Journal of Engineering and Applied Sciences 12 (Special Issue 7): 8025-8029, 2017
ISSN: 1816-949X
© Medwell Journals, 2017

User Mood Prediction on Twitter Network with Sarcasm Detection

Harshita R. Katragadda and Dilipkumar A. Borikar
Department of Computer Science and Engineering,
Shri Ramdeobaba College of Engineering and Management, Nagpur, India

Abstract: Sentiment analysis and sarcasm detection is a specialized area in the field of Information Retrieval (IR) and Natural Language Processing (NLP). To understand the user's mood, opinion, attitude, emotion or view in the text, the study of sentiment analysis is required. Sarcasm is a special kind of sentiment where people use positive words to describe their negative feeling. Detection of sarcasm in the text helps to revamp the efficiency of sentiment analysis. In this study, we surpass simple sentiment classification viz. positive, negative and neutral to aim deeper emotion classification of Twitter data, i.e., for identifying the emotions in six discrete categories (anger, fear, disgust, sadness, happiness, surprise) and propose a generalized approach for determining the mood of the user from the Tweets and thereby detecting sarcasm in the text if any. The hybrid approach has been used here is the combination of machine learning and lexicon based approaches.

Key words: Information retrieval, natural language processing, sentiment analysis, sarcasm detection, Twitter, approach

INTRODUCTION

Textual information on the internet is immensely increasing day by day. Thereby searching and mining the textual information is becoming more challenging. Extracting the information from the text in the proper framework from the massive ocean of content is more complex. The manual efforts are beyond the human direct as it consumes more time. Therefore, automatic categorization and organizing data are more focused. Textual information is primarily divided into two main parts: opinions and facts where facts are only the subjective information while the opinions represent the sentiments.

Twitter, a microblogging service is the biggest medium for the user to express their views and sharing their experiences. As in Twitter, users can Tweet about any topic only within the 140 character limit detection of sentiment analysis becomes difficult. Many organizations are keen in knowing the opinions of the user from these data as this would be beneficial for predicting events, determining the mood of the user, business demands, movie reviews, etc. The main aim is to bind this enormous source of opinionated data to evaluate and identify the emotions in Tweets. The motivation behind this research is to reconnoiter the possibility of identifying emotions of multiple classes rather than virtuously positive or negative sentiment in short texts like Tweets.

Identification of sarcasm is one of the difficulties that still pertain in sentiment analysis. Cambridge dictionary defines sarcasm as "the use of remarks that clearly mean the opposite of what they say, made in order to hurt someone's feelings or to criticize something in a humorous way". While speaking people commonly use certain gestural clues to show sarcasm. But these clues are not present in textual data which increases the difficulty in determining sarcasm. Determining sarcastic statements can be very beneficial for improving the sentiment analysis of data (Pang and Lee, 2008).

As per the research, from 200 microblogging text with general conversation topics, sarcasm text is a rarity. Only 8 of 200 texts contained sarcasm. But for perceptive topics such as brand or politics, the number of sarcasm text significantly increases. There were 52 texts out of 200 texts containing sarcasm.

Sentiment analysis depends on the words that contain certain sentiment score, i.e., whether it deals "positively" or "negatively" with its context. But going beyond simple sentiment analysis and classifying it into deeper emotions, clarity to the emotional expressions and reactions of people is obtained. Rather than identifying it as a simple negative sentiment, the statement given below indicates anger. "He has no right to use abusive words for me!!!! #anger".

However, in some cases, the appearance of the text might be deceiving. In Twitter, sarcastic sentences are

Corresponding Author: Dilipkumar A. Borikar, Department of Computer Science and Engineering
Shri Ramdeobaba College of Engineering and Management, Nagpur, India

used very often. "All your products are incredibly awesome!!!" might be considered as a compliment. However, consider the following Tweet "Did I say incredibly??" The above sentence contains only positive words but still, the sentence expresses a negative sentiment, i.e., there is sarcasm in the text.

For detecting sentiments and mood of the users, researchers used different techniques such as maximum entropy, support vector machine and naive bayes, because these algorithms are proved to be more accurate than any other algorithm for classification of text (Lima *et al*., 2015; Borikar and Chandak, 2016; Pang *et al*., 2002; Soundarya and Manjula, 2016). Additional to that, there is also a lexical resource, Senti Word Net (Baccianella *et al*., 2010) which is a resource of the word with sentiment score. These two methods are also used in our mood detection system which is unlike other research which didn't deal with the sarcasm detection.

Literature review: This study brings out the research contribution in the specific area of Twitter mood detection and sarcasm detection.

Classification of text into positive, negative and neural, i.e., sentiment analysis substantiated to be the initial aspect of this study. Pang and Lee (2008) have conferred about the main obstacle in sentiment analysis and different methods to overcome them. They included various methods to perform sentiment analysis and extraction on structure each dealing with diverse features and characteristics of sentiment analysis. Most of the sentiment analysis workings center on machine learning techniques. The best suitable classifiers proved to be decision tree and Support Vector Machine (SVM). The most general problem is to find valence in the text. Tripathy *et al*. (2015) used Naive Bayes Algorithm and gained precision of 89.53% and SVM classifier obtained a precision of 94.06%. These accuracies were calculated on English database with 2000 samples. As a further leap, Mohammad (2012) proposed a technique using emotion word hash tags for creating a training corpus from Tweets. He also extracted a word-emotion association lexicon from this Twitter corpus which yielded better results when compared with the results of WordNet affect lexicon in the domain of six Ekman's emotion detection from Tweets. Purver and Battersby (2012) proposed a technique in which they made use of conventional markers of emotional content such as hash tags and emoticons as a substitute for explicit labels. The benefit of this method is that it gives a straight access to the reseacher's own envisioned interpretation or emotional state, not being dependent on third-party annotators.

Bharti *et al*. (2015) proposed two approaches to detect sarcasm in the text of Twitter data. The first approach is parsing-based Lexicon generation algorithm and the second approach is to identify sarcasm based on the occurrence of the interjection word. The combination of two approaches is made and compared where the results of the combined approach yielded better results when compared with the existing method. Tsur *et al*. (2010) proposed a semi-supervised approach. Two different lexical features were used, namely pattern-based and punctuation-based to build a weighted k-nearest neighbor classification model for identifying sarcasm in the text.

MATERIALS AND METHODS

Proposed research: The system mainly deals with the mood detection of a user from Tweets and to find sarcasm in it if any and the following diagram illustrates the steps involved in it (Fig. 1).

Twitter data collection: A total of 500 Tweets were extracted based on any context from Twitter and manually labeled. The raw datasets accumulated are then employed to a filtering process called data preprocessing.

Data pre-processing: Preprocessing basically aims to process and present the Tweets in an organized format and converts the text into machine understandable form. It reduces the vocabulary of terms used in the text messages thereby making the further processing easy. Preprocessing includes stop word removal, expansion and correction of words, noise data removal, pos tagging.

Noise data removal: Noise data basically refers to unwanted data in the Tweets, i.e., the words or symbols which are of no use. An example of noise data is #music, https://nfbjhbfio, etc.

Expansion and correction of words: Here, the numeric character is converted into alphabet as well as the words which are misspelled are corrected. Repeated letters are also removed in this step. Such as '2day' will be converted to 'today', 'becoz' will be converted to 'because' and so on.

Stop word removal: Stop Words are basically used to get the grammar of the sentence, they in general, do not contain any valuable information. Such words are filtered out before further processing. An example of stop words are in, how, what, to, are, also, etc.

Parts of Speech (POS) tagging: POS tagging is done after cleansing the data and it is also known as grammatical tagging. POS tagging helps in identifying the part of

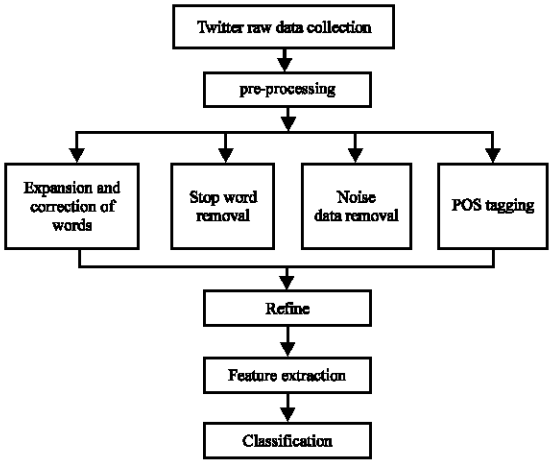

Fig. 1: Architecture of the mood and sarcasm detection system

speech of each word in the sentence. For example, the sentence 'I love ice-cream' after POS tagging will become 'I/PRP love/VBP ice-cream/NNP'.

Refining: In refining module, Tweets are differentiated either into subjective or into objective. The system scans the complete text for finding out the words that contain some sentiment score, i.e., if the word in the Tweet possesses some sentiment weight, either positive or negative then such Tweet will be classified as subjective and will go for further processing. The objective statements which possess no sentiment score will be discarded from further processing.

For finding out the sentiment score Senti Word Net is used. For example, "come and get this product" does not have any word that carries sentiment score and thereby this sentence will be classified as objective and the statement "come and get this awesome product" will be classified as subjective as it consists a word "awesome" that has sentiment score.

Feature extraction: Drawing out proper features is very essential as it would be responsible for the precision of the system. The proposed model uses several features, namely.

Lexical features:
- Unigram
- Bigram

Linguistic features:
- Content words
- The No. of abbreviations (The *et al*., 2015)
- Intensifier-adverb, adjectives
- Interjection-yay, oh, wow, yeah, nah, aha, etc.

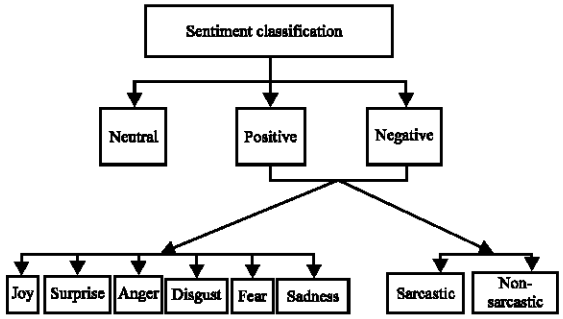

Fig. 2: Flow of classification component

Orthographic features:
- Question mark
- Exclamation mark
- Quotes-" ", ' '

Pragmatic features:
- Smiles
- Emoticons
- Replies
- Classification component

Two classification steps involved in the classification component. The first step is to classify mood of the user into Ekman (1992) six basic emotions, i.e., joy, surprise, anger, disgust, fear and sadness. The second step involves classifying sarcasm of positive as well as the negative text. The flow of the classification component is as shown in Fig 2.

Mood detection and sarcasm detection are conducted using a hybrid approach that permits the benefits of the lexical and machine learning approaches to be used. The lexicon approach is basically implemented by searching the words in an emotion dictionary and computing their emotional scores. The addition of the machine learning classifier has permitted overcoming certain limitations of the lexical method by bagging the relative information in more intricate sentences such as those with multi-entity topics, conjectures.

A profusion of machine learning algorithm such as SVM, Naive Bayes, decision-tree, etc. has been used for sentiment analysis and sarcasm detection. Support vector machine is usually used in automated text classification because it is very effective as it maps each and every feature set into a two-dimensional plane and build a model that is based on a linear line that isolates the class from the mapped feature set (Poria *et al*., 2014).

Emo SenticNet along with SVM is used for classifying mood of the user from Tweets. Emo SenticNet (ESN) (Ekman, 1992) is an emotion lexicon which is an

extension to SenticNet and WordNet Affect. ESN has 13171 words and each word is allocated to one or more emotion categories.

For identifying sarcasm in the Tweets, SentiWordNet along with SVM is used. SentiWordNet is an extension to WordNet such that all synsets can be allocated with a score concerning the negative, positive or objective connotation.

RESULTS AND DISCUSSION

Experimental data: The data which is used in the experiment was collected manually from Twitter. Training data consist a total of 200 Tweets which contains 50 neutral texts, 75 positive texts and 75 negative texts. Testing data consists 500 contains 120 neutral texts, 190 positive texts and 190 negative texts. The data collected for training and testing covers various topics such as food, movie, games, places and politics. From each topic, 100 Tweets were collected and manually labeled on the bases of whether the sentence is sarcastic or non-sarcastic and finally labeled with the mood of the user.

Experimental outcome and analysis: Three different experiments were conducted. The initial experiment is on mood classification, the second experiment is on sarcasm classification and the third classification is combined classification of mood detection and sarcasm detection.

Experiments on mood detection: The accuracy of mood detection was evaluated using three different models which were lexical learning approach, machine learning approach and finally the hybrid approach.

Table 1 illustrates that the lexicon-based model gives lower result in mood detection. The hybrid approach gives the highest accuracy when compared with lexicon-based approach and SVM.

Experiments on sarcasm detection: The accuracy of sarcasm detection was evaluated using the lexicon, SVM and hybrid approach.

Table 2 illustrates that the lexicon based model gives lower result in mood detection. The hybrid approach gives the highest accuracy.

Experiments on mood detection and sarcasm detection: The accuracy of mood detection and sarcasm detection was evaluated using the lexicon, SVM and hybrid approach.

Table 3 illustrates that the hybrid approach gives the highest accuracy when compared to SVM and lexicon approach.

Table 1: Experimental result of classification of mood detection

Models	Accuracy (%)
Lexicon	54.2
SVM	72.7
Hybrid	80.1

Table 2: Experimental result of classification of sarcasm detection

Models	Accuracy (%)
Lexicon	58.3
SVM	73.6
Hybrid	83.4

Table 3: Experimental Result of combined classification of mood detection and sarcasm detection

Models	Accuracy (%)
Lexicon	52.2
SVM	69.8
Hybrid	77.3

CONCLUSION

The experimental results concluded that the use of hybrid approach improves the accuracy of general mood detection by 7.4 and 9.8% for sarcasm detection and 7.5% for combined result mood and sarcasm detection. It has shown good results. The output of the present research can be used to enrich the performances of mood detection and sarcasm detection in future researches. Likewise, manual analysis of abundant sentences is costly. Therefore, engendering rules automatically becomes essential.

REFERENCES

Baccianella, S., A. Esuli and F. Sebastiani, 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Proceedings of the 7th Conference on International Language Resources and Evaluation, May 17-23, 2010, European Language Resources Association, Valletta, Malta, pp: 2200-2204.

Bharti, S.K., K.S. Babu and S.K. Jena, 2015. Parsing-based sarcasm sentiment recognition in Twitter data. Proceedings of the 2015 IEEE-ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'15), August 25-28, 2015, IEEE, Paris, France, ISBN:978-1-4503-3854-7, pp: 1373-1380.

Borikar, D.A. and M.B. Chandak, 2016. An approach to sentiment analysis on unstructured data in big data environment. Proceedings of the 2016 International Conference on Smart Trends for Information Technology and Computer Communications (SmartCom'16), August 6-7, 2016, Springer, Jaipur, India, pp: 169-176.

Ekman, P., 1992. An argument for basic emotions. Cognition, Emotion, 6: 169-200.

Lima, A.C.E., L.N.D. Castro and J.M. Corchado, 2015. A polarity analysis framework for Twitter messages. Appl. Math. Comput., 270: 756-767.

Mohammad, S.M., 2012. Emotional Tweets. Proceedings of the 1st Joint Conference on Lexical and Computational Semantics Vol. 1, Main Conference on Shared Task Vol. 2 and 6th International Workshop on Semantic Evaluation, June 07-08, 2012, ACM, Montreal, Canada, pp: 246-255.

Pang, B. and L. Lee, 2008. Opinion mining and sentiment analysis. Found. Trends Inf. Retrieval, 2: 1-135.

Pang, B., L. Lee and S. Vaithyanathan, 2002. Thumbs up?: Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing Vol. 10, July 6-7, 2002, Association for Computational Linguistics, Stroudsburg, Pennsylvania, pp: 79-86.

Poria, S., A. Gelbukh, E. Cambria, A. Hussain and G.B. Huang, 2014. EmoSenticSpace: A novel framework for affective common-sense reasoning. Knowl. Based Syst., 69: 108-123.

Purver, M. and S. Battersby, 2012. Experimenting with distant supervision for emotion classification. Proceedings of the 13th Conference on European Chapter of the Association for Computational Linguistics, April 23-27, 2012, Association for Computational Linguistics, Avignon, France, ISBN:978-1-937284-19-0, pp: 482-491.

Soundarya, V. and D. Manjula, 2016. Fuzzy classification techniques for effective sentiment analysis using Twitter data. Asian J. Inf. Technol., 15: 887-890.

The, J.E., A.F. Wicaksono and M. Adriani, 2015. A two-stage emotion detection on Indonesian Tweets. Proceedings of the 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS'15), October 10-11, 2015, IEEE, Depok, Indonesia, ISBN: 978-1-5090-0363-1, pp: 143-146.

Tripathy, A., A. Agrawal and S.K. Rath, 2015. Classification of sentimental reviews using machine learning techniques. Procedia Comput. Sci., 57: 821-829.

Tsur, O., D. Davidov and A. Rappoport, 2010. ICWSM-A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10), May 23, 2010, Association for the Advancement of Artificial Intelligence, Palo Alto, California, pp: 162-169.