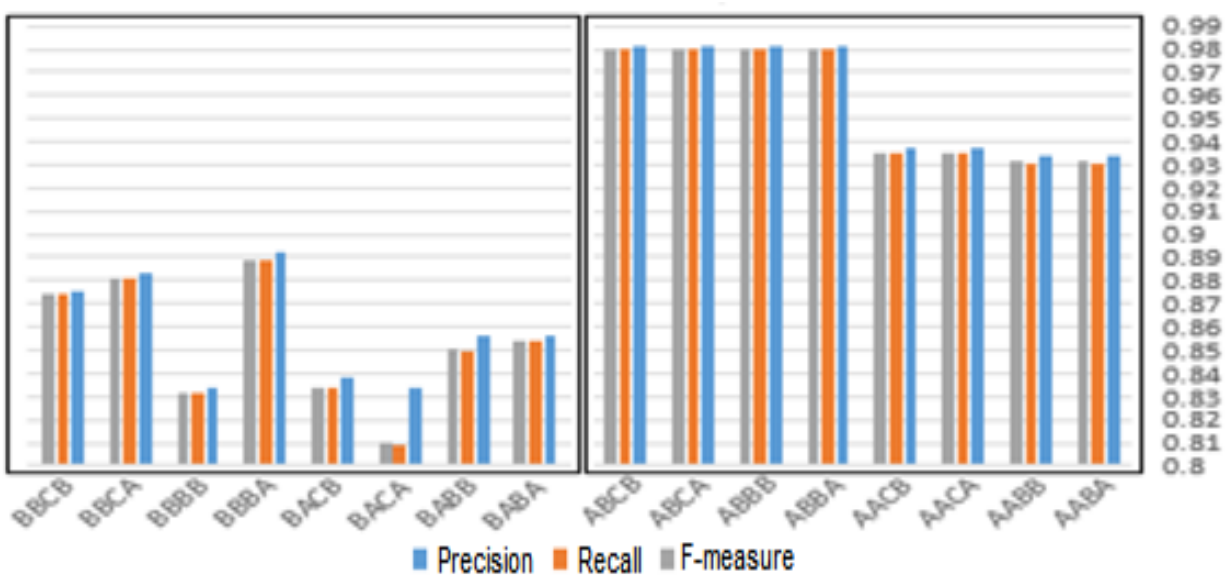


الشكل 39: Symbols Explain

A ◉ Tf-IDF B ○ Modi Tf-IDF C ○ Wei-TwF	One-option	الفقرة 4-3-d
A ◉ النفي باستخدام مبادلة القيم B ○ النفي باستخدام خاصة جديدة	One-option	1- Negation_Swap 2- Negation_Feature (الفقرة 5-2-1-3)
A ☐ NRC-Emotion-Lexicon B ☑ COR-Emotion-Lexicon C : A+B	Multi-option	(28)-Lexicon NRC-Emotion قاموس سيف COR-Emotion-Lexicons القواميس المبنية من المدونة المنمّطة بالمشاعر (الفقرة 4-3-b)
A ◉ 1-gram B ○ 2-gram	One-option	1-gram 2-gram

الشكل 40: Symbols Explain

بعد بناء مجموعة النماذج التي تتنوع ضمن طيف يشمل كل الخيارات السابقة مع اعتماد التعامل مع الكلمة بعد تجذيعها باستخدام ISRI، قمنا بتدريب المصنف CRF على كل منها وكانت النتائج – الشكل 41:



الشكل 41: مقارنة نتائج المصنف CRF مع كل من النماذج

				Average			Sadness			Joy			Surprise		
S1	**S2**	**S3**	**S4**	**P**	**R**	**F**	P	R	F	P	R	F	P	R	F
A	**A**	**C**	**A**	**0.934**	**0.931**	**0.932**	0.833	0.877	0.854	0.932	0.907	0.919	0.986	0.954	0.97
A	**A**	**C**	**B**	**0.934**	**0.931**	**0.932**	0.833	0.877	0.854	0.932	0.907	0.919	0.986	0.954	0.97
A	**A**	**B**	**A**	**0.937**	**0.935**	**0.935**	0.843	0.909	0.875	0.952	0.927	0.94	0.993	0.947	0.97
A	**A**	**B**	**B**	**0.937**	**0.935**	**0.935**	0.843	0.909	0.875	0.952	0.927	0.94	0.993	0.947	0.97
A	**B**	**C**	**A**	**0.981**	**0.98**	**0.98**	0.968	0.974	0.971	0.98	0.987	0.983	0.993	0.967	0.98
A	**B**	**C**	**B**	**0.981**	**0.98**	**0.98**	0.968	0.974	0.971	0.98	0.987	0.983	0.993	0.967	0.98
A	**B**	**B**	**A**	**0.981**	**0.98**	**0.98**	0.95	0.981	0.965	0.974	0.987	0.98	0.993	0.967	0.98
A	**B**	**B**	**B**	**0.981**	**0.98**	**0.98**	0.95	0.981	0.965	0.974	0.987	0.98	0.993	0.967	0.98
B	**A**	**C**	**A**	**0.856**	**0.854**	**0.854**	0.854	0.76	0.804	0.825	0.841	0.833	0.881	0.921	0.9
B	**A**	**C**	**B**	**0.856**	**0.85**	**0.851**	0.729	0.857	0.788	0.86	0.854	0.857	0.938	0.796	0.861
B	**A**	**B**	**A**	**0.834**	**0.809**	**0.81**	0.905	0.617	0.734	0.915	0.715	0.803	0.841	0.868	0.854
B	**A**	**B**	**B**	**0.838**	**0.834**	**0.834**	0.804	0.747	0.774	0.816	0.854	0.835	0.901	0.776	0.834
B	**B**	**C**	**A**	**0.892**	**0.889**	**0.889**	0.914	0.825	0.867	0.794	0.921	0.853	0.912	0.888	0.9
B	**B**	**C**	**B**	**0.834**	**0.832**	**0.832**	0.836	0.825	0.83	0.837	0.848	0.842	0.777	0.849	0.811
B	**B**	**B**	**A**	**0.883**	**0.881**	**0.881**	0.822	0.87	0.845	0.855	0.901	0.877	0.883	0.842	0.862
B	**B**	**B**	**B**	**0.875**	**0.874**	**0.874**	0.813	0.844	0.828	0.847	0.881	0.864	0.848	0.842	0.845
		Average		**0.908**	**0.904**	**0.904**	0.866	0.864	0.862	0.901	0.901	0.900	0.931	0.903	0.916

				Disgust			Fear			Anger		
S1	S2	S3	S4	P	R	F	P	R	F	P	R	F
A	A	C	A	0.873	0.935	0.903	0.981	0.994	0.987	1	0.919	0.958
A	A	C	B	0.873	0.935	0.903	0.981	0.994	0.987	1	0.919	0.958
A	A	B	A	0.91	0.922	0.916	0.952	0.994	0.972	0.971	0.905	0.937
A	A	B	B	0.91	0.922	0.916	0.952	0.994	0.972	0.971	0.905	0.937
A	B	C	A	1	0.994	0.997	0.957	0.981	0.969	0.986	0.98	0.983
A	B	C	B	1	0.994	0.997	0.957	0.981	0.969	0.986	0.98	0.983
A	B	B	A	0.981	1	0.99	0.987	0.975	0.981	1	0.973	0.986
A	B	B	B	0.981	1	0.99	0.987	0.975	0.981	1	0.973	0.986
B	A	C	A	0.771	0.877	0.821	0.934	0.899	0.916	0.871	0.824	0.847
B	A	C	B	0.835	0.857	0.846	0.932	0.861	0.895	0.843	0.872	0.857
B	A	B	A	0.788	0.844	0.815	0.928	0.899	0.913	0.616	0.912	0.736
B	A	B	B	0.85	0.844	0.847	0.884	0.867	0.875	0.768	0.919	0.837
B	B	C	A	0.903	0.903	0.903	0.933	0.88	0.906	0.895	0.919	0.907
B	B	C	B	0.87	0.779	0.822	0.865	0.848	0.856	0.817	0.845	0.831
B	B	B	A	0.935	0.844	0.887	0.918	0.924	0.921	0.882	0.905	0.893
B	B	B	B	0.918	0.877	0.897	0.907	0.93	0.919	0.914	0.865	0.889
			Average	0.899	0.907	0.903	0.94	0.93	0.938	0.90	0.913	0.907

جدول 8 : مقارنة نتائج المصنف CRF مع كل من النماذج

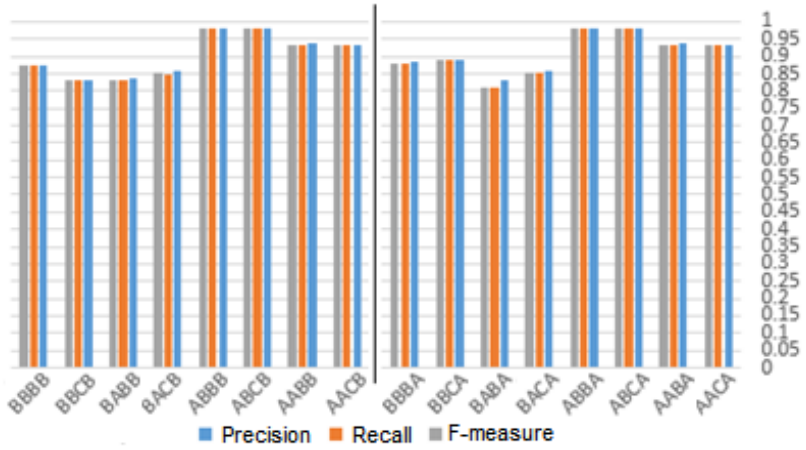
بالعودة إلى نتائج المصنفات عند تدريبها واختبارها على أشعة الواصفات التي تأخذ كل كلمات المدونة كسمات رئيسية (فقرة 6-1)، نستنتج أن تدريب المصنفات على أمثلة التدريب بعد تحسين شعاع الواصفات قد حسّن من نتائج التصنيف بـ 60-70% مما كانت عليه.

6-2-1- أثر نمط الأساس N-gram

نقوم هنا بدراسة أثر الشكل المضاف "1-gram + 2-gram" لكلمات أمثلة التدريب على نتائج التصنيف

بالمقارنة بين نتائج التصنيف عند التعامل مع الأساس 1-gram مع نتائج التصنيف عند التعامل مع الأساس 2-gram نستنتج وفقاً للجدول 9 أن التعامل مع الأساس 2-gram قد حسن وسطي النتائج بـ 0،5% أكثر ما كانت عليه عند التعامل مع الاساس 1-gram. أما من حيث النتائج المطلقة فإن نتائج أفضل حالة من 1-gram هي نفسها نتائج أفضل حالة من 2-gram.

يوضح الشكل 42 هذه النتائج ويستعرضها.



الشكل 42: استعراض نتائج المصنف CRF عند اختلاف الأساس

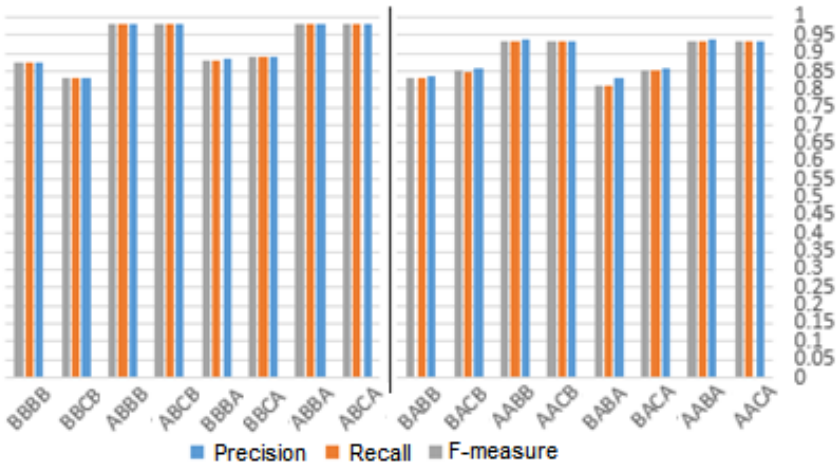
S1	S2	S3	S4		Average P	Average R	Average F
A	A	C	**A**	CRF	0.934	0.931	0.932
A	A	B	**A**	CRF	0.937	0.935	0.935
A	B	C	**A**	CRF	**0.981**	**0.98**	**0.98**
A	B	B	**A**	CRF	**0.981**	**0.98**	**0.98**
B	A	C	**A**	CRF	0.856	0.854	0.854
B	A	B	**A**	CRF	0.834	0.809	0.81
B	B	C	**A**	CRF	0.892	0.889	0.889
B	B	B	**A**	CRF	0.883	0.881	0.881
				Average	0.912	0.907	0.907
A	A	C	**B**	CRF	0.934	0.931	0.932
A	A	B	**B**	CRF	0.937	0.935	0.935
A	B	C	**B**	CRF	**0.981**	**0.98**	**0.98**
A	B	B	**B**	CRF	**0.981**	**0.98**	**0.98**
B	A	C	**B**	CRF	0.856	0.85	0.851
B	A	B	**B**	CRF	0.838	0.834	0.834
B	B	C	**B**	CRF	0.834	0.832	0.832
B	B	B	**B**	CRF	0.875	0.874	0.874
				Average	0.904	0.902	0.902

جدول 9 : جدول نتائج المصنف CRF عند اختلاف الأساس

6-2-2- أثر نموذج النفي

في هذه الفقرة نقوم بدراسة أثر اسلوب التعامل مع حالات النفي الواردة في التغريدة على نتائج التصنيف (الفقرة 3-1-2-5)

الجدول 10 يوضح أن التعامل مع حالات النفي من خلال إضافة واصفة خاصة ضمن شعاع الواصفات تشير إلى ورود أداة للنفي ضمن التغريدة، قد حسن وسطي نتائج المصنفات بـ 4% مما كانت عليه عند التعامل مع حالات النفي من خلال عكس الدلالة العاطفية للكلمة التي تتأثر بأدوات النفي ضمن التغريدة. يستعرض الشكل 43 هذه النتائج ويوضح ذلك.



الشكل 43: مقارنة نتائج المصنف CRF عند اختلاف نموذج التعامل مع حالة النفي

S1	S2	S3	S4		Average P	Average R	Average F
A	**A**	C	A	CRF	0.934	0.931	0.932
A	**A**	B	A	CRF	**0.937**	**0.935**	**0.935**
B	**A**	C	A	CRF	0.856	0.854	0.854
B	**A**	B	A	CRF	0.834	0.809	0.81
A	**A**	C	B	CRF	0.934	0.931	0.932
A	**A**	B	B	CRF	**0.937**	**0.935**	**0.935**
B	**A**	C	B	CRF	0.856	0.85	0.851
B	**A**	B	B	CRF	0.838	0.834	0.834
				Average	0.890	0.884	0.885
A	**B**	C	A	CRF	**0.981**	**0.98**	**0.98**
A	**B**	B	A	CRF	**0.981**	**0.98**	**0.98**
B	**B**	C	A	CRF	0.892	0.889	0.889
B	**B**	B	A	CRF	0.883	0.881	0.881
A	**B**	C	B	CRF	**0.981**	**0.98**	**0.98**
A	**B**	B	B	CRF	0.981	0.98	0.98
B	**B**	C	B	CRF	0.834	0.832	0.832
B	**B**	B	B	CRF	0.875	0.874	0.874

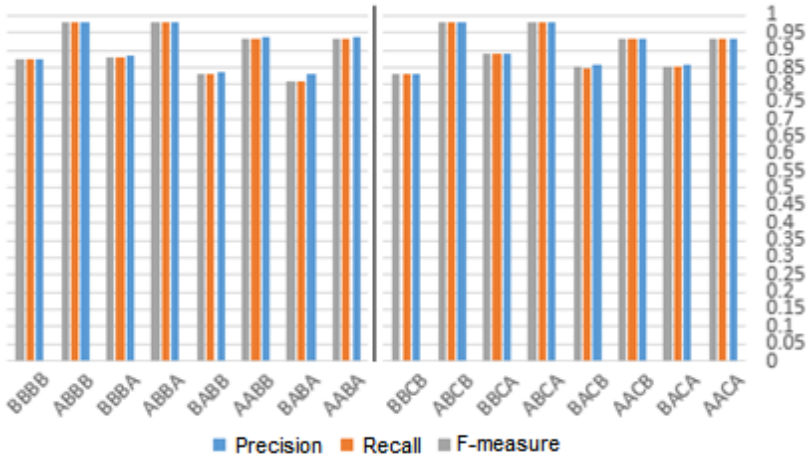
Average	0.926	0.924	0.924

جدول 10: مقارنة نتائج المصنف CRF عند اختلاف نموذج التعامل مع حالة النفي

6-2-3- أثر الاستعانة بقاموس سيف

نقوم في هذه الفقرة بدراسة تأثير الاستعانة بقاموس سيف في عملية البحث عن أثر الكلمات ضمن الأصناف العاطفية إلى جانب القواميس المبنية من المدونة.

نلاحظ في الجدول 11 أن الاستعانة مع قاموس سيف، قد أدى إلى تحسّن وسطي النتائج بـ 0.3% أكثر مما كانت عليه عند الإقتصار على استخدام القواميس المبنية من تغريدات أمثلة التدريب في عملية البحث عن أثر الكلمات على الأصناف العاطفية. يستعرض الشكل 44 هذه النتائج ويوضح ذلك.



الشكل 44: مقارنة نتائج دقة المصنف CRF عند الاستعانة بقاموس سيف إلى جانب القواميس المبنية من المدونة

S1	S2	S3	S4		Average P	Average R	Average F
A	A	**C**	A	CRF	0.934	0.931	0.932
B	A	**C**	A	CRF	0.856	0.854	0.854
A	A	**C**	B	CRF	0.934	0.931	0.932
B	A	**C**	B	CRF	0.856	0.85	0.851
A	B	**C**	A	CRF	**0.981**	**0.98**	**0.98**
B	B	**C**	A	CRF	0.892	0.889	0.889
A	B	**C**	B	CRF	0.981	0.98	0.98
B	B	**C**	B	CRF	0.834	0.832	0.832
				Average	0.908	0.905	0.906
A	A	B	A	CRF	0.937	0.935	0.935
B	A	B	A	CRF	0.834	0.809	0.81
A	A	B	B	CRF	0.937	0.935	0.935
B	A	B	B	CRF	0.838	0.834	0.834
A	B	B	A	CRF	**0.981**	**0.98**	**0.98**
B	B	B	A	CRF	0.883	0.881	0.881

A	B	B	B	CRF	0.981	0.98	0.98
B	B	B	B	CRF	0.875	0.874	0.874
				Average	0.908	0.903	0.903

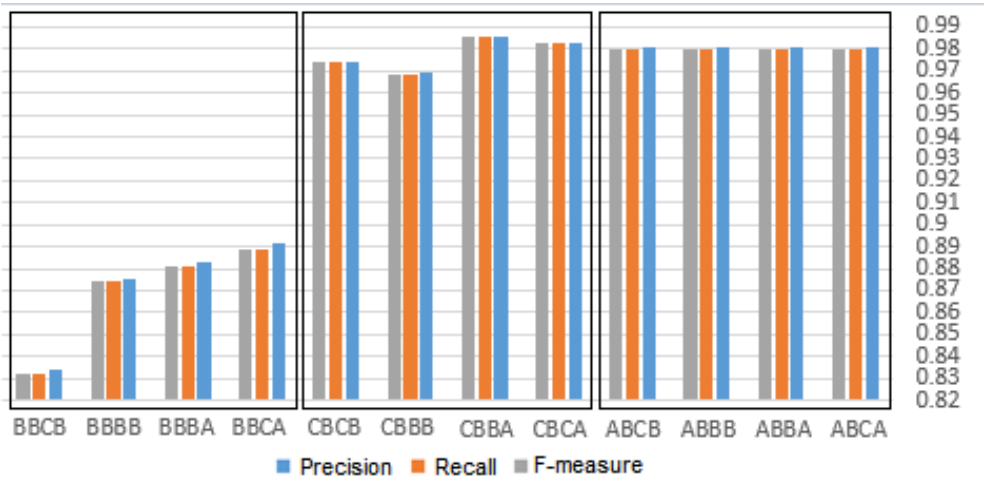
جدول 11: مقارنة نتائج دقة المصنف CRF عند الاستعانة بقاموس سيف إلى جانب القواميس المبنية من المدونة

6-2-4- أثر نموذج التوزين على دقة المصنفات

نقوم هنا بدراسة أثر طريقة احتساب أوزان الكلمات ضمن الأصناف العاطفية (4-3-d) على نتائج التصنيف.
بالمقارنة بين نتائج المصنِفات مع اعتماد نموذجي التوزين (Modified TF-IDF) & (TF-IDF) وبمقارنة النتائج مع دقة التصنيف عند استخدام نموذج التوزين (Weighed-TwF) نلاحظ النتائج التالية – وفقاً للجدول 12:

1- استخدام نموذج التوزين TF-IDF في عملية إيجاد قيمة تأثير كل كلمة من كلمات المدونة ضمن كل صنف عاطفي، قد حسّن من وسطي نتائج التصنيف بـ 10-11% مما كانت عليه عند اعتماد النموذج Modified TF-IDF.
2- استخدام نموذج التوزين TF-IDF في عملية إيجاد قيمة تأثير كل كلمة من كلمات المدونة ضمن كل صنف عاطفي، قد حسّن من وسطي نتائج التصنيف بـ 0.3% مما كانت عليه عند اعتماد النموذج Weighted-TwF.

يستعرض الشكل 45 هذه النتائج.

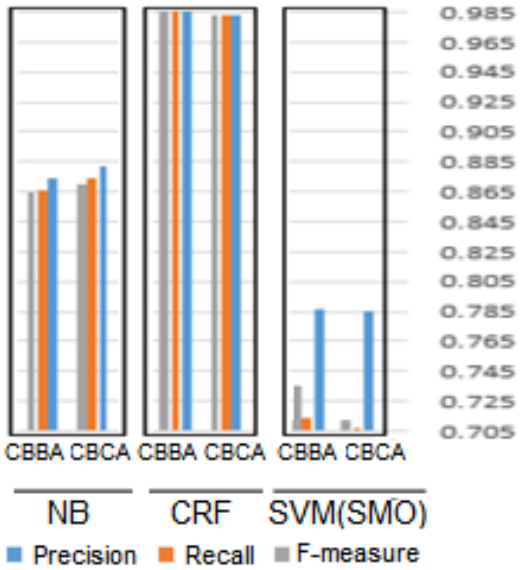


الشكل 45: مقارنة نتائج المصنف CRF عند اختلاف نموذج التوزين

				Average		
S1	S2	S3	S4	P	R	F
A	B	C	A	0.981	0.98	0.98
A	B	B	A	0.981	0.98	0.98
A	B	B	B	0.981	0.98	0.98
A	B	C	B	0.981	0.98	0.98
		Average		0.981	0.98	0.98
C	B	C	A	**0.983**	**0.983**	**0.983**
C	B	B	A	**0.986**	**0.986**	**0.986**
C	B	B	B	0.969	0.968	0.968
C	B	C	B	0.974	0.974	0.974
		Average		0.978	0.977	0.977
B	B	C	A	0.892	0.889	0.889
B	B	B	A	0.883	0.881	0.881
B	B	B	B	0.875	0.874	0.874
B	B	C	B	0.834	0.832	0.832
		Average		0.871	0.869	0.869

جدول 12: مقارنة نتائج المصنف CRF عند اختلاف نموذج التوزين

فيما يلي مقارنة لأداء لنتائج مجموعة من المصنفات على أفضل نموذجين (CBBA،CBCA **C**) تستخدم طريقة التوزين TF-IDF. الشكل 46. نلاحظ أن المصنف CRF أظهر أفضل نتائج.



الشكل 46: مقارنة نتائج المصنفات SMO، CRF، NB مع أفضل ثلاث نماذج عند اعتماد نموذج التوزين TF-IDF

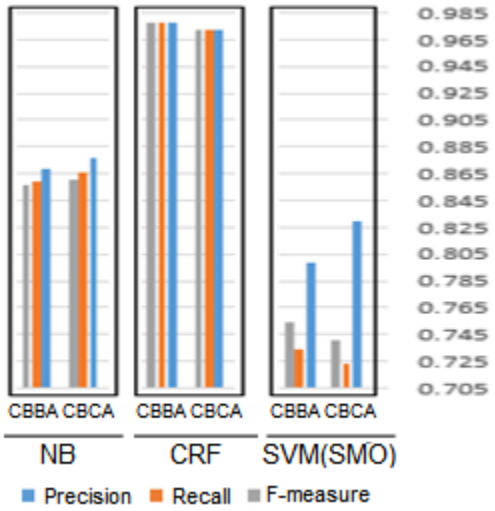
S1	S2	S3	S4		Average P	Average R	Average F
C	B	C	A	SMO	0.785	0.707	0.712
C	B	B	A	SMO	0.787	0.714	0.735
C	B	C	A	CRF	0.983	0.983	0.983
C	B	B	A	CRF	**0.986**	**0.986**	**0.986**
C	B	C	A	NB	0.882	0.874	0.87
C	B	B	A	NB	0.874	0.866	0.864
				Average	0.882	0.855	0.858

جدول 13 : مقارنة نتائج مصنفات SMO، CRF، NB مع أفضل ثلاث نماذج عند اعتماد نموذج التوزين TF-IDF

6-2-5- أثر اضافة سمات خاصة بالسب والتعجب والاستفهام ضمن شعاع الواصفات

نقوم باستعراض أثر كلمات السب وعلامات اللإستفهام والتعجب على دقة التصنيف، فبعد تعديل شعاع الواصفات ليأخذ بعين الإعتبار كل من كلمات السب وعلامات التعجب والإستفهام التي قد ترد ضمن التغريدة وذلك من خلال إضافة واصفات خاصة بكل منها "الجدول 14". وبالمقارنة مع النتائج "الجدول 13" التي تشير إلى دقة المصنِفات قبل إضافة سمات خاصة بالسب والتعجب والاستفهام، نستنتج أن هذه الإضافة قد حسنت من وسطي نتائج المصنفات بـ 0.2%

يستعرض الشكل 47 هذه النتائج.



الشكل 47: مقارنة نتائج المصنفات SMO، CRF، NB مع أفضل نموذجين وذلك بعد تمثيل حلات السب، التعجب والاستفهام ك سمات ضمن شعاع الواصفات

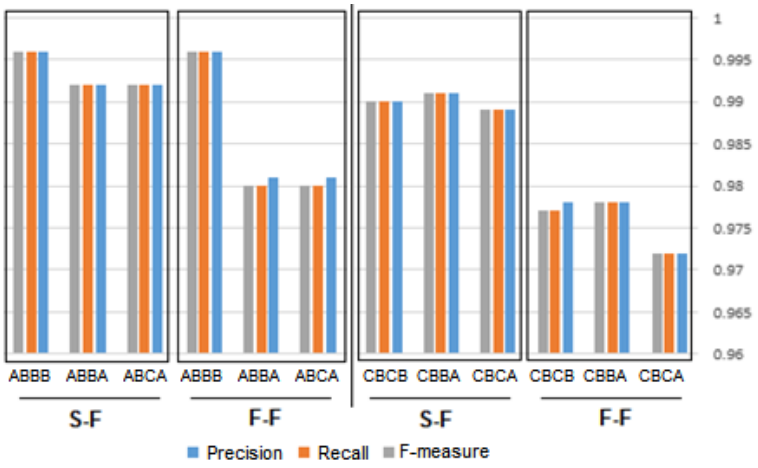
S1	S2	S3	S4		Average P	R	F
C	B	C	A	SMO	0.83	0.723	0.741
C	B	B	A	SMO	0.799	0.734	0.754
C	B	C	A	CRF	0.972	0.972	0.972
C	B	B	A	CRF	**0.978**	**0.978**	**0.978**
C	B	C	A	NB	0.876	0.866	0.86
C	B	B	A	NB	0.869	0.859	0.857
				Average	0.887	0.8553	0.860

جدول 14 مقارنة اداء مصنفات SMO، CRF، NB مع أفضل نموذجين وذلك بعد تمثيل حلات السب، التعجب والاستفهام ك سمات ضمن شعاع الواصفات

6-2-6- أثر التعامل مع شكل الكلمة

في هذه الفقرة نستعرض تأثير أسلوب التعامل مع شكل الكلمات على نتائج التصنيف، فبعد إضافة السمات الخاصة بالسب والتعجب والاستفهام ضمن شعاع الواصفات. نستعرض نتائج دقة التصنيف لأفضل ثلاث نماذج وذلك بهدف المقارنة بين حالة التعامل مع الشكل الكامل للكلمة Full Form word(F-F) أو مع الجذع الخاص بها Stem form word(S-F).

نستنتج ان التعامل مع الشكل الكامل للكلمة Full Form Word ضمن عملية بناء أشعة الواصفات قد حسّن وسطي النتائج بـ 0.8-1.5% مما كانت عليه عند التعامل مع الشكل المجذع للكلمة.



الشكل 48: مقارنة نتائج المصنف CRF مع أفضل ثلاث نماذج وذلك عند التعامل مرة مع الكلمة بشكلها الكامل F-F ومرة عند التعامل مع الكلمة بشكلها المجذع S-F

S1	S2	S3	S4		Class	Average P	Average R	Average F
C	B	C	A	S-F	CRF	0.972	0.972	0.972
C	B	B	A	S-F	CRF	0.978	0.978	0.978
C	B	C	B	S-F	CRF	0.978	0.977	0.977
					Average	0.976	0.975	0.975
C	B	C	A	F-F	CRF	0.989	0.989	0.989
C	B	B	A	F-F	CRF	0.991	0.991	0.991
C	B	C	B	F-F	CRF	0.99	0.99	0.99
					Average	0.99	0.99	0.99
A	B	C	A	S-F	CRF	0.981	0.98	0.98
A	B	B	A	S-F	CRF	0.981	0.98	0.98
A	B	B	B	S-F	CRF	**0.996**	**0.996**	**0.996**
					Average	0.986	0.985	0.985
A	B	C	A	F-F	CRF	0.992	0.992	0.992
A	B	B	A	F-F	CRF	0.992	0.992	0.992
A	B	B	B	F-F	CRF	0.996	0.996	0.996
					Average	0.993	0.993	0.993

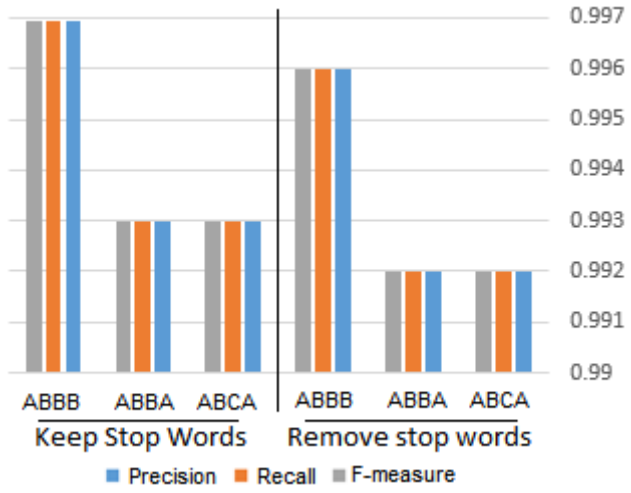
جدول 15: مقارنة نتائج المصنف CRF مع أفضل ثلاث نماذج وذلك عند التعامل مرة مع الكلمة بشكلها الكامل F-F ومرة عند التعامل مع الكلمة بشكلها المجذعS-F

6-2-7- أثر الاحتفاظ بـ كلمات التوقف Stop words

نقوم هنا بدارسة تأثير كلمات التوقف على نتائج التصنيف من خلال مقارنة النتائج مرةً من خلال التصنيف بعد حذف كلمات التوقف من أمثلة التدريب ومرةً بعد الإبقاء عليها.

فبعد إضافة السمات الخاصة بالسب والتعجب والاستفهام ضمن شعاع الواصفات. نستعرض نتائج دقة التصنيف لأفضل ثلاث نماذج وذلك بهدف بالمقارنة بين حالة حذف كلمات التوقف Stop words أو الإبقاء عليها.

نلاحظ في الجدول 16 أن الإبقاء على كلمات التوقف قد حسَّن من وسطي نتائج التصنيف بمقدار 0.1% أكثر مما كانت عليه عندما حذفنا كلمات التوقف من أمثلة التدريب. يستعرض الشكل 49 هذه النتائج.



الشكل 49: مقارنة نتائج المصنف CRF عند اختلاف نموذج التعامل مع كلمات التوقف

S1	S2	S3	S4		Class	Average P	R	F
A	B	C	A	Remove SW	CRF	0.992	0.992	0.992
A	B	B	A	Remove SW	CRF	0.992	0.992	0.992
A	B	B	B	Remove SW	CRF	**0.996**	**0.996**	**0.996**
					Average	0.993	0.993	0.993
A	B	C	A	Keep SW	CRF	0.993	0.993	0.993
A	B	B	A	Keep SW	CRF	0.993	0.993	0.993
A	B	B	B	Keep SW	CRF	0.993	0.997	0.993
					Average	0.993	0.994	0.993

جدول 16: مقارنة نتائج المصنف CRF عند اختلاف نموذج التعامل مع كلمات التوقف

6-2-8- تعميم نموذج التدريب

وهكذا، عند اختبار المصنف CRF الذي تم تدريبه على مدونة التدريب التي تم تمثيلها شعاعيا باستخدام النموذج ABCB "والذي ثبت أنه أفضل النماذج وفقا للنتائج السابقة" وعند اختبار هذا المصنف على مدونة الإختبار "مدونة تحتوي 280 تغريدة موزعه ضمن ستة أصناف عاطفية، هذه التغريدات تختلف تماماً عن التغريدات التي تم تدريب المصنِف عليها والتي تم بناء القواميس انطلاقاً منها"، نلاحظ الدقة التالية:

	Precision	**Recall**	**F-Measure**	Class
	0.661	0.804	0.726	**Sadness**
	0.679	0.809	0.738	**Joy**
	0.828	0.462	0.593	**Surprise**
	0.695	0.837	0.759	**Disgust**
	0.484	0.596	0.534	**Fear**
	0.862	0.521	0.649	**Anger**
Avg.	0.7	0.669	0.664	

جدول 17: نتائج تعميم المصنف CRF

نلاحظ أن المصنِف يتفوق في علمية تنبؤ الصنف العاطفي "حزن" منه على الصنف العاطفي "تفاجؤ" وذلك يرتبط بمدى احتواء القاموس العاطفي الموافق للصنف العاطفي على كلمات مميزة للشعور، فكلما كانت كلمات القاموس العاطفي أكثر تمييزاً للعاطفة كلما كانت نتائج التنبؤ المرتبط بهذا الصنف أكثر دقة.

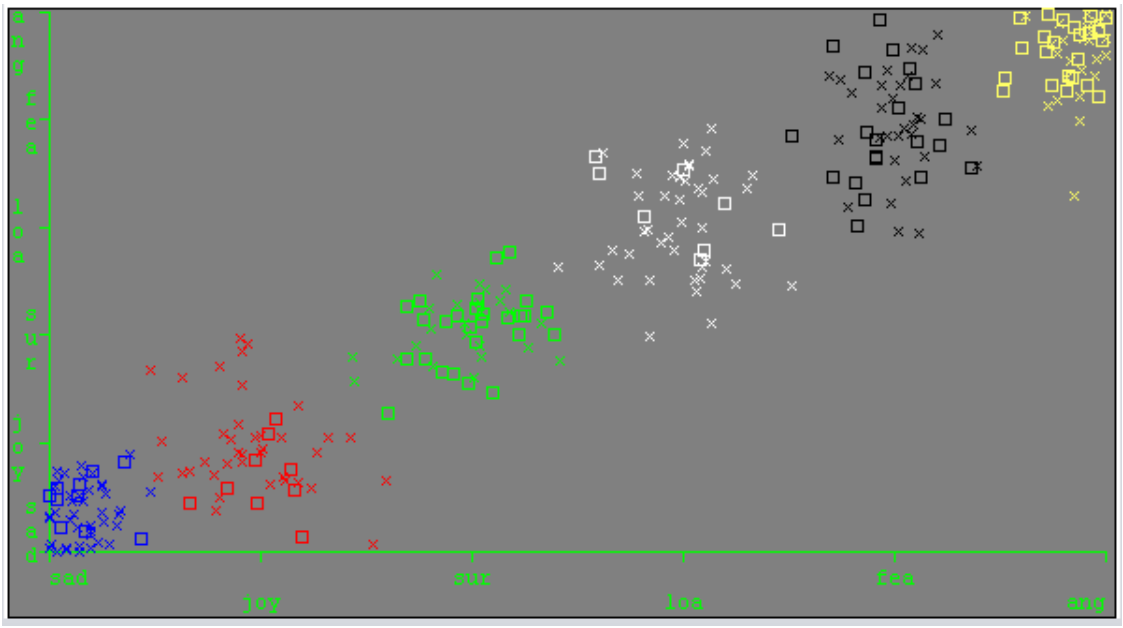
على سبيل المثال: نلاحظ أن كلمة "تعيس" من القاموس العاطفي الذي يشير الى الحزن، أكثر تمييزاً للحزن من الكلمة "القتل" التي تنتمي الى نفس القاموس – يستعرض الجدول 18 بعض هذه الكلمات.

	حزن	**فرح**	**تفاجؤ**	**اشمئزاز**	**خوف**	**غضب**
تعيس	1031515	0	0	0	0	0
القتل	873018.8	0	0	0	873018.8	0
دمعه	873018.8	873018.8	0	0	0	0
حرام	773018.7	0	0	773018.7	0	0
كريه	0	0	0	1031515	0	0
نجاح	773018.7	773018.7	0	0	773018.7	0

الجدول 18: استعراض مدى تمييز بعض الكلمات لصنف العاطفي

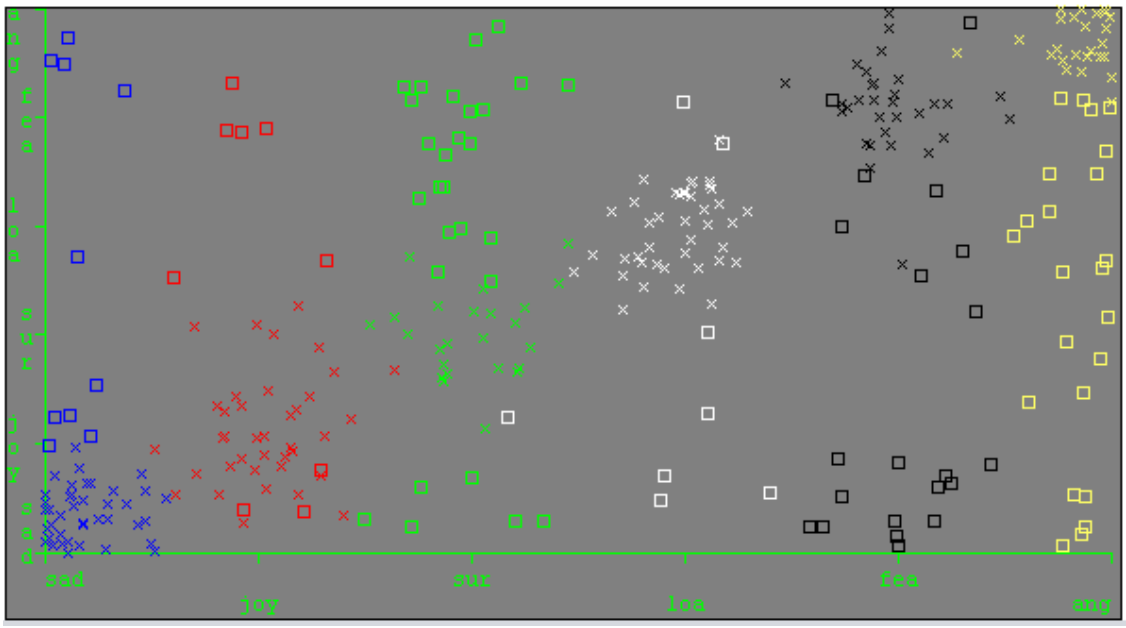
بهدف إجراء توضيح بياني لدقة المصنِف CRF نستعرض الشكلين التاليين:

- يعرض الشكل 50 تمثيلا بيانياً لـ "أشعة بيانات التدريب التي تمثل المدونة المنمّطة بالمشاعر وفقا للنموذج ABCB" ضمن فضاء ثنائي البعد.



الشكل 50: تمثيل "أشعة أمثلة التدريب – نموذج ABCB" ضمن فضاء ثنائي البعد

- أما الشكل 51 فإنه يعرض تمثيلا بيانياً لـ "نتائج خوارزمية التصنيف CRF على أشعة بيانات التدريب التي تمثل المدونة المنمّطة بالمشاعر وفقا للنموذج ABCB " ضمن فضاء ثنائي البعد



الشكل *51*: تمثيل "دقة المصنف CRF على أشعة أمثلة التدريب – نموذج ACBC" ضمن فضاء ثنائي البعد

الخاتمة وآفاق مستقبلية

في هذا البحث استخدمنا ما يزيد عن 1320 تعليق ومنشور تم جمعها بشكل آلي من موقع التواصل الإجتماعي تويتر، وتم تصنيفها يدوياً وذلك لبناء مدونة نصية منمّطة بالمشاعر تتعامل مع ستة عواطف "الحزن، الفرح، التفاجؤ، الإشمئزاز، الخوف والغضب". ثم عملنا على تمثيل هذه البيانات "التغريدات" ضمن فضاء شعاعي كل تغريده من هذه المدونة تم تمثيلها كشعاع بيانات ذي بنية رقمية تشير بمدلولها إلى نفس الصنف العاطفي الذي كان يشار اليه ضمن النموذج النصي لهذه التغريدة، خلال ذلك حاولنا الإستفادة من البنية النحوية للكلمة مثل محاولة العمل على الكلمة مرةً بشكلها الكامل full form ومرةً أخرى بشكلها المجذع Stem form، ثم عملنا إلى تهجين أفضل الشكلين مع النهج N-grams في عملية تمثيل بيانات التدريب. بالإضافة إلى الإستفادة من الخصائص الإسلوبية للكلمات ضمن بيانات التدريب والبحث عن أفضل نموذج رياضي لالتقاط أوزان الكلمات ودلالاتها ضمن الأصناف العاطفية بهدف تمثيل كل من هذه البيانات "التغريدات" كشعاع ضمن فضاء شعاعي Vector Space Model بعد ذلك قمنا بتدريب مجموعة من المصنفات على بيانات التدريب بشكلها الشعاعي وانتهينا إلى اختبار نتائج التدريب باستخدام المعايير Precision and Recall and F-measure" وحققنا دقة F-measure 66.9%.

بعد تحقيقنا للأهداف المرجوة من المشروع والتي جئنا على ذكرها في سياق التقرير:

- يمكن تطوير برنامج بقوم بإصدار تقارير يومية بنسبة المنشورات والتعليقات ذات الدلالات العاطفية المتنوعة والتي تنشر على مواقع التواصل الاجتماعي ضمن الصفحات الوطنية والتي تحاكي حياة المواطن السوري وهمومة مما يساعد في معرفة أي القرارات أو القوانين التي تثير استياء المواطن السوري أو تلاقي ترحيبه، وما هي محفزات ومثبطات العمل لديه، وهذا بالتأكيد ينعكس على صوابية اتخاذ القرارات المستقبلية. فعالم اليوم هو بلا منازع عالم التواصل الاجتماعي على الشبكة العنكبوتية، وكل من يرغب في البقاء قريباً من مواطنية أو زبائنه وعملائه هو بحاجة لأن يكون على تماس مباشر معهم، والتنافس اليوم هو على إرضاء الجمهور وكسب أكبر شريحة من المؤيدين سواء على الصعيد السياسي أو التجاري أو الإجتماعي أو حتى الأكاديمي.
- يمكننا مستقبلاً العمل توسيع حجم بيانات التدريب من خلال الحصول على المزيد من التغريدات ذات الدلالة العاطفية باللهجة السورية، والتوسع بالفئات العاطفية المدروسة مثل "الحب والندم".
- كما يمكن أيضا توسيع المعاجم الخاصة لدينا (كلمات الشكر والكلمات المستخدمة في الحالة الجدلية، والتعابير...)

في النهاية نأمل أن نكون قد وفقنا في هذا العمل لتقديم مشروع بحث أكاديمي متكامل يضاف إلى مكتبة الجامعة الافتراضية السورية ليثري مواضيعها في مجال يعتبر من أكثر مجالات البحث نشاطاً واهتماماً في هذه الأيام، وتزداد أهميته مع الدور المتزايد الذي تلعبه شبكات التواصل الإجتماعي في حياتنا اليومية. وبنشرنا لخلاصة العمل وأدواته ضمن إحدى المجلات المحكّمة المختصة بمجال البحث العلمي نأمل بأن نرتقي باسم جامعتنا وأن يشار إلى أبحاثها العلمية في دراسات العديد من الباحثين.

References

1. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.

10. Rabie, O., & Sturm, C. (2014). Feel the heat: Emotion detection in Arabic social media content. In the International Conference on Data Mining, Internet Computing, and Big Data (BigData2014) (pp. 37-49). The Society of Digital Information and Wireless Com.

11. Calvo, R. A., & Mac Kim, S. (2013). Emotions in text: dimensional and categorical models. Computational Intelligence, 29(3), 527-543.

12. Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Transactions on Information Systems (TOIS), 26(3), 12.

13. Al-Shalabi, R., & Obeidat, R. (2008, March). Improving KNN Arabic text classification with n-grams based document indexing. In Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt (pp. 108-112).

14. Abd El Salam, A. L., HAJJAR, M., & ZREIK, K. Classification of Arabic Information Extraction methods.

15. Danisman, T., & Alpkocak, A. (2008, April). Feeler: Emotion classification of text using vector space model. In AISB 2008 Convention Communication, Interaction and Social Intelligence (Vol. 1, p. 53).

16. Do, H. J., & Choi, H. J. (2015, October). Korean Twitter Emotion Classification Using Automatically Built Emotion Lexicons and Fine-Grained Features. In PACLIC.

17. Davidov, D., Tsur, O., & Rappoport, A. (2010, August). Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of the 23rd international conference on computational linguistics: posters (pp. 241-249). Association for Computational L.

18. Barbosa, L., & Feng, J. (2010, August). Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 36-44). Association for Computational Linguistics.

19. Elhawary, M., & Elfeky, M. (2010, December). Mining Arabic business reviews. In Data Mining Workshops (ICDMW), 2010 IEEE International Conference on (pp. 1108-1113). IEEE.

2. P. Ekman, “Facial expression and emotion,”American Psy- chologist, vol. 48, no. 4, pp. 384–392, 1993.

20. Canales, L., & Martínez-Barco, P. (2014). Emotion Detection from text: A Survey. Processing in the 5th Information Systems Research Working Days (JISIC 2014), 37.

21. Shivhare, S. N., & Khethawat, S. (2012). Emotion detection from text. arXiv preprint arXiv:1205.4944.

22. Strapparava, C., & Valitutti, A. (2004, May). WordNet Affect: an Affective Extension of WordNet. In LREC (Vol. 4, pp. 1083-1086).

23. Qadir, A., & Riloff, E. (2014). Learning Emotion Indicators from Tweets: Hashtags, Hashtag Patterns, and Phrases. In EMNLP (pp. 1203-1209).

24. Suin Kim, JinYeong Bak, and Alice Oh. 2012.Discovering emotion influence patterns in online social network conversations. SIGWEB Newsl., (Autumn):3:1–3:6, September.

25. Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).

26. 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).

27. Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., ... & Roth, R. (2014, May). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In LREC (Vol. 14, pp. 1094-1101).

28. NRC Hashtag Emotion Lexicon - Version 0.2 - 2013. Saif Mohammad (saif.mohammad@nrc-cnrc.gc.ca). http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

29. Danisman, T., & Alpkocak, A. (2008, April). Feeler: Emotion classification of text using vector space model. In AISB 2008 Convention Communication, Interaction and Social Intelligence (Vol. 1, p. 53).

3. Parrott, W.G, “Emotions in Social Psychology,” in Psychology Press, Philadelphia 2001 .

30. Emotion Detection and Recognition Market by Technology (Bio-Sensor, NLP, Machine Learning), Software Tool (Facial Expression, Voice Recognition), Service, Application Area, End User, and Region - Global Forecast to 2021.

31. Twitter Data set for Arabic Sentiment Analysis. April, 2014. (a) Creator: N. A. Abdulla and N. Mahyoub (b) Donor: N. A. Abdulla (naabdulla11@cit.just.edu.jo) “collection of 2000 labelled tweets (positive tweets and negative ones) These tweets written in b.

32. AraSenti Lexicon. May, 2016 (a) Nour AlTwairesh (Twairesh@ksu.edu.sa) (b) Hend Alkhaifa (hendk@ksu.edu.sa) (c) abdulMalik Alsalman (salman@ksu.edu.sa) “All resources created and used in Arabic Sentiment Analysis of Arabic Tweets. Includes Sent.

33. Seol, Y. S., Kim, D. J., & Kim, H. W. (2008, July). Emotion recognition from text using knowledge-based ANN. In ITC-CSCC: International Technical Conference on Circuits Systems, Computers and Communications (pp. 1569-1572).

34. Hajar, M. (2016). Using YouTube Comments for Text-based Emotion Recognition. Procedia Computer Science, 83, 292-299.

35. Shaheen, S., El-Hajj, W., Hajj, H., & Elbassuoni, S. (2014, December). Emotion recognition from text based on automatically generated rules. In Data Mining Workshop (ICDMW), 2014 IEEE International Conference on (pp. 383-392). IEEE.

36. K. Toutanova, Klein D., Manning C., Singer Y., StanfordPOStagger, [Online].Available:http://nlp.stanford.edu/software/tagger.shtml,Stanford, 2003.

37. Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D.Manning. "Generating typed dependency parses from phrase structure parses." Proceedings of LREC. Vol. 6. 2006.

4. http://www.6seconds.org/2017/04/27/plutchiks-model-of-emotions/.

5. Donalek, C. (2011). Supervised and Unsupervised learning.

6. Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction (Vol. 1, No. 1). Cambridge: MIT press.

7. Aggarwal, C. C., & Zhai, C. (Eds.). (2012). Mining text data. Springer Science & Business Media.

8. Kamber, M., Han, J., & Pei, J. (2012). Data mining: Concepts and techniques. Elsevier.

9. Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian Joint Conference on Artificial Intelligence (pp. 1015-1021). Springer Berlin He.

Table of Figures