Global Journal of Computer Sciences

Volume 05, Issue 2, (2015) 90-98

http://sproc.org/ojs/index.php/gjcs

Daily and hourly mood pattern discovery of Turkish twitter users

Mete Celik *, Engineering Faculty, Computer Engineering Department, Erciyes University, Kayseri 38350, Turkey.

Ahmet Sakir Dokuz, Engineering Faculty, Computer Engineering Department, Nigde University, Nigde 51245, Turkey.

Suggested Citation:

Celik, M. & Dokuz, A., S. (2015). Daily and hourly mood pattern discovery of Turkish twitter users. *Global Journal of Computer Sciences*. *5*(2), 90-98. doi: http://dx.doi.org/10.18844/gjcs.v5i2.183

Received 1 August, 2015; revised 23 August, 2015; accepted 12 October, 2015.

Selection and peer review under responsibility of. Prof. Dr. Doğan İbrahim, Near East University, Cyprus.

Abstract

Massive amount of data-related applications and widespread usage of web technologies has started big data era. Social media data is one of the big data sources. Mining social media data provides useful insights for companies and organizations for developing their services, products or organizations. This study aims to analyze Turkish Twitter users based on daily and hourly social media sharings. By this way, daily and hourly mood patterns of Turkish social media users could be revealed in positive or negative manner. For this purpose, Support Vector Machines (SVM) classification algorithm and Term Frequency – Inverse Document Frequency (TF-IDF) feature selection technique was used. As far as our knowledge, this is the first attempt to analyze people's all sharings on social media and generate results for temporal-based indicators like macro and micro levels.

Keywords: big data, social media, text classification, svm, tf-idf term weighting, daily and hourly mood patterns.

*ADDRESS FOR CORRESPONDENCE: **Mete Celik,** Engineering Faculty, Computer Engineering Department, Erciyes University, Kayseri 38350, Turkey. *E-mail address:* mcelik@erciyes.edu.tr / Tel.: +90-352-207-6666

1. Introduction

Social media networking sites gained popularity in the last decade. These platforms provide infrastructure for people to interact with each other and share personal information. Also, these sites are data sources for information scientists. Social media data can be used for several areas such as business, social interactions, and user feedbacks. Mining social media data provides useful insights for companies and organizations for developing their services, products or organizations.

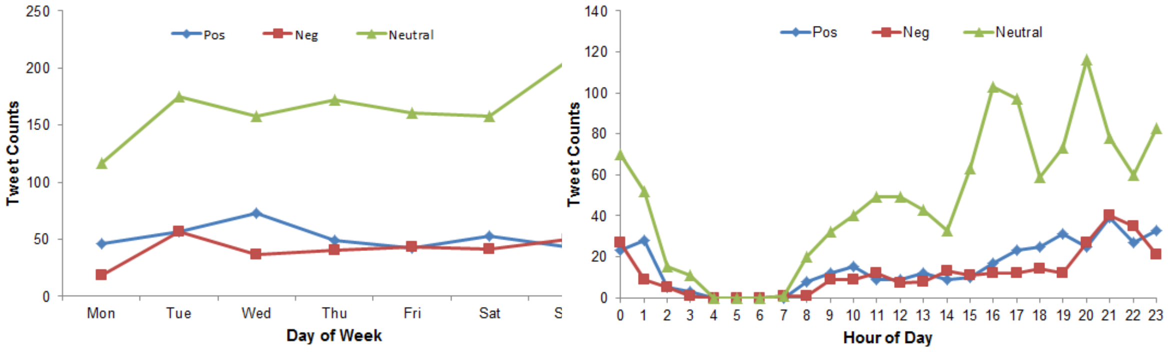
Although social media data can produce valuable information, they have several challenges [1]. First of all, social media data are at an enormous size and it's growing significantly. Second, the data are user-generated, and thus it's unstructured and very noisy. Third, social media data are dynamically changing. These challenges force researchers to develop new and novel algorithms for dealing with this kind of data.

Social media data can be used in several application domains such as sentiment analysis and opinion mining, recommender system development, graph analysis, spam detection, and trends discovery etc. Mood detection is also a part of social media mining. Mood detection is important for researches like brand monitoring, user sentiment analysis and so on.

Mood can be identified as "a strong form of sentiment expression, conveying a state of the mind such as being happy, sad or angry" [1]. Analyzing mood on social media networks can be beneficial for several application areas. Also, detecting people's mood for a special time of day or for a day of week can provide insights of societal behavior.

Mood detection literature can be broadly classified into two parts. The first group deals with detection of positive and negative moods which is called two-class mood detection [1]. The second group deals with multi-class mood detection problem and aims to classify user sentiments based on over 100 different moods [1]. While two-class mood detection generally aims to gather positive or negative thoughts about some product, multi-class mood detection aims to gather different moods and changes of moods.

This study aims to classify the moods of Turkish Twitter users as positive or negative based on daily and hourly time slices. Two definitions are done for daily and hourly time slices, macro and micro level analyses. Macro level analysis refers to analyze users' moods based on day of week scope, and micro level analysis refers to analyze mood based on hour of day scope. These analyses will provide important daily and hourly mood habits of social media users. These temporal benchmarks are selected because mood is generally affected by temporal parameters like work days or weekends, daytime or night. In addition people's mood change based on their relaxing times or stressed times. These criterions will provide important information about social media users, and indirectly people's temporal mood patterns.



(a) (b)

Figure 1. a) Macro and b) micro results of a specific user

Figure 1 a) and b) shows macro and micro level analysis results of a typical user. As can be seen from Figure 1, the number of tweets change daily and hourly. Maximum number of tweets is Tuesday and Sunday, and minimum number of tweets is Mondays for this user. Also this user has maximum tweets at night time, from 20 to 00, and minimum tweets at the first hours from midnight, from 3 to 7. If we analyze percentage of positive and negative tweet counts for macro and micro results, we can observe that this user has positive mood on Wednesdays, and hasn't particular negative mood for a day in average. For micro results, this user has positive mood from 16 to 19, and hasn't particular negative mood for a specific hour. These results reveals positive mood for this user and this users' own statistics can be used for target-oriented advertisement, or concert recommendation, or simply requesting to fill a survey in these hours depending on not to reject.

In this study, first, user data are collected with using Twitter APIs and a preprocessing step is performed. Then, Support Vector Machine (SVM) classification algorithm is used for mood classification. After, macro and micro level analyses are done based on classified tweets, whether positive or negative, and tweet's day and hour as macro and micro levels.

2. Literature Review

Mood detection is a part of sentiment analysis Sentiment analysis approaches can be divided into two broad groups; lexicon based unsupervised approaches and machine learning based approaches. Lexicon based approaches use a large corpora to classify test instances, while machine learning based approaches use classification algorithms.

Lexicon based unsupervised sentiment analysis approaches, which is proposed by Turney [2], try to label data instances using dictionaries. This approach needs dictionary content to be detailed. Lexicon based approaches can be seen as advantageous because of its' unsupervised nature. However this approach is domain and language dependent and the performance is based on dictionary size. O'Connor et al. [3] used this approach for polling using Twitter and got satisfying results. Tumasjan et al. [4] used Twitter for predicting election results. For Turkish, Vural et al. [5] and Gezici et al. [6] studied lexicon-based sentiment analysis.

Machine learning based sentiment analysis approach, which is proposed by Pang, Lee and Vaithyanathan [7], tries to label data instances using machine learning algorithms. For this approach, a training dataset and a powerful classification algorithm is needed. The algorithms are trained based on training dataset and new data instances are classified by the trained model. Go et al. [8] used Twitter for polarity classification with using several machine learning algorithms. Bifet and Frank [9] studied analyzing streaming data on Twitter based on sentiment knowledge discovery job. Jiang et al. [10] studied target-dependent sentiment classification which takes a specified query and tries to classify it to a positive, negative or neutral class by using SVM classification algorithm. For Turkish, Erogul [11] and Tantug [12] studied machine learning based supervised sentiment analysis.

In this study, different from the literature, macro and micro level definitions are done and users' moods are classified based on these levels. For this purpose, all social media data are taken into account for users and so the dataset size is big. As far as our knowledge, this is the first attempt to analyze people's all sharings on social media and generate results for temporal-based indicators like macro and micro levels.

3. The Dataset

Social media networking sites provide data collection APIs for developers. In this context, Twitter REST API [13]and Streaming API [13] were used for gathering data from Twitter. Also Twitter4j Java library was used for performing queries and getting results from APIs.

First, in this study, we picked Turkish and Turkey oriented tweets. Twitter determines tweet language and so Turkish tweets can be picked. For Turkey oriented tweets, *location* information is used. At this point, Streaming API was used and eligible streaming tweets with the query are collected. These tweets are used for determining Turkish Twitter users. Then, REST API was used and these users' followers are gathered and if the follower has a location inside Turkey. After the data collection step, 1000 unique users were randomly selected for this study. The resulting dataset contains 1000 users with 833.285 tweets.

3.1. Preprocessing

Gathered dataset is unstructured and so a preprocessing step is needed. First of all, to gather meaningful results from macro and micro analyses, the selected users should be active users. To achieve this, the users whose tweet count is less than 20 are extracted. At this part 99 users are removed from the collection because they do not satisfy required tweet number.

Main interest of this study is tweets that are user specific and include sentiment. Thus, retweets are extracted. Also, the links inside tweet contents are extracted too. Stop words and mentions, which are starting with @ character and refer another social media user, are removed from tweet contents.

3.2. Training data

Because of its application and language-oriented nature, labeled data is an important problem for social media datasets [14]. One approach is to label the dataset manually but for millions of tweets this is really time-consuming job. One alternative is distant supervision [8], which uses emoticons for labelling the data instance. Emoticons are characters which present some personal mood and are frequently used by social media users to show their momentary mood.

In this study distant supervision is used for labelling users' tweets as positive or negative. But a third class is required for this study, neutral class. For this purpose, two dictionaries are generated which includes positive and negative words in Turkish. If a tweet includes neither a positive nor a negative word, then this tweet is identified as neutral tweet. 5.000 training tweets for every class is identified based on distant supervision and the resulting training data contains 15.000 training instances.

4. Method

In this study, classification technique of SVM is used for mood extraction from tweet contents. It is one of the popular and powerful classification algorithm. Before SVM algorithm is performed, tweet contents should be converted into a feature vector format. For this purpose, TF-IDF feature selection technique is used.

Algorithm 1 shows the algorithmic steps of this study. First, data are retrieved from Twitter. For this purpose, a geographical search is performed and the results are checked whether their location is in Turkey. This checking operation is done with location information from Twitter user profiles. Next, passive users, who have less than 20 tweets, are eliminated from the collection. Then, a preprocessing step is performed which is explained at Section 3.1. In step 4, terms, or word contents of tweets, are extracted from tweets. After terms are extracted, TF-IDF values are calculated. In step 6, SVM classification algorithm is performed with explained parameters in Section 4.2. Lastly, experimental analyses are performed based on SVM's results.

```
1. Retrieve Turkish Twitter users with Twitter APIs
2. Eliminate passive users from the collection.
3. Preprocess the tweets of these users.
4. Extract terms from the tweets.
5. Calculate TF-IDF values for every term.
6. Run SVM algorithm.
7. Perform experimental analyses based on SVM's classification results.
```

Algorithm 1. Algorithmic steps of this study.

In this section, first TF-IDF term weighting approach [15] and then SVM classification algorithm [16] is explained. Also the parameter specifications of SVM for this study are explained.

4.1. TF-IDF term weighting

TF-IDF term weighting [15] is one of the most powerful information retrieval feature selection technique, but it's also widely used in text mining applications for feature selection. TF-IDF could successfully rank terms from given dataset and determine how important a term is to a document in the dataset. The importance of the term increases with the usage in the document, but is balanced with the usage of documents in the collection.

TF, IDF and TF-IDF formulas can be given as follows:

$$TF\ (t,d_i) = frequency(t,d_i). \quad (1)$$

$$IDF(t,D) = log(N/n_t). \quad (2)$$

$$TF\text{-}IDF(t,d_i) = TF(t,d_i)*IDF(t,D). \quad (3)$$

Term frequency is calculated as the frequency of term *t* in the given document d_i (Formula 1). Inverse document frequency is calculated as total number of documents count *N* is divided to document number n_t which includes term *t* (Formula 2). As a result TF-IDF is calculated as the product of TF and IDF values (Formula 3).

4.2. SVM classification algorithm

Support Vector Machines (SVM) is a supervised classification algorithm which is introduced by Boser et al [16]. SVM is successfully applied to bioinformatics, text mining, image and speech recognition, environmental sciences, and so on.

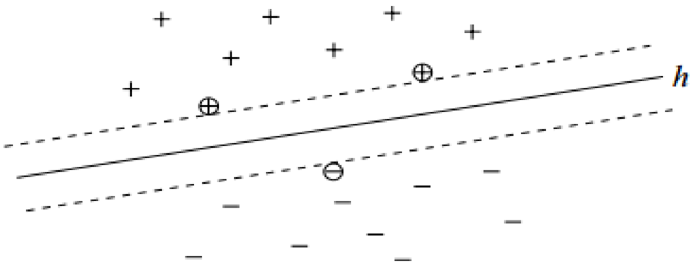


Figure 2. Best hyperplane with the margins [17]

SVM algorithm constructs a hyperplane or a set of hyperplanes in a high dimensional feature space. The selection of best hyperplane is determined with the distance to closest training examples. As shown in Figure 1, the h hyperplane is the best splitting hyperplane.

SVM algorithm is based on structural risk minimization principle, rather than empirical risk minimization principle. The idea behind structural risk minimization is to find a hypothesis h on the training examples and apply found hypothesis on test examples. This approach provides the algorithm more accuracy and freedom from the training and test examples.

For this study, SVM type is determined as C_SVC with kernel type of radial bases function. Three-class classification isn't possible, so we ran three two-class classifications. The results are evaluated and the highest class score for an instance is determined as class label.

5. Experimental Evaluation

In this section, experimental evaluation of SVM algorithm on social media data is presented. Experiments are done based on these questions:

- Which mood behavior do Turkish users show daily?
- Which mood behavior do Turkish users show hourly?
- Which days and hours are the most positive and most negative for Turkish users?

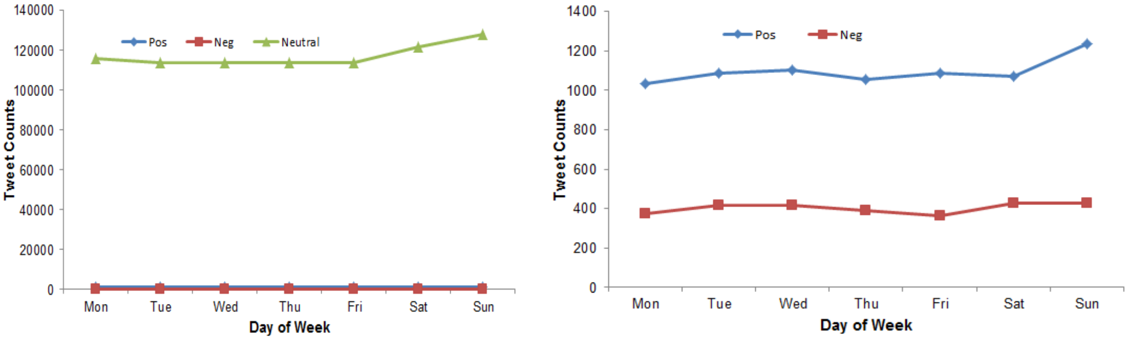
First question is entitled as macro analysis, and second question is entitled as micro analysis. These questions reveal mood patterns of Turkish Twitter users based on daily and hourly bases. With these results, Turkish social media users' moods for special day or hour can be observed and detailed sociological analyses can be done. These analyses implicitly points to the Turkish public's general tendency about their mood on daily and hourly bases.

The experiments are done on Intel Core i7 CPU with 3.40 GHz, and 8 GB RAM.

5.1. Macro analysis

Macro analysis results reveal daily mood patterns of Turkish social media users, thus indirectly Turkish people. This analysis will provide information whether Turkish people's mood changes among different days.

The main idea behind this analysis is to extract daily mood changes for Turkish social media users. This analysis will provide detailed analysis results for every day of week. The analysis results and comments are given below.



(a) (b)

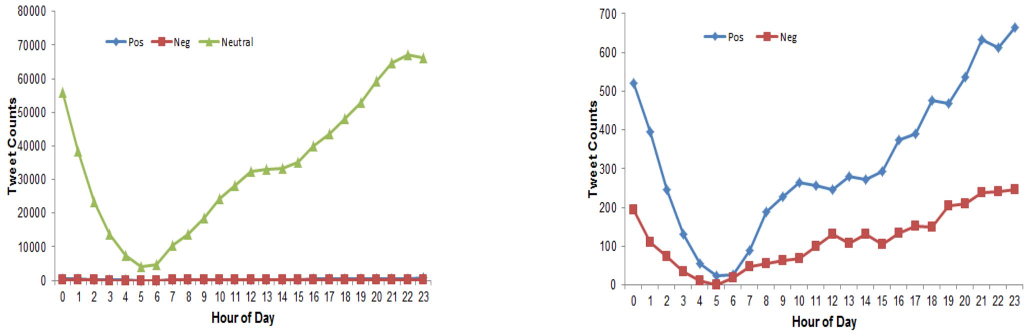
Figure 3. Macro counts a)with and b) without neutral tweets

Figure 3 a) and b) shows macro analysis results based on daily mood information. As shown in Figure 3-a), the neutral set dominates positive and negative sets, and thus hides details. So Figure 3-b) gives more information about daily changes of mood patterns. The positive set is twice of the negative set for daily bases. Also, the changes in positive set differ from the negative set at some days, i.e. Sunday. Sunday is the most important day for this analysis. Sundays, Turkish social media users are happier than other days. This information is logical, because Sundays are weekends in Turkey. Also the decrease of negative set in Fridays can be logical because government agencies start weekend from Fridays. Based on this information, we can foresee that Turkish social media users, and thus Turkish people are happiest at Sundays and minor changes occur in other days.

5.2. Micro analysis

Micro analysis results reveal hourly mood patterns of Turkish social media users. This analysis will provide information whether Turkish people's moods change based on hourly changes. The analysis results and comments are given below.

The main idea behind this analysis is to extract hourly mood changes for Turkish social media users. This analysis will provide detailed analysis results for every hour of day. The analysis results and comments are given below.



(a) (b)

Figure 4. Micro counts a)with and b) without neutral tweets

Figures 4 a) and b) shows micro analysis results based on hourly mood information. Similar with macro results, neutral tweets dominates positive and negative sets. From Figure 4 b), we can observe our actual micro results. From Figure 4 b), we can see that both sets follow similar curve led by positive set. The main reason for this type of curve is night times. Turkish social media users rest from 3 to 7, and these times can be defined as dead hours for analysis. The most active hours are before midnight. At these hours, Turkish social media users are happier than other hours. Also another important time slice is 8-12. At this slice, positive mood patterns are more frequent than negative mood patterns. The reason for this result is morning mood is generally positive. Based on this information, Turkish social media users and indirectly Turkish people are happier at night times before midnight and between 8 and 12.

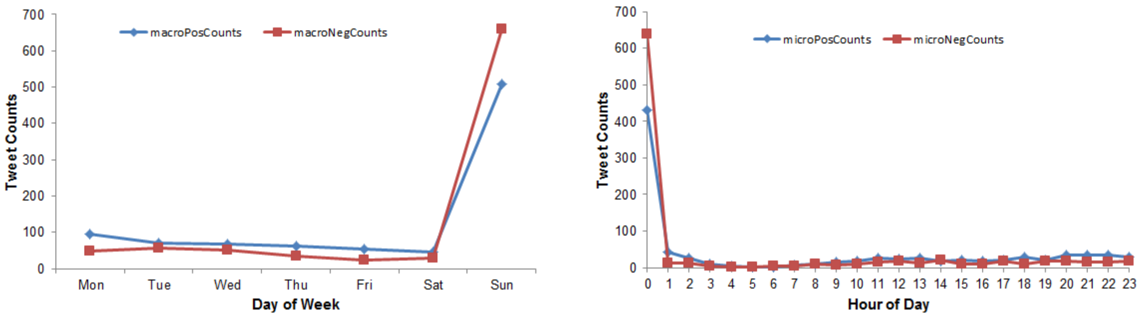
5.3. Most positive and negative day and hour analysis

In this analysis, we observed users for their most positive and most negative days and hours. For example Table 1 shows daily analysis result of the sample user from Introduction section.

The most positive day for this sample user is Wednesday because s/he shares most positive tweets at this day. For general analysis result, Wednesdays are incremented with 1 for this user. Also most negative day for this user is Tuesday because s/he shares most negative tweets at this day. Similar with positive, for general results, Tuesdays are incremented with 1 for this user. Different from macro and micro results, the main idea behind this analysis is for extracting most positive and negative habits for every Turkish social media users and cumulatively Turkish people.

Table 1. Daily analysis results for sample user.

Days	Positive	Negative	Neutral
Monday	46	18	117
Tuesday	57	57	175
Wednesday	73	37	158
Thursday	49	40	172
Friday	42	43	161
Saturday	53	41	158
Sunday	43	50	206



(a) (b)

Figure 5. Most positive and negative day analysis

Figure 5 a) and b) show the analysis results. The results for this analysis are compatible with macro and micro level analysis results. The most positive and also the most negative day is Sundays for Turkish people. The reason for negative set being above the positive set is the necessity for selecting a most positive and a most negative day for every user. The most positive and also the most negative hour is 0 midnight. Based on the results, we can say that Sunday and 0 midnight are the most important and socially most active day and hour for Turkish social media users, and thus Turkish people.

6. Conclusions

This study aims to reveal Turkish people's mood behaviors based on two important temporal aspects; macro (day of week) and micro (hour of day). These two aspects are selected because user mood is generally affected by daily and hourly changes. Identifying a person's mood on a special day or hour can benefit several applications relating people's moods. For example advertisements can be made to the person on his/her cheerful day or hour.

For the future, we are planning to add several classification algorithms to evaluate which is more suitable for mood pattern discovery. We needed more people's data for more accurate classification but unfortunately our computational resources are insufficient. The methods like cloud computing are being researched for extending computational resources.

References

[1]Nguyen, T., Phung, D., Adams, B., & Venkatesh, S., (2014). Mood sensing from social media texts and its applications, *Knowl Inf Syst, 39*(3), 667-702.

[2]Turney, P.D., (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 417-424*

[3]O'Connor, B., Balasubramanyan, R., Routledge, B.R., & Smith, N.A., (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 122-129*

[4]Tumasjan, A., Sprenger, T.O., Sandner, P.G., & Welpe, I.M., (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 178-185*

[5]Vural, A.G., Cambazoglu, B.B., Senkul, P., & Tokgoz, Z.O., (2013). A Framework for Sentiment Analysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish, *in Computer and Information Sciences III, Springer London, 437-445*

[6]Gezici, G., Yanikoglu, B., Tapucu, D., & Saygın, Y., (2012). New Features for Sentiment Analysis: Do Sentences Matter? *SDAD 2012 The 1st International Workshop on Sentiment Discovery from Affective Data,. 5-15*

[7]Pang, B., Lee, L., & Vaithyanathan, S., (2002). Thumbs up?: sentiment classification using machine learning techniques, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, 2002, pp. 79-86*

[8]Go, A., Bhayani, R., & Huang, L., (2009). Twitter sentiment classification using distant supervision, *CS224N Project Report, Stanford, 1, 1-12.*

[9]Bifet, A., & Frank, E., (2010). Sentiment Knowledge Discovery in Twitter Streaming Data, *in Discovery Science, Springer Berlin Heidelberg, 1-15*

[10]Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T., (2011). Target-dependent Twitter sentiment classification, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1,* 151-160

[11]Erogul, U., (2009). *Sentiment Analysis in Turkish*, METU, Ankara.

[12]Tantug, A.C., (2010). Document Categorization with Modified Statistical Language Models for Agglutinative Languages, *International Journal of Computational Intelligence Systems, 3*(5), 632-645.

[13] Received November 02, 2015 from: https://dev.twitter.com/

[14]Read, J., (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification, *Proceedings of the ACL Student Research Workshop,* 43-48

[15]Sparck Jones, K., (1972). A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation, 28*(1), 11-21.

[16]Boser, B.E., Guyon, I.M., & Vapnik, V.N., (1992). A training algorithm for optimal margin classifiers, *Proceedings of the fifth annual workshop on Computational learning theory,* 144-152