Title:	Documents Emotions Classification Model Based on TF-IDF Weighting Measure
Author(s):	Amr Mansour Mohsen Hesham Ahmed Hassan **Amira M. Idrees**
Journal/Conference:	ICDAR 2016 : 18th International Conference on Document Analysis and Recognition
Publication details:	Proceedings of the ICDAR 2016 https://www.waset.org/conference/2016/01/johannesburg/ICDAR/home
Publication Date:	January, 12-13, 2016
Publisher	World Academy of Science Engineering and Technology
Place	Johannesburg, South Africa

Paper Title:	**Documents Emotions Classification Model Based on TF-IDF Weighting Measure**
Main Domain:	Machine Learning – Text Mining
Sub-Domain:	Emotion Detection in Text
Problem:	Emotions classification of text documents is applied to reveal if the document expresses a determined emotion from its writer. The classification of situations can be among different emotions, some examples are [Anger, Disgust, Fear, Joy, and Sad] which come under two general categories, they are positive or negative. As blogs, reviews, and social networks play an important role for revealing the opinions and feelings considering many topics or products, therefore, analyzing these resources and extracting the opinions of their owners has become a vital field of research. An example of using blogs when a company such as "HTC" offered a new mobile edition in the market and needed to analyze the impact of its mobile on the users. As the analysis of hundreds blogs has been written about specific products or topics manually is too difficult, therefore, following a more productive approach for sentimental classification and opinion mining is required.
Context:	The research focuses on the classifications of emotions in short text documents such as blogs, reviews, and posts. The previous work on detecting emotions is classified to three main categories which are 1) Information retrieval techniques; 2) Lexicon based techniques, 3) Machine Learning Techniques. in this paper, we discuss an approach that aims to integrate the previously mentioned techniques targeting enhancing the classification accuracy. The research included applying the pre-processing for the dataset, pre-processing included stemming, removing stop words, and preparing the training and testing data. Then, the attributes for classification has been determined, applying the proposed integrated technique is then performed and compared with other research that are applied using the same dataset to prove the originality of the proposed approach.
Solution approach:	The proposed solution started by detecting the dataset ISEAR which will be used in the research. ISEAR dataset consists of about 7666 sentences examples which are classified by different emotions (anger, disgust, fear, joy, sadness, shame and guilt). Preprocessing on the dataset is performed such as stop words removal and part of speech tagging. Then the data set

	is prepared for the classification approach by preparing the training and testing data. The first phase of the classification approach is lexicon enrichment, in this research, NRC lexicon is used which is enriched using regular expressions technique for extracting more keywords that belongs to one of the current emotion classes. the second step is to compute the term frequency index document frequency (TF-IDF) for all terms in the training set so we have two dimensional matrix as training examples and features of the training set. Each cell in the matrix represents Tf-Idf value. The third phase is the features preparation for learning. Seven features are considered, five of them represents the occurrences of the terms in the lexicon appeared in the training examples and two senti words values. For each sentence, the weight of each word in the sentence is calculated by multiplying its occurrences in each emotion with the Tf-idf measure for this term. This calculation determines the first five attributes, with the senti words' attributes that are calculated by the use of training examples computing in the SentiWordNet lexicon. Two more attributes are calculated by determining the positive and negative sentiment values for each term in the document. The final phase is using different machine learning techniques for evaluating the proposed approach to prove its applicability to provide more accurate emotion classification.
Contribution:	This research successfully developed a sentimental classification approach base on three main techniques, they are 1) Information retrieval techniques; 2) Lexicon based techniques, 3) Machine Learning Techniques The research contribution is summarized as follows: 1- presenting an integrated approach for sentimental analysis and classification 2- Comparing the proposed work with other previous work using the same dataset. 3- The proposed approach proved the enhancement in the precision, recall and f-measure results by applying different experiments in different directions. 4- The proposed approach proved its advancement on other recent work that has been proposed by researchers.

إسم البحث	**نموذج لتصنيف الأنفعالات فى الوثائق مبنى على مقياس التكرارات الترجيحى**
ملخص المشكلة	تتلخص فكرة البحث فى القدرة على إستنتاج و تصنيف الإنفعالات فى النص المكتوب بدقة عالية للمواقف المختلفه. هذا النص من الممكن أن يكون مكتوبا في مدونات أو مكتوب في مواقع التواصل الإجتماعي وتعود أهمية هذا البحث للتأثير القوى الذى تحدثة الآراء المختلفه التى يعلنها أصحابها على هذه المواقع والتى يمكن أن تؤثر على آراء الآخرين وهذا يؤكد التأثير القوى لهذه الآراء على المنتجات والشركات والخدمات المختلفه. كمثال عندما يقول شخص "لقد أسئت فهمي من أحد أصدقائي", فهنا نحتاج لتصنيف هذا الموقف وتحديد الإنفعال الذى تتضمنه هذه الجملة طبقا للإنفعالات المختلفه مثل (الغضب والاشمئزاز والخوف والفرح والحزن والخجل والشعور بالذنب..)
سياق البحث	يتضمن هذا البحث دراسة الأبحاث السابقة القائمة على تحليل الإنفعالات ووجد أنها تأتي تحت ثلاث أشياء أساسية 1) تقنيات استرجاع المعلومات , 2) مبني على المعاجم , 3) تقنيات التعلم الالي. تم فى هذا البحث العمل من خلال تلك الثلاث وسائل لتحليل الإنفعالات للوصول لأعلى درجة دقه فى تحديد الإنفعالات. تم أستخدام مجموعه المواقف المكتوبة كنص حر و المصنفة مسبقاً تنتمى لISEAR تتكون من 7666 موقف مصنفين طبقاً لسبع إنفعالات(الغضب والاشمئزاز والخوف والفرح والحزن والخجل والشعور بالذنب). وتم تجهيز هذه النصوص. كما تم تعديل معجم NRC والخاص بالإنفعالات. ثم تم المزج بين الإتجاهات المختلفه للوصول إلى درجة دقة عالية فى التصنيف وهى حساب تكراره الكلمات فى النص طبقاً ل TF_IDF وإستخدام تقنيات Machine Learning من أجل إصدار نموذج مدرَب من الحاسوب. وتم تطبيق الإتجاه المقترح ومقارنته بالأعمال السابقة للباحثين الدوليين الذين أتموا أبحاثهم بإستخدام نفس مجموعة المواقف وأثبتت النتائج أن البحث يقدم درجة دقة أعلى.
أسلوب البحث	الحل المقترح لتصنيف الإنفعالات في النص المكتوب يبدأ بإستخدام مجموعة من المواقف المكتوبة و المصنفة مسبقاً من أجل إختبار هذا المقترح. هذه المجموعة من البيانات تنتمي لISEAR تتكون من 7666 موقف مصنفين طبقاً لسبع إنفعالات(الغضب والاشمئزاز والخوف والفرح والحزن والخجل والشعور بالذنب). فى هذا البحث تم التركيز على المواقف المصنفة طبقاً لخمس إنفعالات و هم (الغضب والاشمئزاز والخوف والفرح والحزن). تم عمل تجهيزات لتلك النصوص مثل تقطيع الكلمات من الجمل و إزالة الكلمات التي ليس لها معنى و لم تؤثر كثيراً على الجملة و إزالة بعض علامات الترقيم و معرفة السياق النحوي لكل كلمة. تم إستخدام معجم NRC والخاص بالإنفعالات. تم إعادة تنظيم للمعجم وإثراءه ببعض الكلمات عن طريق التعرف على السياق و البحث به مرة أخرى في المواقف لأخذ كلمات جديدة تتعلق بهذا السياق. تم تقييم أهمية الكلمات طبقاً ل TF_IDF وحساب أهمية الكلمة طبقا للمقياس الخاص بها تمهيدا لإستخدامه فى خطوة التصنيف. ثم تم إستخدام تقنيات Machine Learning من أجل إصدار نموذج مدرَب من الحاسوب و إختبار مدى دقته على مجموعة أخرى من البيانات. تم تنفيذ عدة تجارب وعرض نتائجها والتى تؤكد قدرة النموذج المقدم فى البحث على التصنيف بدقة أعلى من الأبحاث السابقه.
النتائج المستخلصة	1- إستخدام معجم خاص بالإنفعلات و عمل تكيف للمعجم علي بعض الكلام في مواقف متغيرة والقدرة على إثرائه بكلمات جديدة مبنية على وجود بعض الكلمات التي تكون موجوده من قبل يحسن جودة قاموس الإنفعالات. 2- إستخدام تكرار الكلمات لمعرفة أهمية تلك الكلمات مع تحديد نسبة الترجيح في مجموعة الأمثلة التي يتدرب عليها الحاسوب ليستطيع تعلم و إختيار الإنفعال المناسب طبقا للكلام المكتوب. 3-تنفيذ عدة مقارنات مع بعض الأبحاث الأخرى والتى تؤكد زيادة نسبة الدقة و جودة إسترجاع البيانات الصحيحة.