CONCEPTUAL FRAMEWORK FOR SARCASM DETECTION FOR ENGLISH TEXT

Riya Das, Shailey Kadam, Chetan Kalra, Vijeta Nayak and Dr. Sharvari Govilkar
Department of Computer Engineering, Mumbai University, PCE, New Panvel, India

Abstract— In recent years, with the increasing popularity of social media sites, people express themselves by communicating with each other in the form of texts. Without facial expression and vocal sounds it becomes very difficult to extract the exact meaning and intentions of the text. Sentiments, sarcasm and other elements present in spoken language are lost. So understanding the sentiments of the text becomes very important. Hence, one of the basic idea in sentimental analysis is to understand the sarcasm used in the statements. Our project mainly focuses on the detection and comparative analysis of three machine learning algorithms such as Random Forest, Support Vector Machine and Naive Bayes Classifier. Finally, from the results obtained after applying these algorithms on input data which comprises of emoticons, hashtags, punctuation marks the best approach for sarcasm detection can be selected.

Index Terms — **Hashtags, Punctuation Marks, Emoticons, Random Forest Classifier, Support Vector Machine (SVM), Naive Bayes Classifier**

1. INTRODUCTION

Sentiment analysis is one of the field in Natural Language Processing that deals with people's sentiments, attitude, and emotions from text. It is one of the most widely studied field in text mining. Sentiment analysis have many domains which needs to be analyzed such as consumer product reviews, review of any hotel, movie or social events. A common task in analyzing these domains is to classify the document or statement into positive or negative sentiments. There are many challenges in Sentiment Analysis and one of them is Sarcasm Detection.

Sarcasm is a mode of satirical wit depending for its effect on bitter, caustic, and often ironic language that is usually directed against an individual. It is often used for the purpose of criticism and mockery. Sometimes sarcasm is an intended humour. It is not implicitly understood by most individuals. Therefore, sarcasm detection is the need of an hour to understand the context in which a person is expressing his views. It helps to distinguish sentence into sarcastic and non-sarcastic sentences. Major conflicts which occurs are due to misunderstanding of a written text that can be avoided through sarcasm detection system.

The pitch and the tone of the speaker used in a sentence help to perceive the context of the given text. For example "I had a great time with you in traffic." The sentence wanted to convey the message that he never enjoyed the traffic but during speech with the help of tone and expression of a person we can directly say that it is a negative message but in written text it is very difficult to understand whether it is sarcasm or not. Thus, the premise of a sentence is best understood by speech but a written text creates a lot of misunderstanding and confusion. In order the perceive the actual sentiment behind the written text and avoiding the conflicts among people. Therefore, analyzing any statement becomes our first priority.

Social Media websites, review on any product, chat applications are the best sources from which statements can be analyzed. Its anonymity provides the perfect ground to detect the presence of sarcasm in it. However, extensive usage of slang words, abbreviations, mixed language, number of punctuation marks, hashtags and emotions affect the meaning of the sentence. Hence a system is designed to analyze these given statements based upon the features like handling hashtags, punctuation marks, emoticons, etc. To enhance the result of sentiment analysis.

Our objective is to use the concept of machine learning in order to train and test various sentences. Social Media is the most budding platform with a great global outreach and an important source for sentiment analysis in social media analytic. Hence, this paper presents a method for detecting sarcasm in given text. Since, this project mainly focuses on English text, the most important process is to remove all other mixed languages present in the given statement. This is done by script validation and filtering of pre-processing block. Before training any dataset first step is to clean the data which is done by preprocessor block by removing stop words and HTML tags. This clean data is then used to train classifier such as Random Forest, SVM and Naive Bayes Classifier. The dataset is divided into approximately 70-30% in order to train and test data to get a confusion matrix which provides us with an estimate result about the training of our dataset. This paper also deals with comparing the result obtained from the above mentioned machine learning algorithms to find out which classifier gives a better result and accuracy so that the best classifier can be used in social media analytics in order to improve the understanding of the overall sentiment present in the statements.

The scope of the system would be to find the Sarcasm present in English Language Only. The contents are taken from Social media sites like twitter, Product based sites like Amazon, etc. This project deals with machine learning approach. There is always a mystery whilst encountering any kind of review. People tend to use sarcasm just for the sake of mocking a person's work or criticizing his efforts.Therefore, this system plays a vital role in sarcasm detection so that the essence of the sentence could be understood more effectively.

The recipients of the system would be organizations which use social media monitoring such as public opinion, reviews and rating of the product which provide valuable information about emerging trends and what consumers and clients think about specific topics, brands or products.and also with the rapid development of craze TV series, use of sarcasm in daily life has become more common and prominent. Besides this, use of Hashtags and emoticons have been rapidly increasing.Therefore, it has become a need of an hour for all these companies to understand the progress of their products in the market and among their clients.

2. LITERATURE SURVEY

Whiting, A. and D. Williams [1], proposed the paper that explored the uses and gratification of the consumer that they receive from social media. Based on the study of 25 interviews of people using gratifications, 10 uses and gratifications are listed with their usage are as follows:- Social interaction - 88%, Information seeking - 80%, Pass time - 76%, Entertainment - 64%, Relaxation - 60%, Expression of opinions - 56%, Communicatory utility - 56%, Convenience utility - 52%, Information sharing - 40% and Surveillance/knowledge about others - 20%. This application of uses and gratifications theory to social media not only proves to be rich and comprehensive understanding of the reasons of the consumers to utilize social media but it effectively contributes to the business and social media marketing and also helps in communicating with the potential customers by fulfilling their needs .

Hiroshi Shimodaira [11], classified the documents with its contents, and of the words of which they are composed of. Two document models - Bernoulli and Multinomial were used for the classification. The Zero Probability Problem is overcome by Laplace's law of succession or add one smoothing, that adds a count of one to each word type. Naive Bayes approximation can be used for document classification, by constructing distributions over words. The classifiers require a document model to estimate P(document | class). 1. Bernoulli document model : a document is represented by a binary feature vector, whose elements indicate absence or presence of corresponding word in the document. 2. Multinomial document model : a document is represented by an integer feature vector, whose elements indicate frequency of corresponding word in the document.

B. pang and L. Lee [9], stated the General challenges for opinion mining and Sentiment analysis which are: Contrasts with standard fact-based textual analysis, Factors that make opinion mining difficult. The mentioned Key Concepts are sentiment Polarity and degree of positivity, Subjectivity detection and opinion identification, Joint-topic sentiment analysis, Viewpoints and perspectives. Then the various features taken into considerations are Term Presence vs Frequency, Term-based features beyond term unigrams, Parts of speech, Syntax, Negation and Topic-oriented features. Afterwards Impact of labeled data is observed and obtained. Here, the unsupervised approach used are Unsupervised lexicon induction. The classification based on relationship information are: Relationships between sentences and between documents, Relationships between discourse participants, Relationships between product features, Relationships between classes and Incorporating discourse structure. Special considerations for extraction are: Identifying product features and opinions in reviews and Problems involving opinion holders. Basically, the aim to use all the techniques is achieved, but no conclusion either positive or negative can be made for the algorithms used.

Alec Go, Richa Bhayani and Lei Huang [5], proposed a different approach of Distant Supervision as they removed all emoticon and non-word tokens while training their algorithms. They found that removing the non-word tokens allowed the classifiers to focus on other features like classification. They used tweets ending in positive emoticons like ":)" ":-)" as positive and negative emoticons like ":(" ":-(" as negative. They applied Naive Bayes, Maximum Entropy, and Support Vector Machine algorithms to classify Twitter sentiment which resulted to be in the range of 80% accuracy. They concluded that the unigram model outperforms all other models used, specifically bigrams and POS features do not help. Suggestion of using Maximum Entropy classifier was provided to obtain best result of 83% with both Unigrams and Bigrams during classification. They suggested that domain- specific tweets, handing neutral tweets, sentiment analysis in regional language and utilizing emoticon data in the test set must be considered to make proposed algorithm error resistance and with higher accuracy.

Luciano Barbosa and Junlan Feng [3], understood the usage of meta - information along with feature based model which gave description and information regarding hash tag, punctuation and emoticons. From the discussion,and observation an algorithm was designed which featured that for a given word in a tweet, they mapped these words to it's part-of-speech using a part-of-speech dictionary and opinion based messages contains adjectives or interjection. Along with this mapping, the word also mapped with its subjectivity. The algorithm used a two-step classification method : first training a classifier to distinguish between subjective and objective tweets and secondly training another classifier to differentiate between positive and negative sentiment. The accuracy rate was calculated on the popular list of words collected from various websites by comparing between various approaches such as ReviewSA, Unigrams, TwitterSA. These methods, reduced the error rate was from 46% to 23%. The main limitation of this approach was in the cases of sentences that contain antagonistic sentiments and web vocabulary.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau [2], discussed SVM with Unigram based, feature based and tree kernel based model. They collected data source from various websites and generated their sample sentences. The dataset used was trained with polarity and tweets contained emoticons and noisy labels. For evaluation of data collected they used a 5-fold cross validation model. For binary classification of sentence various techniques were developed which gave result as : For Only unigram classification it had result of 71.35%, Kernel Tree Technique gave an accuracy of 73.98%, Unigram + Senti-feature technique found to be 75.39% accurate, Kernel + senti-feature technique resulted into 74.61%. Same techniques were used for ternary classification which gave 56.58, 56.31, 60.6 and 60.50% accurate results. They therefore concluded that unigram+senti-feature gave result with maximum accuracy for binary classification. Finally from the analysis it was stated that sentiment analysis for Twitter data is not that different from sentiment analysis for other genres.

Ashwin Rajadesingan and Reza Zafarani Arizona and Huan Liu Arizona [10], identify the traits using the user's past tweets. SCUBA framework for Behavioral modeling Approach was being used to analyze the user's past tweets and categorize it as: Sarcasm as a contrast of sentiments in which divisions were made based on Contrasting connotations and Contrasting present with the past, Sarcasm as a complex form of expression where readability was considered. Observations are, with no historical information, accuracy of 79.38%, is obtained. considerable gain (+4.14%) in performance is obtained by observing past 30 tweets. But including more past history still gives significant results. Results have derived that SCUBA is effective in detecting sarcastic tweets.

Dmitry Davidov, Oren Tsur and Ari Rappoport [4], discussed a supervised classification framework that provided a way to utilize tagged data and emoticons for classification. It calculated the contribution of different feature types for sentiment classification and it was shown that the framework successfully identified sentiment types of untagged tweets. This quality was confirmed by human judges. They developed a methodology in which they used four basic feature types for sentiment classification : single word features, n-gram features, pattern features and punctuation features. All these feature types are combined into a single feature vector. They also used surface patterns to classify the words into into high frequency words and content words. For each feature vector construction they developed and used k-nearest neighbours (KNN) strategy for classification and with help of Euclidean distance to matching vectors were calculated. The Amazon Mechanical Turk (AMT) service was used to obtain a list of the most commonly used and unambiguous ASCII smileys to train and generate the data sets. They also discussed about algorithms that would help them to find dependencies and overlapping between different sentiment types represented by all smileys and hashtags.

N. Kourtellis, J. Finnis, P. Anderson, J. Blackburn, C. Borcea, and A. Iamnitchi [8], introduced a peer-to-peer service (Prometheus, a P2P service that enables socially-aware applications by providing decentralized, user-controlled social data management). They emulated Prometheus the workload of two socially-aware applications and one social sensor based on previous system characterizations. The social-based mapping of users onto peers leads to significant improvements, especially for the 30 users/peer case. 15% of the invocations finishing faster when compared to the random case (some invocations can finish in half the time).

Haruna Isah, Paul Trundle, Daniel Neagu [6], proposed a product safety framework using text mining and sentiment analysis. They utilized the framework to gather and analyse views and experiences of users of drug and cosmetic products. They also demonstrated how to develop product safety lexicon and training data. Naive Bayes Classifier is used for implementation obtaining 83% of accuracy for twitter. Since, this research is work in progress yet can be used for users, product manufacturers, regulatory and enforcement agencies.

Hassan Saif, Yulan He and Harith Alani [7], proposed a novel approach of adding sentiment at each topic levels along with lexicon based pattern matching algorithm. For each extracted entity from tweets, the algorithm added the semantic concept as an additional feature and measured the correlation between the concept and negative/positive sentiment. The model used techniques such as Unigram, POS, Sentiment at topic level and semantics. Unigram and POS techniques were used to find whether sentences are positive or negative as these sentences were gathered from Stanford Twitter Sentiment Corpus. These sentences when tested, gave accuracy around 71.5% and 75.53% respectively. Data collected from Health Care Reform were trained and tested using sentiment at topic level which evaluated to be 77.02%. OMD used n-fold cross validation to detect the semantics that evaluated to be 77.18%. All these techniques used were based on binary classification. To implement these techniques three different approaches were incorporating for the analysis; replacement, augmentation, and interpolation.

So we can conclude that the approach for the Sarcasm Detection started with Lexicon based approach. Then Machine Learning was used in the mid years. As machine learning approach requires more amount of time to train the dataset, so a hybrid approach consisting of both lexical and machine learning approach was implemented which gave results in optimized amount of time.

The main inference obtained was, Twitter is not different from other social medias and therefore same approach can be used for analyzing the sarcasm present in the data from other sources as well. Therefore, this project mainly focuses on machine learning approach as it is better way to obtain whether sentences are sarcastic or not to increase its result and accuracy.

3. PROPOSED ARCHITECTURE

In this, we would be discussing about the proposed system architecture. The input of the system would be reviews or simply some content from various Social Media Sites Amazon, Zomato, etc. and tweets from twitter, etc..The first step is to clean the raw input so that a standardized format of content is obtained. From this clean data obtained we have constructed our dataset which is used in training phase to train the various machine learning classifier.

The system proposed will mainly focuses on English text, therefore the important part before cleaning the content to get dataset, it is necessary to check whether the content is in English text only. This is done by script validation and filtering of pre-processing block. Next step is to clean the content by removing all URLs present in the data, unwanted HTML tags and converting the whole data in the dataset into lower case. This cleaned data is then converted into standard format to get our cleaned data set which would be in the form of data matrix with two columns as review and labels.

The proposed system would use three machine learning algorithms to train our classifier such as Random Forest, Support Vector Machine (SVM) and Naive Bayes Classifier. The system during training phase, builds a classifier by analyzing the training data and associated label with each class and develops a pickle file which consists of all the features extracted by the model in training phase.

During testing the system accepts the input from the user and compares with the features stored in pickle and predict whether given input sentence is sarcastic or not.

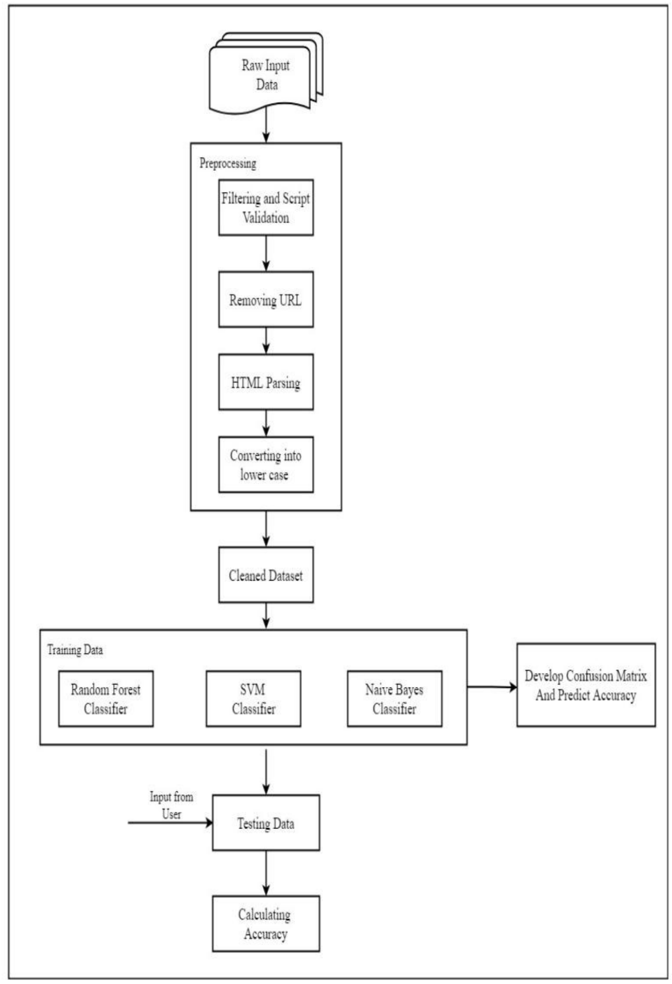

Figure 1 : Proposed Architecture

The main aim of our system is to compare these machine learning algorithm to find which algorithm can be further used to detect sarcasm during text analytic.

4. CONCLUSION

Thus, the system will be built in order to detect the presence of sarcasm not only in twitter tweets but also on various contents collected from various social media sites. As our system is implemented only for English text script validation and filtering is done with help of UTF-8. Before generating the dataset for training and testing it is necessary to clean all raw inputs. This would be implemented by using pre-processing block. It also uses three supervised machine learning algorithm such as

Random Forest, Support Vector Machine and Naive Bayes Classifier to train the classifier on the given dataset collected from social media sites in order to find certain features and storing it separately in a file which would be used later for prediction during evaluation process. Depending upon the nature of data and classification algorithm used, the system will then use the major chunk of data for training and remaining for testing.

Acknowledgment

We would like to thank Dr. Madhumita Chatterjee, Head of Computer Engineering Department for the invaluable support. We would also like to show our gratitude towards Dr. Sandeep M. Joshi, Principal, PCE, New Panvel for his invaluable support and for providing an outstanding academic environment. Last but not the least we would like to thank Prof. Dhiraj Amin and all other staff members of the Department of Computer Engineering for their critical advice and guidance without which this project would not have been possible. Also, we would like to say that it has indeed been a fulfilling experience for working out this project topic.

REFERENCES

[1] Whiting A and D Williams. Why people use social media: a uses and gratications approach. Qualitative Market Research: An International Journal, 2013

[2] Ilia Vovsha Owen Rambow Apoorv Agarwal, Boyi Xie and Rebecca Passonneau. Sentiment analysis of twitter data. In Proceedings of the ACL 2011 Workshop on Languages in Social Media, pages 30-38, 2011.

[3] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In Proceedings of COLING, pages 36-44, 2010.

[4] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. 2010.

[5] Bhyani R. Go, A. and L Huang. Twitter sentiment classification using distant supervision. Technical report, CS224N Project Report, Stanford, 2009.

[6] Daniel Neagu Haruna Isah, Paul Trundle. Social media analysis for product safety using text mining and sentiment analysis. 2015.

[7] Yulan He Hassan Saif and Harith Alani. Semantic sentiment analysis of twitter. 2011.

[8] P. Anderson J. Blackburn C. Borcea N. Kourtellis, J. Finnis and A. Iamnitchi. Prometheus user-controlled p2p social data management for socially aware applications. 2010.

[9] B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2008.

[10] Ashwin Rajadesingan, Reza Zafarani Arizona, and Huan Liu Arizona. Sarcasm detection on twitter: A behavioral modeling approach. 2015.

[11] Hiroshi Shimodaira. Text classification using naive bayes. 2015.