Are Emoticons Good Enough to Train Emotion Classifiers of Arabic Tweets?

Conference Paper · July 2016

DOI: 10.1109/CSIT.2016.7549459

4 authors:

Wegdan A. Hussien
Jordan University of Science and Technology

Yahya Tashtoush
Jordan University of Science and Technology

Mahmoud Al-Ayyoub
Jordan University of Science and Technology

Mohammed N. Al-Kabi
Zarqa University

Are Emoticons Good Enough to Train Emotion Classifiers of Arabic Tweets?

Wegdan A. Hussien,* Yahya M. Tashtoush,* Mahmoud Al-Ayyoub* and Mohammed N. Al-Kabi†
* Jordan University of Science and Technology, Irbid, Jordan
Emails: wahussien13@cit.just.edu.jo,{ yahya-t, maalshbool }@just.edu.jo
† Zarqa University, Zarqa, Jordan
Email: malkabi@zu.edu.jo

***Abstract*—Nowadays, the automatic detection of emotions is employed by many applications across different fields like security informatics, e-learning, humor detection, targeted advertising, etc. Many of these applications focus on social media. In this study, we address the problem of emotion detection in Arabic tweets. We focus on the supervised approach for this problem where a classifier is trained on an already labeled dataset. Typically, such a training set is manually annotated, which is expensive and time consuming. We propose to use an automatic approach to annotate the training data based on using emojis, which are a new generation of emoticons. We show that such an approach produces classifiers that are more accurate than the ones trained on a manually annotated dataset. To achieve our goal, a dataset of emotional Arabic tweets is constructed, where the emotion classes under consideration are: anger, disgust, joy and sadness. Moreover, we consider two classifiers: Support Vector Machine (*SVM*) and Multinomial Naive Bayes (*MNB*). The results of the tests show that the automatic labeling approaches using SVM and MNB outperform manual labeling approaches.**

Keywords—Arabic emotion annotation; Arabic emotion analysis; emojis; sentiment analysis

I. INTRODUCTION

During the last two decades, the Internet evolution has led to a transition from the static web to a dynamic web, where the users can get information and share their opinions, ideas, knowledge and even media content [1]. Consequently, social networks emerged and many social services have grown rapidly and attracted millions of people in a short time. Hence, these social networks become an important part in people's life in which they communicate and share their daily events. Micro-blogging is a growing form of communication that can be noticed in many social services, where users are allowed to publish short messages. Twitter[1] is one of the most popular micro-blogging services that makes it a valuable source for many Natural Language Processing (NLP) studies in different languages, since it supports more than 35 natural languages. Twitter monthly active users exceed 350 million, 80% of these users live outside the US, and active on mobile according to Twitter's usage fact Web page. The popularity of Twitter is widely varied in the 18 Arab countries listed by Alexa.com during April, 2016. Twitter ranked fourth in Yemen, sixth in Oman and Saudi Arabia, seventh in Syria, and eighth in Kuwait, ninth in Bahrain, Libya, and Sudan, tenth in Qatar, twelfth in Iraq and United Arab Emirates, fifteenth in Jordan and Palestinian territories, nineteenth in Algeria and Egypt, twenty third in Tunis, and Twitter ranked twenty ninth in Morocco.

Sentiment Analysis (SA) studies are flourished during the last 15 years, and it constitutes an active important part of NLP studies. SA is used to determine whether the text is subjective (positive or negative), or objective which means fact (neutral). On the other hand, Emotion Analysis (EA) is concerned with detecting emotions from text depending on the underlying emotion framework used. Many frameworks exist but we will mention two of the most commonly used ones: the Ekman's [2] six basic emotions {Anger, Disgust, Fear, Joy, Sadness and Surprise}, and the Plutchik's [3] eight basic emotions which are Ekman's six emotions, anticipation and trust.

The field of EA is quite different from the popular field of Sentiment Analysis (SA), where the concern is to determine the polarity of the sentiment (usually, positive vs negative) conveyed in a certain text. One might view some emotions as sentiments on a finer-granularity. As an example, consider the 'sadness' and 'anger' emotions. People can easily consider them as negative sentiments, but clearly, they are different emotions. EA can thus be done as an additional layer on top of the (relatively) simpler SA. Another difference is related to how many classes each problem considers and whether a text excerpt can bear more than one class or not (multi-label classification). While SA can be binary, ternary or multi-way (e.g., a 5-star ranking system), EA is much more open to interpretation. Although only six or eight emotions are considered basic/primary emotions, the number of different emotions that can be considered by EA can be much larger. Dealing with larger number of classes increases the complexity of the problem. Moreover, the boundary between two emotions may not be clear. In fact, the same text excerpt might convey different emotions at the same time making it a multi-label text classification problem.[2] On the other hand, in the field of SA, the issue of conveying different sentiments is usually ignored (conflicting sentiments) or dealt with within the context of aspect-based sentiment analysis.[3]

In this study, we aim at showing that training a classifier to detect emotions on automatically annotated tweets (based on emojis) is actually better than training it on a manually annotated tweets. To do so, we collect emotional Arabic tweets. We focus on four basic emotions: anger, disgust, joy and sadness. Then we extract two training sets: one with the tweets containing emojis and one with randomly selected tweets to be

[1] http://www.twitter.com

[2] See [4, 5] for more information about multi-label text classification.

[3] http://goo.gl/l3V9NZ

annotated manually. We also extract a third set to be used for testing. The testing set is manually annotated. We train the same classifiers on the two training sets and test each of them on the same testing sets. The results show that the automatically trained classifiers are better than the human trained ones. This result is significant as it might change the way researchers approach problems like EA of social media content. It will also allow them to transition into much larger data sizes making their systems more practical and trustworthy.

This paper is organized as follows. Section II introduces some related works on EA. Section III describes the methodology followed to automatically labeled our data followed by the steps needed before classification. The experiments and results are presented and discussed in Section IV. Finally, Section V concludes our work.

II. RELATED WORKS

Most of the existing researches on EA and SA focused on the English language and the work in EA field started earlier for English [6]. Few studies handled the problems of EA and SA for Arabic language. Over the past few years, many works conducted on SA of Arabic posts. Surveys of [7, 8, 9] give very good summaries. On the other hand, many studies exploited emoticons as their annotation approach using data from different languages such English [10], Chinese [11] or Arabic [12].

One of the earliest works on EA of Arabic text [13], is conducted by El Gohary et al., where they constructed a model to detect emotions in Arabic children stories using a lexicon-based approach. The model they used is based on a moderate size lexicon to identify the emotional expressions at different levels (word, sentence and document). El Gohary et al. dataset consists of 100 documents (2514 sentences) in which they used 65 documents as training and the rest of 35 documents as testing. The training dataset is annotated manually using the six basic emotions of Ekman [2] (joy, sadness, anger, fear, disgust and surprise), in addition to two more categories: neutral category (carried no emotion) and mixed (carried more than one emotion). They started their approach with preprocessing the text to remove stop words and perform stemming. After performing the preprocessing steps, the Vector Space Model (*VSM*) is built to measure the similarities between the sentences and the constructed six emotion lexicons using Cosine measure. Finally, their model yields 65% of f-measure.

Omneya and Sturm [14] showed that emotions can be detected automatically from Arabic tweets after applying Arabic preprocessing. They started their work collecting 1,776 tweets related to the Egyptian revolution in 2011 and manually annotated each tweet using Ekman's [2] six basic emotions in addition to a lexicon of words for each emotion extracted from the dataset. The annotation process is performed by labeling each tweet with corresponding emotion category using an average of 15 human annotators. They excluded any tweet with agreement less than 50% between annotators, so 1,605 tweets left to be used in their study. The authors performed basic preprocessing steps (to remove non-Arabic letters, punctuations and multiple spaces) in addition to the removal of stop words

TABLE I. EMOTION CATEGORIES AND TOP USED EMOJIS IN EACH CLASS.

Category	A Sample of top used emojis
Joy	😍 ❤️ 👍 😊 😘
Sadness	😭 💔 😩 😢 😔
Anger	😠 😡 😤 😒 👿
Disgust	😷 😖

and stemming. They use the unigram model to represent their dataset and use the WEKA software as suite of machine learning algorithms to find the sets of words that are highly correlated with each emotion from the first 1,012 annotated tweets. These lists represent a seed for emotional lexicons that were later expanded manually. Finally, the authors compared the effectiveness of two common classifiers, which are Support Vector Machine (*SVM*) and Naïve Bayes (*NB*).

Abd Al-Aziz et al. [15] proposed a method to detect emotions using a combination of lexicon-based approach and Multi-Criteria Decision Making (MCDM) approach. Their dataset is consisted of 1552 tweets collected from two different sources and the five lexicons of emotion words are manually constructed for each of the five emotion classes (happiness, sadness, anger, fear and disgust). Then, the five emotion lexicons are evaluated by two judges based on some criteria and Cohen's Kappa coefficient is performed to measure the agreements between the two judges. After that, based on the emotion lexicons, the emotion scoring algorithm is used to represent each tweet as a vector consisting of five emotion scores. Finally, Co-Plot, which is special form of MCDM, is used to classify tweets to their emotion state using 2-D graphical representation. The importance of this approach is in its ability to handle tweets with mixed emotions.

One of the most recent work, Al-A'abed and Al-Ayyoub [16] developed a lexicon-based approach to detect emotions from Arabic text. They make use of an existing emotion lexicon called EmoLex [17] which considers the Plutchik's eight emotions. Primarily, EmoLexis created for English that consisted of 14,182 terms and then translated to 20 different languages including Arabic. After removing terms conveying no emotions and duplicates caused by translation, the lexicon is reduced to 4,279 Arabic terms which is also used in [18]. Finally, they evaluate their approach using 39 text excerpts collected from different online resources and achieved an accuracy of 89.7%.

III. METHODOLOGY

This section explains in details the steps conducted to accomplish this study. First, the data collection and labeling approaches are presented. The next step is preprocessing and feature extraction. The Bag-Of-Words (BOW) splits each document into a set of terms (words) and computes the feature

TABLE II. SAMPLE OF LABELING TWEET BASED ON EXISTING EMOJIS.

#	Tweet	Category
1	RT @777Ropi 😟شي_ينرفزك_ قلة الادب والوقاحه وعدم الاحترام خصوصا مع كبار السن # #It_makes_me_angry lack of manners, rudeness and disrespect, especially with the elderly.	(😟 = -5) anger
2	الله يرحمه ويغفر له ويسكنه فسيح جناته يارب💔. #استشهاد_الملازم_فيصل_الطوب May Allah bless his soul, forgive him and put him in paradise. #Martyrdom_ lieutenant_Faisal_Altoob	(💔 = -3) sadness
3	❤❤😭زد_رصيدك59 مايعجبني من البرنامج بكبره الا فقرة الشيخ مشاري الله يجزاه الجنة # What I like about the whole program is only a scene of Sheikh Mishari, may Allah put him in paradise.	(😭= -4) (❤❤=3*2) joy

vectors based on their frequencies. Once the features are computed, the manually labeled data is split into a training set and a testing set. Each approach is trained using classification algorithm that builds a model of it. Finally, the same testing set is used to evaluate the models of the two labeling approaches.

A. COLLECTING DATA

We collect our data from Twitter using trending hashtags. It consists of 134,194 Arabic tweets collected from August 2015 to February 2016. Using emoticons can carry strong sentiment [19] and some users express their opinions and moods using emoticons when posting tweets. A large number of Twitter users use small digital images called emojis which are a new generation of emoticons represented as a Unicode symbols. Since we only focus on four categories of emotions: anger, disgust, joy and sadness, we select emojis that are related to these four emotion categories. There are 845 emojis in Twitter, but we only take into account the top used emojis. First, we scan through the dataset in order to estimate the top frequently used emojis. Afterward, these emojis are categorized into one of the four categories of emotions adopted in our study. Table I exhibits a sample of emojis distributed among the four categories of emotions.

B. AUTOMATIC LABELING APPROACH

After assigning the emojis into their categories, we label each tweet according to the emojis it contains. The AFINN lexicon [20] is used to guarantee that one category is assigned to each emoji. The AFINN lexicon is a list of English words that are manually labeled and assigned a weight that ranges from -5 to +5, where the +/- signs represent positive and negative sentiments, respectively. In addition, there is a list of emojis with their weights listed in the AFINN lexicon. We also used some emojis that are not listed in the AFINN lexicon. To assign weights to them in a way consistent with what already have, we follow a very simple idea. Since each emoji has its own name, description and usage as presented in emojipedia,[4] we search for the emoji's name in the AFINN lexicon and assign its weight to the emoji. For example, the broken heart emoji (💔) is not existed in AFINN lexicon but it is used a lot to express sadness emotion. So, we assign to it the weight of the word "heartbroken" as listed in the AFINN lexicon. With all emojis assigned weights, we can proceed to annotate each tweet based on the sum of the weights of the emojis it contains.

TABLE III. THE NUMBER OF TWEETS IN EACH CATEGORY OF AUTOMATICALLY LABELED DATA.

Category	Total number of tweets
Joy	10467
Sadness	7878
Anger	2874
Disgust	1533
Total	22,752

TABLE IV. THE NUMBER OF TWEETS IN EACH CATEGORY OF MANUALLY LABELED DATA WITH TRAIN/TEST PERCENTAGES.

Category	Number of tweets	Training (80%)	Testing (20%)
Joy	630	504	126
Sadness	415	332	83
Anger	620	496	124
Disgust	360	288	72
Total	2025	1620	405

Table II shows few examples of how tweets are automatically labeled. The first tweet is labeled as an angery tweet due to the presence of an anger emoji as indicated in the lexicon, with a weight score (-5). The second tweet presented in Table II is labeled as a sad tweet due to the presence of a sadness emoji, with a weight score (-3). The third tweet is considered as a joy tweet due to the presence of two joyful emojis and one sad emoji, with a net weight score $((3 \times 2 = 6) - 4 = 2)$.

We have collected around 122,000 tweets from Twitter using a crawler, then we extracted 22,752 tweets that contain emojis and divided them into four emotion categories. We end up with 10,467 joy tweets, 7878 sadness tweets, 2874 anger tweets and 1533 disgust tweets. Table III shows the number of tweets in each category.

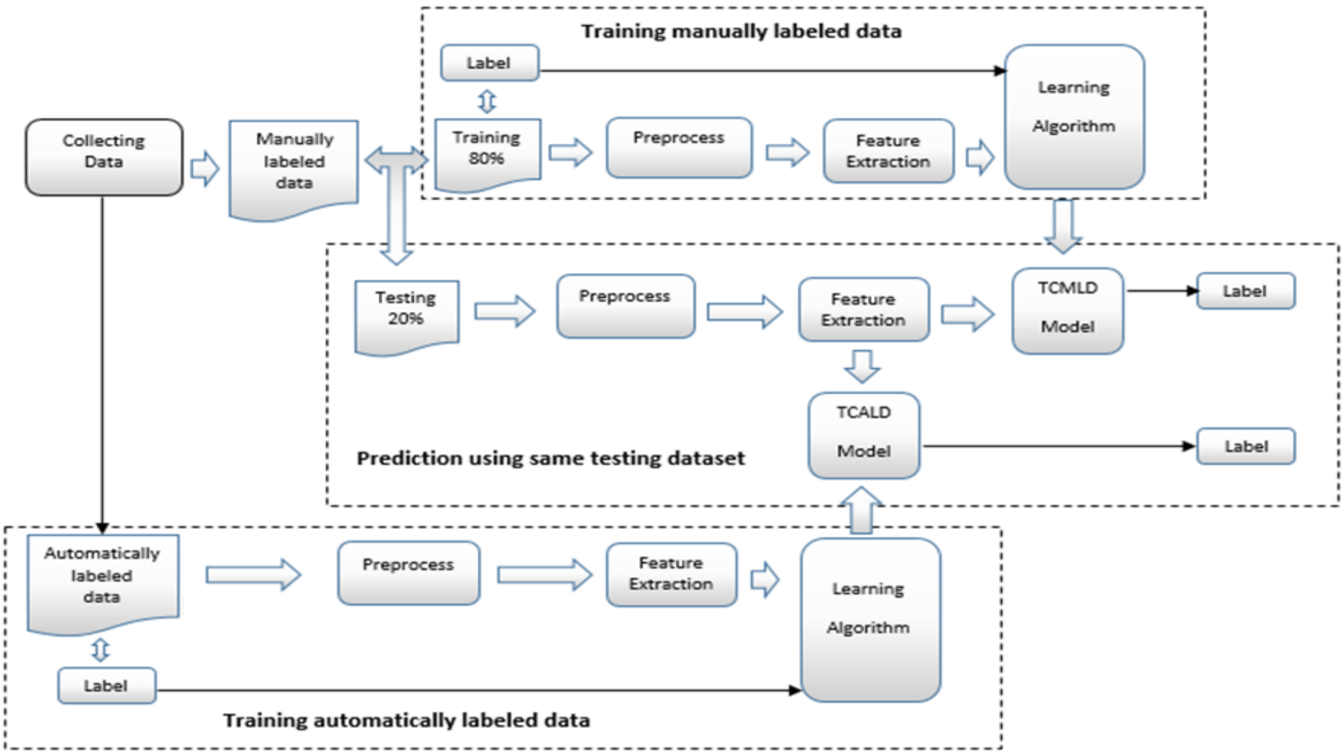

Fig. 1. System architecture

C. MANUAL LABELING APPROACH

We select 2025 tweets from our collected data that are free of emojis and manually label into the four emotion categories mentioned before. These tweets are distributed among the four categories as follows: 630 joy, 415 sadness, 620 anger and 360 disgust. Then, these 2025 tweets are divided into 1620 tweets as training data (constituting 80%), and 405 tweets as test data (constituting 20%). Table IV shows the number of tweets in each class with train/test percentages.

D. PREPROCESSING

To improve the efficiency of our work, several preprocessing steps are employed using regular expressions:

- Removing non Arabic characters.
- Removing diacritics.
- Removing definite articles (e.g. "ال", "بال", "كال", …)
- Removing special characters (@,$,!,…).
- Removing numbers.
- Normalizing (e.g. convert " إ "," أ " to " ا ").
- Removing stop words.
- Removing hashtag at the end of the tweet [21].

E. FEATURE EXTRACTION

We use the BOW approach as feature extraction. BOW relies on the occurrences of terms within a document. To provide more meaningful features, the term frequency-inverse document frequency (TF-IDF) weighting technique is applied which yields higher weights to terms appearing frequently in the given tweet and rarely in other tweet. Since the BOW approach usually generates a high-dimensional feature vectors, so dimensionality reduction techniques should be applied. But because the length of tweet is limited to 140 character, we did not need to apply any dimensionality reduction technique and we used the default value.

F. SYSTEM ARCHITECTURE

After collecting our dataset as mentioned earlier, the manually labeled data is split into training dataset (80%) and testing dataset (20%). First, the manually annotated training dataset is preprocessed and the feature extraction is performed. Then, the training phase takes place producing a trained model called Trained Classifier for Manually Labeled Data (TCMLD). After producing the model, the testing dataset is used to evaluate TCMLD model. The same process is applied on the automatically annotated training dataset to produce the Trained Classifier for Automatically Labeled Data (TCALD). Fig. 1 shows in details the classification of manually and automatically labeled data using the same testing dataset.

G. CLASSIFICATION

Generally, Naive Bayes (*NB*) and Support Vector Machine (*SVM*) classifiers are known to perform very well for the general text classification problem [22]. Since we are dealing with multi-class problem in this study, we find out that Multinomial Naive Bayes (*MNB*) outperforms *NB* with multi-class text classification. Therefore, we experimented with the following two classifiers.

- Support Vector Machine (*SVM*).
- Multinomial Naive Bayes (*MNB*).

[4] http://www.emojipedia.org

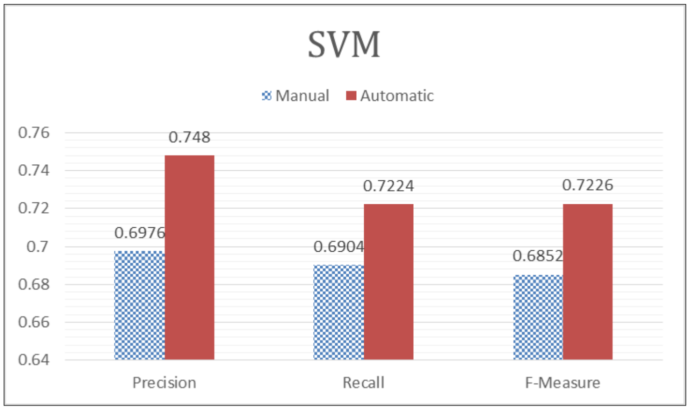

Fig. 2. Precision, recall and f-measure for manually and automatically labeled data using SVM.

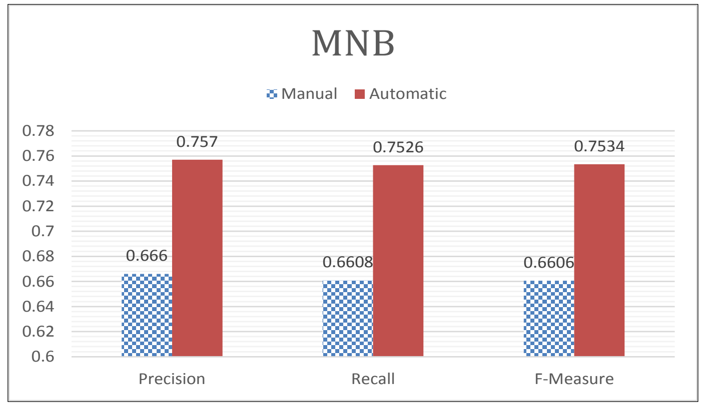

Fig. 3. Precision, recall and f-measure for manually and automatically labeled data using MNB.

IV. EXPERIMENTS AND RESULTS

A set of experiments is conducted to evaluate the automatic labeling approach. These experiments are conducted using one of the most popular data mining libraries, Weka [23].

In order to ensure fairness, we repeat our experiments five times, where in each time we change the manually annotated training and testing sets. We calculate the average precision, recall and *F*-measure for the five folds as illustrated in next section.

Fig. 2 and Fig. 3 show the results of our experiments. Fig. 2 shows that the F-measure of the automatically trained SVM classifier is about 4% higher than that of the human trained SVM classifier. The same trend is evident in Fig. 3, but with a higher advantage for the automatically trained classifier compared with the human trained one.

V. CONCLUSION

In this work, we study the problem of emotion analysis in Arabic social media. Twitter, which is one of the most popular micro-blogging services, is used to construct the dataset of Arabic tweets. These tweets are labeled manually and automatically. We employed emojis which are new generation of emoticons to automatically label tweets into four emotion categories: anger, disgust, joy and sadness.

The effectiveness of automatic labeling is studied using two machine learning classifiers Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB). The results of the tests showed that the automatic labeling approaches using SVM and MNB outperformed manual labeling approaches.

In future, we plan to study more emotion categories using other machine-learning algorithms. Also we will try to make the system built in this study available online as a web service.

VI. REFERENCES

[1] L. Akritidis, D. Katsaros and P. Bozanis, "Identifying influential bloggers: Time does matter," in *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*, 2009.

[2] P. Ekman, "An argument for basic emotions," *Cognition & emotion,* vol. 6, no. 3-4, pp. 169-200, 1992.

[3] R. Plutchik, Emotion: A psychoevolutionary synthesis, Harpercollins College Division, 1980.

[4] J. Read, "Scalable multi-label classification," *Hamilton, New Zealand: University of Waikato,* 2010.

[5] N. A. Ahmed, M. A. Shehab, M. Al-Ayyoub and I. Hmeidi, "Scalable multi-label arabic text classification," in *Information and Communication Systems (ICICS), 2015 6th International Conference on*, 2015.

[6] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing*, 2008.

[7] M. Korayem, D. Crandall and M. Abdul-Mageed, "Subjectivity and sentiment analysis of arabic: A survey," in *Advanced Machine Learning Technologies and Applications*, Springer, 2012, pp. 128-139.

[8] N. Al-Twairesh, H. Al-Khalifa and A. Al-Salman, "Subjectivity and sentiment analysis of Arabic: Trends and challenges," in *Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on*, 2014.

[9] M. Biltawi, E. Wael, S. Tedmori, A. Hudaib and A. Awajan, "Sentiment Classification Techniques For Arabic Language: A Survey," in *In the 7th International Conference on Information and Communication Systems (ICICS).IEEE*, 2016.

[10] A. Go, R. Bhayani and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford,* vol. 1, p. 12, 2009.

[11] J. Zhao, L. Dong, J. Wu and K. Xu, "Moodlens: an emoticon-based sentiment analysis system for chinese tweets," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012.

[12] E. Refaee and V. Rieser, "Can we read emotions from a smiley face? Emoticon-based distant supervision for

subjectivity and sentiment analysis of arabic twitter feeds," in *5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data, LREC*, 2014.

[13] A. F. El Gohary, T. I. Sultan, M. A. Hana and M. M. El Dosoky, "A computational approach for analyzing and detecting emotions in Arabic text," *International Journal of Engineering Research and Applications (IJERA),* vol. 3, pp. 100-107, 2013.

[14] O. Rabie and C. Sturm, "Feel the heat: Emotion detection in Arabic social media content," in *The International Conference on Data Mining, Internet Computing, and Big Data (BigData2014)*, 2014.

[15] A. M. A. Al-Aziz, M. Gheith and A. S. Eldin, "Lexicon Based and Multi-Criteria Decision Making (MCDM) Approach for Detecting Emotions from Arabic Microblog Text," in *2015 First International Conference on Arabic Computational Linguistics (ACLing)*, 2015.

[16] M. Al-A'abed and M. Al-Ayyoub, "A Lexicon-Based Approach for Emotion Analysis of Arabic Social Media Content," in *In The International Computer Sciences and Informatics Conference (ICSIC)*, 2016.

[17] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word--emotion association lexicon," *Computational Intelligence,* vol. 29, no. 3, pp. 436-465, 2013.

[18] K. Alsmearat, M. Shehab, M. Al-Ayyoub, R. Al-Shalabi and G. Kanaan, "Emotion analysis of arabic articles and its impact on identifying the author’s gender,”," *IEEE/ACS AICCSA,* 2015.

[19] S. Aoki and O. Uchida, "A method for automatically generating the emotional vectors of emoticons using weblog articles," in *Proc. 10th WSEAS Int. Conf. on Applied Computer and Applied Computational Science, Stevens Point, Wisconsin, USA*, 2011.

[20] F. A. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903,* 2011.

[21] M. Hasan, E. Rundensteiner and E. Agu, "Emotex: Detecting emotions in twitter messages," *In SocialCom-Stanford, May 2014.,* 2014.

[22] A. Alwajeeh, M. Al-Ayyoub and I. Hmeidi, "On authorship authentication of arabic articles," in *Information and Communication Systems (ICICS), 2014 5th International Conference on*, 2014.

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, no. 1, pp. 10-18, 2009.