Malaya Journal of Matematik, Vol. S, No. 1, 26-29, 2018

https://doi.org/10.26637/MJM0S01/05

# Information extraction from text messages using data mining techniques

Sartaj Ahmad[1]* and Rishabh Varma[2]

**Abstract**

We are living in an era of increased pressure and mental disorders. The increased level of stress and pressure results in inclination of the number of people showing suicidal tendencies and thus a larger number of people are committing suicide. Stress can be caused due to family dispute, job dissatisfaction, health issues, etc. In the world of modern computing, people feel free to share their views and feelings over social media with peers and family members via services such as messaging. Due to the reserved nature and busy schedules of people it is extremely difficult to interact with peers and family members in person, therefore social media platforms are considered as the most used platform for personal conversations. The aim of this paper is to estimate the suicidal tendencies of a person by applying data mining techniques to the text messages a person sends to the associated people. By analysing the components of the text messages (key words and emoticons) we can estimate the suicidal tendencies of a person so that necessary steps can be taken in order to save the life of the subject.

**Keywords**

Text mining, knowledge discovery, sentiment analysis, opinion mining.

[1,2]*KIET Group of Institutions, Ghaziabad-201206, Uttar Pradesh, India.*

***Corresponding author**: [1]sartajahmad2u@gmail.com; [2]rishabhvarma22@gmail.com

**Article History**: Received **24** December **2017** ; Accepted **21** January **2018**

©2018 MJM.

## Contents



## 1. Introduction

The need for applying data analysis techniques to text massages arrives from the ever increasing suicide rates in different parts of the world. Saving the life of humans is the task of prime importance for a nation. In order to save the life of people their sentiments must be known and inferred so that the required steps can be taken on time. The best way to know about the sentiment of a person is by applying data mining techniques [1] to the text messages a person sends. If a person shows symbols of hyper stress then informing the people close to that person can help in saving the life of the subject. In [11, 13] Text processing is applied on the text obtained from the user. Text pre-processing involves tokenization, stop-word-removal and stemming and some other techniques. Tokenization involves splitting the text in the form of words called tokens. Tokenization is used to identify keywords in the stream of texts. Stop-word-removal is the process of removal of words which do not convey a special meaning in the document like the, and, this ... etc. Stemming is done to obtain the root word of the data and remove suffixes like -ing, -ion, etc.

This paper focuses on sentiment analysis for predicting the stress level of a person. The prediction model comprises of SVM and K-NN algorithms. This is done by feeding the system with a data set for training the system. This framework can be used in different scenarios regarding other domains. This approach can be used to predict the results of elections when applied at larger scale and for multiple subjects. It is highly effective in predicting the results regarding different opinions of people. It can be used to get prior knowledge about terror attacks or unorganized violent protests [6, 21].

Emoticons are a very important part of any textual conversation over the internet. It is also known that they are the most

expressive part of any text message as they convey the real essence of the conversation between the two counterparts [26]. Hence, it is of prime importance to analyse the emoticons used in any text message so that the real sentiment of the text is accessible.

## 2. Proposed Methodology

The proposed methodology helps to save life of people who maybe undergoing problems of hyper stress or any other factor that may prove fatal to them.

The aim is to extract information from the text messages of the user and use it for different purposes such as sentiments analysis. The model also includes the analysis of emoticons in order to completely parse the statements.

### 2.1 Data set description

The data is obtained by extracting all the text messages send by the subject. This can be achieved from multiple sources such as Facebook, Whatsapp, etc. All the messages send through these messaging services are stored in a database where we can apply our model and analyse the sentiments. The data set will contain text form of data and emoticons. No other form of data such as images will be analysed through the model.

### 2.2 Model Components

#### Sentiment analysis

In this component the data is assigned a sentiment such as positive or negative and the extent of it by performing data pre-processing using SVM algorithm [23].

• Text Pre-processing [2]
The processes involved in text pre-processing are.
*Tokenization*: Every new message is split into meaningful words called tokens. Example - "Morning walk is a bliss" is converted to "Morning" "walk" "is" "a" "bliss".
*Data standardization*: It involves converting all words in the message in standard form, converting all words in lower case [24].
Example. "The market is near Puneet's house" is converted to "the market is near puneet's house".
*Emoji conversion:* The emoticons present in the text messages are assigned a keyword based on the expression they convey [16].

The emoticons are classified into following two categories:
*Positive emoticons*: these are the emoticons which convey positive sentiment and are replaced by positive words based on the symbol [25].

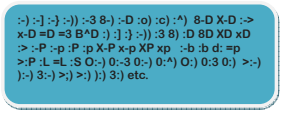


**Figure 1.** Positive Emoticons

*Negative emoticons*: these emoticons reflect the sad or disturbed sentiments of the subject and are thus replaced by negative words.

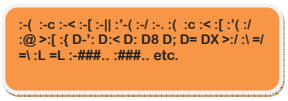


**Figure 2.** Negative Emoticons

*Stop-word-removal*: All the words in the message which do not convey a special meaning are removed like a, the, then, etc[15].
*Stemming*: It involves obtaining the root word corresponding to every word by dropping suffixes ling -ing, -ion, etc[7, 14].
*Abbreviation analysis*: Replacing the abbreviations present in the message by their full forms. Example FB by facebook, GM by good morning, etc.

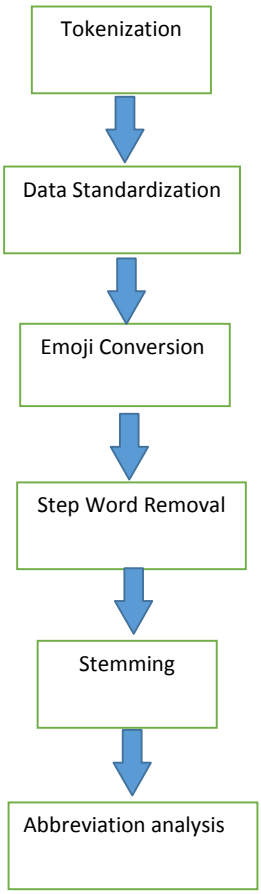


**Figure 3.** Steps in Text Message Processing

• *N*-gram [20]
The next step after data pre-processing is *N*-gram features extraction. *N*-gram is a series of *n* tokens. *N*-gram is a model very widely used in NLP tasks [3]. The model creates N-grams from the messages in the data set to extract keyword features from the data set.

For $n = 3$ a sequence of three-words for each message is generated. The process of $N$-gram increases the efficiency and accuracy of the classification step because of the feature extracted from three sequence of token combination.
Example. “What is your name” is analysed as “what is your” “is your name”.

• Term Frequency [19]
The number of times a token occurs in each data sample is called its term frequency. Words having high frequency have better relationship with the sample.

• Inverse Document Frequency
Idf factor is used to diminish the weight of words that occur very often in the data set and to increase the weight of words that occur rarely [20].

• Support Vector Machines [4]
The resulting stream of words after the text pre-processing step is processed by SVM Algorithm in order to classify the messages as “normal” or “critical” sentiment. The process is applied on every message in data set in order to classify the chat as one among “normal” and ”critical” sentiment. Thus we will get a sentiment associated with the messages associated with the user. SVM’s are supervised learning models which are used for classification and regression analysis of data used. A SVM model represents examples as points in space, different classes of examples are divided by a certain gap which must be as wide as possible. New examples when mapped into the space are predicted to belong to a class of examples based on which side of the gap they fall [10, 18, 22].

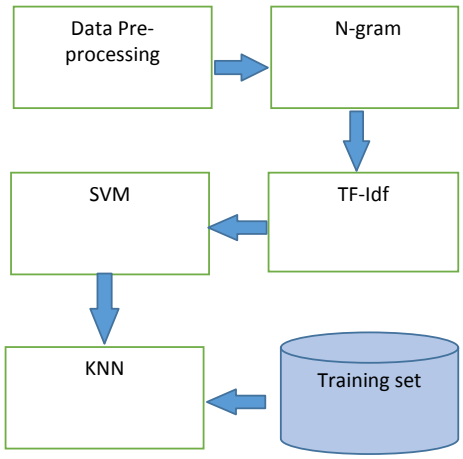


**Figure 4.** Different Steps in Data Processing and Analysis

• KNN Algorithm
The output obtained from Support Vector Machines Algorithm [9] are clusters of two sentiments with class labels “normal” and “critical”. Based on the output KNN algorithm is applied in order to deduce the overall sentiments of the subject. The input for KNN algorithm is the sentiments associated with all the chats that the subject is involved in. The last step is to predict the sentiment of the person based on the collected feature set. Data is divided into training and testing sets, and KNN algorithm is used to predict the sentiment [5]. KNN algorithm is a method for classifying data based on the nearest training sets in the feature space. The class label is assigned the same class as the nearest K instances in the training set. KNN is a type of lazy learner strategy. KNN algorithm is considered a flexible and simple classification technique based on machine learning concepts.

## 3. Result Analysis

The result obtained from the proposed model gives the estimated sentiment prediction of the subject based on the text messages sent by the user. The resulting output can be used in many situations, the mental disorders and stress level is estimated and therefore in case of “critical” sentiments the peers and family members of the subject can take actions to encourage, motivate and uplift the emotional stature of the subject thus resulting in the harmony and peace of mind of the subject. Therefore such sentiment analysis models are a requirement for shaping the society into a happening place.

## 4. Future scope

The proposed model can be used in situations where sentiment analysis is required to achieve the desired result and use it for various different purposes such as critic reviews for hotels, [17] movies, videos, etc. Sentiment analysis methods till now have been used to detect the polarity in the thoughts and opinions of all the users that access social media. Businesses are very interested to understand the thoughts of people and how they are responding to all the products and services around them. Companies use sentiment analysis to evaluate their advertisement campaigns and to improve their products. Companies aim to use such sentiment analysis tools in the areas of customer feedback, marketing, CRM, and e-commerce.

## 5. Conclusion

The proposed model takes input from the data set created by accumulating all the text messages send by the subject. All the messages may be from different social media platforms such as facebook, whatsapp, etc. The messages are then pre-processed to obtain the key words from the data sets. After pre-processing we use probabilistic language models like n gram. Associating weights to the data set using TF-Idf increases overall efficiency of classifying algorithms. The next step is to use the classifying algorithms to classify the conversations as “normal” or “critical”. [8] First a supervised algorithm is used which is SVM as it proves to be highly efficient for such computations and then an unsupervised algorithm is used which in turn increases the efficiency drastically, in our case we use the KNN algorithm. Thus we propose to give a highly efficient method [12] of finding the sentiment of the person by analysing the text messages and also processing emoticons. Emoticons [25] are very common tokens in any text message

in the new world, therefore we must also focus on efficient ways to analyse them. We have converted emoticons to textual form for our computation processes. Thus this model is a requirement and a life saviour in the modern world.

## References


[1] K. Tan, Steinbach, *Introduction to Data Mining*, 2006.

[2] C. Paper, Preprocessing techniques for text mining preprocessing techniques for text mining, *J. Emerg. Technol. Web Intell.*, 2016.

[3] D. Lyon and B. Cedex, *N*-grams based feature selection and text representation for Chinese Text Classification Zhihua WEI, *Int. J. Comput. Intell. Syst.*, 2(4), 365–374, 2009.

[4] J.C.B. Christopher, A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, 2(2)(1998), 121–167.

[5] E.-H. Sam Han, G. Karypis and V. Kumar, *Text Categorization Using Weight Adjusted k-nearest Neighbor Classification*, Springer, 2001.

[6] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[7] D.A. Hull et al., Stemming algorithms: A case study for detailed evaluation, *JASIS*, 47(1)(1996), 70–84.

[8] H. Isozaki and H. Kazawa, Efficient support vector classifiers for named entity recognition, *in Proceedings of the 19th international conference on Computational linguistics, Vol. 1, Association for Computational Linguistics*, 2002.

[9] M. James, *Classification Algorithms*, Wiley-Interscience, 1985.

[10] T. Joachims, *Text Categorization With Support Vector Machines: Learning With Many Relevant Features*, Springer, 1998.

[11] M. Kantardzic, *Data Mining: Concepts, Models, Methods and Algorithms*, John Wiley & Sons, 2011.

[12] L.S. Larkey and W.B. Croft, Combining classifiers in text categorization, *in Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, ACM*, (1996), 289–297.

[13] E.D. Liddy, *Natural Language Processing*, 2001.

[14] J.B. Lovins, Development of a stemming algorithm, *MIT Information Processing Group, Electronic Systems Laboratory*, 1968.

[15] C. Silva and B. Ribeiro, The importance of stop word removal on recall values in text categorization, in Neural Networks, 2003, Proceedingsof the International Joint Conference on, Vol. 3 (2003), *IEEE*, 1661–1666.

[16] List of Text Emoticons, retrieved 20 July 2012.

[17] L. Jin et al., A text classifier of english movie reviews based on information gain, Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence (ACIT-CSI), 2015 3rd International Conference on, *IEEE*, 2015.

[18] Cong, Yingnan, Yao-ban Chan and Mark A. Ragan, A Novel Alignment-Free Method for Detection of Lateral Genetic Transfer Based on TF-IDF, *Scientific Reports*, 6(2016): 30308. PMC. Web. 12 Oct. 2017.

[19] Trstenjak, Bruno, Sasa Mikac and Dzenana Donko, KNN with TF-IDF based Framework for Text Categorization, *Procedia Engineering*, 69(2014), 1356–1364.

[20] P.F. Brown et al., Class-based n-gram models of natural language, *Computational Linguistics*, 18(4)(1992), 467–479.

[21] A. Pak and P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, *LREc*., 10 (2010). 2010.

[22] R.S. Ramya et al., Feature Extraction and Duplicate Detection for Text Mining: A Survey, *Global Journal of Computer Science and Technology*, 16(5)(2017).

[23] R. Feldman, Techniques and applications for sentiment analysis, *Communications of the ACM*, 56(4)(2013), 82–89.

[24] I.B. Mohamad and D. Usman, Standardization and its effects on *K*-means clustering algorithm, *Research Journal of Applied Sciences, Engineering and Technology*, 6(17)(2013), 3299–3303.

[25] D. Thompson and R. Filik, Sarcasm in written communication: Emoticons are efficient markers of intention, *Journal of Computer-Mediated Communication*, 21(2)(2016), 105–120.

[26] T.A. Jibril and M.H. Abdullah, Relevance of emoticons in computer-mediated communication contexts: An overview, *Asian Social Science*, 9(4)(2013), 201–203.



⋆⋆⋆⋆⋆⋆⋆⋆⋆

ISSN(P):2319-3786

Malaya Journal of Matematik

ISSN(O):2321-5666

⋆⋆⋆⋆⋆⋆⋆⋆⋆