Sentiment Analysis of Social Networking Sites (SNS) Data using Machine Learning Approach for the Measurement of Depression

*

Anees Ul Hassan, Jamil Hussain, Musarrat Hussain, Muhammad Sadiq, Sungyoung Lee (✉)
Department of Computer Science and Engineering
Kyung Hee University
Suwon, South Korea
{anees, jamil, musarrat.hussain, sadiq, sylee}@oslab.khu.ac.kr

***Abstract*—The advent of different social networking sites has enabled anyone to easily create, express, and share their ideas, thoughts, opinions, and feelings about anything with millions of other people around the world. With the advancement of technology, mini computers and smartphones have come to human pockets and now it is very easy to share your idea about anything on social media platforms like Facebook, twitter, Wikipedia, LinkedIn, Google+, Instagram etc. Due to the tremendous growth in population and communication technologies during the last decade, use of social networks is on the rise and they are being used for many different purposes. One such service for which their use may be explored is an analysis of users post to diagnosis depression. In this paper, we present how to find the depression level of a person by observing and extracting emotions from the text, using emotion theories, machine learning techniques, and natural language processing techniques on different social media platforms.**

***Keywords*—Sentiment Analysis, Social Networking Sites (SNS), Depression Measurements**

I. Introduction

Depression has become the worlds fourth major disease. Compared with the high incidence, however, the rate of depression medical treatment is very low because of the difficulty of diagnosis of mental problems [1]. According to the World Health Organization (WHO) survey in 2012 more than 350 million people were suffering from depression and almost 1 million people with depression end their lives each year [2]. Depression is also called clinical depression or depressive disorder that is a mental disorder characterized the disruption in the mood, sadness, deliberation, loss of interest, feeling of guilt that affects how you feel, thinks, and handles daily activities. Social Networking Sites (SNS) is an online community where people across the world, irrespective of demographic and geographical differences, can make a network with different organizations or individuals to share and express their ideas, opinions, and feelings with each other [3][7]. The advanced research in several studies has focused on confirming social media observations, which increases our collective confidence in these data values that act as a source for monitoring health issues and trends. Twitter is a social media application that allows users to broadcast news, information, and personal updates to other users in tweets or statements of 140 characters or less [4]. Sentiment analysis is the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions, and emotions expressed in a text. It is a way to evaluate written or spoken language to determine if the expression is positive, or negative, or neutral. The applications of sentiment analysis are broad and powerful. The ability to extract sentiment and emotions insights from social data is a practice that is being widely adopted by organizations across the world. In this paper, we present that how to teach a machine to analyze the various grammatical nuances, cultural variations, extract emotions, and find sentiment and meaning behind that words using machine learning techniques. The rest of the paper is organized as follows. Overview of the related work is presented in section 2. Research model and proposed methodology is described in section 3 and discussion and conclusion are presented in section 4.

II. Related Work

This section provides the literature reviews on sentiment analysis of SNS data for depression measurements. In [1] the author proposed a multi-kernel SVM based model to recognize the depressed people and extracted three categories of features, user microblog text, user profile, and user behaviors from their social media to describe users situations. In [2] and [4] the authors proposed a data mining application based on classification techniques i.e decision tree C 4.5 and J 4.8 to predict that who will be a future candidate for developing depression. In [5] the author used hybrid machine learning algorithm technique i.e support vector machine to select the best feature from the training data and then give these features to the artificial neural network to perform document-level

sentiment classification. In [6] the author proposed the hybrid and wrapper approaches of supervised learning classifiers in order to find minimal features subset for document sentiment classification. The sentiment classification task involves analyzing and predicting opinions and sentiments in relation to the polarity sentiment. According to literature review, people use different machine learning classifiers for sentiment classification. No single classifier is good for all kinds of datasets. In this paper, we combine the three different classifiers by using voting approach, in which each feature is assigned a number of votes and choose that label which gets the most votes.

III. Proposed Methodology

There are two types of sentiment classification techniques, binary classification technique and multi-class sentiment classification technique. In binary classification technique each document di in D where $D = \{d1, d2, d3, ...dn\}$ are classified into category C where $C = \{Positive, Negative\}$

and in multi-class sentiment classification the di is classified into category: $C = \{StrongPositive, Positive, Neutral, Negative, StrongNegative\}$ [5]. In our proposed methodology, there are four main components that are preprocessing, feature extraction, meta learning and training data as shown in figure 1.

A. Preprocessing

In this step, we first split the paragraph into sentences and then tokenize the sentences into words. From words vector then we remove the stop words. Stop words are those words that are often useless and do not convey much meaning e.g in English the, much, under, over in, from, on, of, and etc. are stop words. On the remaining words we apply the stemming i.e find the root words e.g the root word of stems, stemmer, stemmed and stemming is stem.

B. Feature Extraction

The detection of features from raw text is related to whether a word or a sequence of words can be a feature or not. The main approaches to the problem of feature representation are bag of words, a bag of opinions, lexicons base, and dictionary based data adjective, adjectival phrases, nouns, and adverbs. Some of the feature extraction methods are the following.

- N-grams Features: N-grams are basically co-occurring words, syllables, multiple words (bigrams, trigrams and more), or phenomes within a given sequence of text or speech.
- Parts of Speech (POS) Tagging: Part of speech tagger is a piece of software that reads text in some language and assigns parts of speech to each word such as nouns, verb adjective etc.
- Negation: Negation words can change sentiment meaning of a word from negative to positive and from positive to negative e.g not bad, not good etc.
- Sentiment Analyzer: Sentiment analyzer analyze positive and negative sentiment of words in given document e.g wonderful express a positive sentiment orientation.

C. Meta Learning / Voting

The simplest way to combine multiple classifiers is to use the voting approach, and choose which ever label gets the most votes. Voting is a well-known aggregation procedure that combines different opinions of voters classifiers into consensus. The voting operator is nested operator which uses majority vote for classification and regression on top of predictions of the inner learners. Here we use three classifiers, Support Vector Machine (SVM), Naive Bayes (NB), and Maximum Entropy (ME) classifiers as inner learners.

- Support Vector Machine (SVM): SVM is a machine learning classifier which can be used for both classification and regression challenges. Single kernel SVM is widely used for data analysis in different domains including social media and linear SVM is known to be one of the best performing methods for text classification. In this algorithm we plot each data item as a point in n-dimensional space with the value of each feature being the value of each coordinate, where n is the number of features we have.
- Nave Bayes (NB): Nave Bayes is a machine learning model which can be used for both classification and regression challenges. Single kernel SVM is widely used for data analysis in different domains including social media. In this algorithm we plot each data item as a point in n-dimensional space with the value of each feature being the value of each coordinate, where n is the number of features we have.

$$P(label|features) = P(label) * P(features|label)/ P(features) \quad (1)$$

In the above equation the P (label) shows the prior probability of the label occurring, which is likelihood that a random feature set will have the label. This is based on the number of training instances with the label compared to the total number of training instances. $P(features|label)$ is the prior probability of a given feature set being classified as that label. This is based on which features have occurred with each label in the training set. P (features) is the prior probability of a given feature set occurring . This is the likelihood of a random feature set being the same as the given feature set and is based on the observed feature sets in the training data. $P(label|features)$ tells us the probability that the given features should have that label. If this feature is high, then we can be reasonably confident that the label is correct for the given features.

- Maximum Entropy: Maximum Entropy is also known as a conditional exponential classifier or logistic regression classifier. The maximum entropy classifier converts labeled feature sets to vectors using encoding. This encoded vector is then used to calculate weights for each feature that can then be combined to determine the most likely

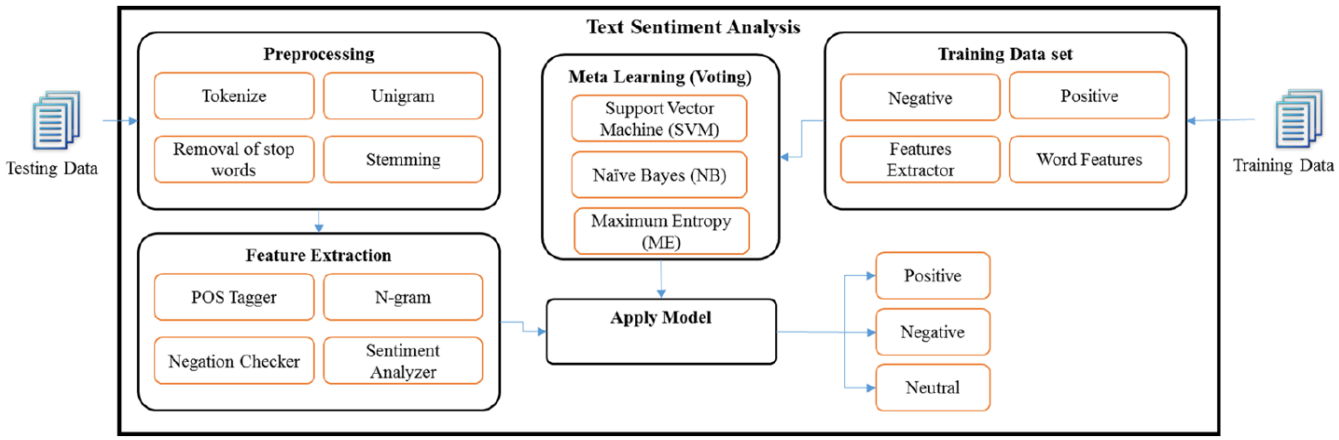

Fig. 1. Sentiment Analysis for the Measurement of Depression.

label for a feature set.

$$H(X) = -\Sigma P(X) log_2 P(X) \quad (2)$$

$$H(X) = \Sigma P(X) log_2 1/P(X) \quad (3)$$

$$H(X) = E[log_2 1/P(X)] \quad (4)$$

The range of entropy is based on the number of outcomes that is $0 \leq H(X) \leq log(n)$

IV. Conclusion

In this paper we have made a comparison among SVM, NB and ME classifiers regarding sentence level sentiment analysis for depression measurement. We have adopted voting model and feature selection technique. We examined the performance of our proposed methods on two datasets, twitter dataset and 20newsgroups. Our experiment indicates that SVM shows superior result as compare to Nave Bayes and Maximum Entropy classifiers. We observed that the accuracy of SVM is 91 %, the accuracy of Nave base is 83 % and the accuracy of Maximum Entropy is 80 %.

TABLE I
Performance Evaluation Table

Classifier Name	Accuracy		
	Accuracy	*Precession*	*Recall*
Support Vector Machine	91 %	83 %	79 %
Naive Bayes	83 %	88 %	82 %
Maximum Entropy	80 %	84 %	79 %

[a]Performance Evaluation Table.

Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-0-01629) supervised by the IITP(Institute for Information communications Technology Promotion), this research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT Future Planning(2011-0030079) and this work was also supported by Institute for Information communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2017-0-00655)

References

[1] Peng, Zhichao, Qinghua Hu, and Jianwu Dang. "Multi-kernel SVM based depression recognition using social media data." International Journal of Machine Learning and Cybernetics (2017): 1-15.

[2] Banitaan, Shadi, and Kevin Daimi. "Using data mining to predict possible future depression cases." International Journal of Public Health Science (IJPHS) 3.4 (2014): 231-240.

[3] Abhyankar, Anjali. "Social networking sites." SAMVAD 2 (2011): 18-21.

[4] Braithwaite, Scott R., et al. "Validating machine learning algorithms for twitter data against established measures of suicidality." JMIR mental health 3.2 (2016).

[5] Tripathy, Abinash, Abhishek Anand, and Santanu Kumar Rath. "Document-level sentiment classification using hybrid machine learning approach." Knowledge and Information Systems (2017): 1-27.

[6] Yousefpour, Alireza, Roliana Ibrahim, and Haza Nuzly Abdel Hamed. "Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis." Expert Systems with Applications 75 (2017): 80-93.

[7] Hussain, Jamil, Maqbool Ali, Hafiz Syed Muhammad Bilal, Muhammad Afzal, Hafiz Farooq Ahmad, Oresti Banos, and Sungyoung Lee. "SNS based predictive model for depression." In International Conference on Smart Homes and Health Telematics, pp. 349-354. Springer, Cham, 2015.