Sarcasm Detection in Twitter using Sentiment Analysis

Bala Durga Dharmavarapu, Jayanag Bayana

ABSTRACT--- Designing efficient and robust algorithms for detection of sarcasm on Twitter is the exciting challenge in opinion mining field. Sarcasm means the person speaks the contradictory of what the individual means, expressing gloomy feelings applying positive words. It helps the retailers to know the opinions of the customers. Sarcasm is widely used in many social networking and micro-blogging websites where people invade others which makes problematic for the individuals to say what it means. In the existing systems, logistic regression technique is used to detect these sarcastic tweets, it has a drawback as it cannot predict for continuous variables. In the proposed methodology Sentiment Analysis, Naive Bayes classification and AdaBoost algorithms are used to detect sarcasm on twitter. By using Naive Bayes classification, the tweets are categorized into sarcastic and non-sarcastic. The AdaBoost algorithm is used to make the weak statement to strong statements by iteratively considering the subset of training data. Sentiment Analysis is used to mine the opinions of customers to identify and extract information from the text. By using these two techniques, sarcastic statements can be easily classified and identified from twitter.

Keywords—Sarcasm, Sentiment Analysis, Naive Bayes Classification, AdaBoost, Twitter, Tweets

I. INTRODUCTION

In recent years Social network sites like Facebook, Instagram and Twitter has acquired extensive popularity and importance. Twitter is one of the largest social platforms where people express their opinions, feelings, views and real-time events such as live tweets etc. Twitter allows the users to register and then read and send messages which are known as tweets. Sarcasm is one of the major challenges faced in Sentiment Analysis. Twitter also enables the users to express their ideas and opinions with each other which enable the companies to know the public opinion on their products or services so that they can provide the real- time customer assistance.

Sarcasm is a form of expressing negative feelings using positive words. Sarcasm is also when people mean something else from what they speak. Sarcasm is used not only to make fun but also for criticizing other people, views, ideas etc. due to which sarcasm is very much used on twitter. Sarcasm can be conveyed in various ways like a direct conversation, speech, text etc. It can be reflected using rating of stars by providing less number of stars.

There are many applications for detecting sarcasm. It is used to let the analyst know the intent of the user and the situation in which it is said. Sarcasm is more prevalent in the places where there are capital letters, emoticons, and exclamation marks etc. Sarcasm detection is one of the prominent tasks in sentiment analysis. On Amazon and shopping websites, it helps to understand the review of the product. The consumer's preferences and opinions can be analyzed in order to understand the market behavior for better consumer experience.

II. LITERATURE SURVEY

Anukarsh G Prasad, Sanjana S, Skanda M Bhat, B S Harish [1] proposed a methodology to detect sarcastic and non-sarcastic tweets based on the slang and emojis used in their tweets. They considered the values for slang and emoji used from the slang dictionary and emoji dictionary. Then these values are compared with different classification algorithms like Random Forest, Gradient Boosting, Adaptive Boost, Gaussian Naive Bayes, Logistic Regression, and Decision Tree, to identify the sarcasm in tweets from the Twitter Streaming API. From all these classification algorithms considered the best classification algorithm is considered and combined with different pre-processing and filtering techniques using emoji and slang dictionary mapping to yield the finest efficiency.

Sana Parveen, Sachin N. Deshmukh, [2] the authors proposed a methodology to identify the sarcasm on twitter using Simple Vector Machine (SVM), Maximum Entropy algorithms. Initially, they collected the data and created into two datasets that are before adding the sarcastic tweets to the training data and after adding sarcastic tweets to the training data. POS tagging was performed using Penn tree bank to tag each word with the associated part of speech. The authors extracted features related to sentiment, punctuation, syntactic, and pattern etc from the training data. After extracting features, classification is done by using SVM, and Maximum Entropy algorithms. Compared to both the algorithms Maximum Entropy gives more accuracy when compared to the SVM algorithm.

Automatic Sarcasm Detection using Feature Selection by ParasDharwal, TanupriyaChoudary, Rajat Mittal, Praveen Kumar [3] the authors define the automatic sarcasm detection as the identification of irony in the tweets. The authors proposed a feature selection approach to recognize the sarcastic statements from the tweets. Initially, the data set is taken and preprocessing is performed by removing URLs. Pattern recognition is used for analyzing sarcastic tweets and product review comments. Feature Selection is used to extract features from existing data. Automatic sarcasm was classified into positive and negative sarcastic

Revised Manuscript Received on June 10, 2019.

BalaDurgaDharmavarapu, Department of CSE, V R Siddhartha Engineering College, Kanuru.A.P , India.

JayanagBayana, Department of CSE, V R Siddhartha Engineering College, Kanuru.A.P , India.

nature identified by using classification algorithms like Logistic Regression and Simple Vector Machine (SVM). The input data is given to the SVM to get an optimal hyperplane used for classification of data. Logistic Regression is a statistical method consisting binary classification that is it can take two values. Logistic Regression for two binary classifications is calculated using an equation. F-score is used to calculate the accuracy and it is considered for n-grams, sentiments, and topics. N-grams give more accuracy when compared to sentiments and topics in terms of F-score.

Sindhu. C, G. Vaidhu, Mandala Vishal Rao [4] in this paper the authors performed a study on detecting sarcasm using different algorithms. Data is retrieved using #sarcasm on Twitter API. Data preprocessing is applied on this data Classification algorithms like Random forest, Gradient boost, Decision Tree, Simple Vector Machine (SVM), and Maximum Entropy are used to calculate the accuracy using precision and recall to identify the sarcasm on twitter and high accuracy is given by the Maximum Entropy algorithm.

Tanya Jain, NileshAgrawal, GarimaGoyal, NiyatiAggrawal [5] in this paper the authors proposed random forest and weighted ensemble algorithms to identify the sarcasm in tweets and pragmatic classifier to detect the emotion-based sarcasm. According to the author, Sarcasm is defined as the combination of positive sentiment or feeling attached to a negative situation. Precision, recall, and accuracy is considered to calculate the efficiency for both Random Forest Classifier and Weighted Ensemble algorithms and found that both the algorithms have nearly equal precision and accuracy.Whereas, weighted ensemble exceeds Random Forest in terms of recall.

Sreelakshmi K, Rafeeque P C [6] proposed a new methodology to identify sarcasm on twitter by using a concept known as context incongruity. They considered different features like the lexical, sentiment, and context features etc to identify the sarcasm. The tweets containing word sarcasm are considered as sarcastic tweets, non-sarcastic tweets are which do not contain the sarcasm word in the tweet are considered. Both these sarcastic and non-sarcastic tweets are gathered and preprocessing is performed by removing URLs, hashtags, and words like sarcasm or sarcastic. Simple Vector Machine (SVM) and Decision Tree classifiers are used to identify the sarcasm.

Manoj Y. Manohar, PallaviKulkarni, [7] the authors proposed a new approach for sarcasm detection as NLP and Corpus-based approach. The objective was to identify the intention to use the sarcastic statement in the tweets by individuals. The authors collected the tweets from the Twitter website and NLP techniques like tokenization, parts of speech (PoS), and lemmatization are performed.NLP techniques on tweets are applied to fetch action words. Once the action words are found from the tweets, these are matched with the corpus of sarcasm data using semantic matching and graph-based matching which gives a score of sarcasm for the given tweet. By this score, the level of sarcasm in the given tweet is detected.

III. PROPOSED METHODOLOGY

The architecture of proposed methodology is shown in Figure1 as follows and mainly consists of three modules such as data preprocessing, data modeling, and classification modules.

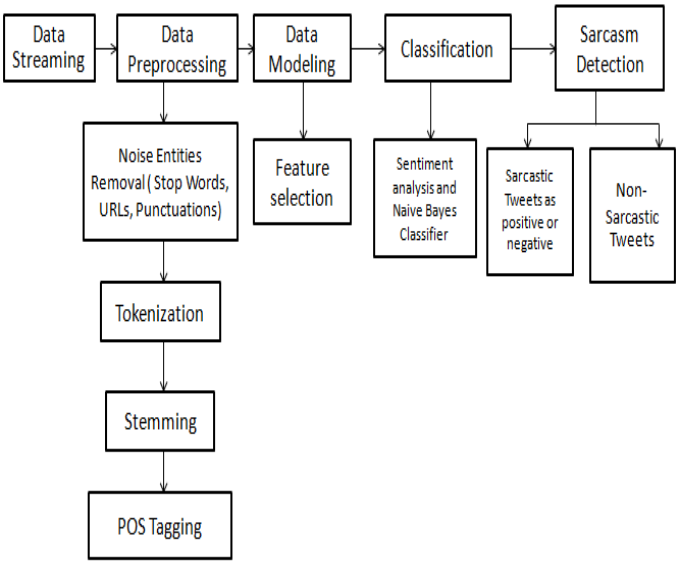

Figure1 Proposed Methodology

Data Preprocessing

In the data preprocessing phase initially, the input data is taken and hashtags are identified. These hashtags are removed from the input data.This module includes Tokenization, Stemming, Noise Removal, and PoS Tagging.

Noise Removal

Noise Removal from the input data is the process of removing the stop words, URLs, punctuation etc. These are removed from data so that the cleansed data can be easily processed for the next phase.

Tokenization

Tokenization is the process of splitting the input data into small units called tokens. The data is divided into tokens in order to easily process the data.

POS Tagging

It is the process of matching a word to its grammatical class, which helps to understand its role within the sentence. Common parts of speech considered in Pos Tagging are Noun, Verb, Adverbs, and Conjunctions, etc.

Part-of-speech taggers mostly take a sequence of words as input and provide a list of tuples as output, where each word is associated with the related tag.

Stemming

Stemming is the process of shortening a word to its root form. By using the stemming technique the number of words can be reduced in the input data. Porter Stemming algorithm is used for the stemming process.

Feature Selection

Feature Selection is the process of extracting the required features from the available input dataset. The required features are like tweet id, tweet, date of the tweet etc.

Classification

In the classification phase, the product reviews collected are classified using Naive Bayes Classification and AdaBoost Classification algorithms in order to identify the sarcasm in the tweets. For tweets, the classification is positive or negative. For the product reviews the classification technique used to identify review considering a scale of 5.

Sentiment Analysis

Sentiment Analysis is the process of identifying whether a tweet is positive, negative. Sentiment Analysis is helpful for the marketers to recognize the public opinion about their company and products, and also to consider customer satisfaction.

IV. RESULTS

The output classifies the sarcastic and non-sarcastic tweets from the given list of tweets. Sentiment Analysis organize the tweets into positive tweet or negative tweet. Naive Bayes Classification algorithm finds the sarcastic tweets by considering the probabilities.

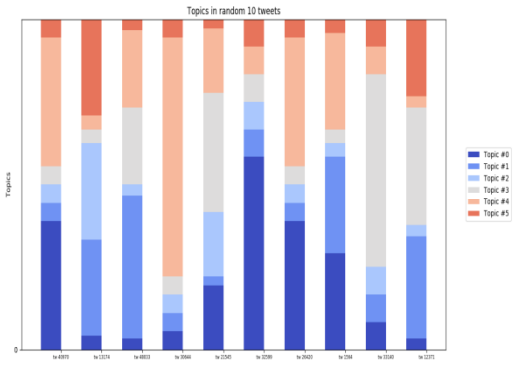

Fig1: Different topics considered for identifying the Sarcasm

Figure 1 shows the topics which are considered for classifying the sarcastic and non-sarcastic tweets and Figure 2 shows the accuracy of the model as training accuracy and validation accuracy as follows.

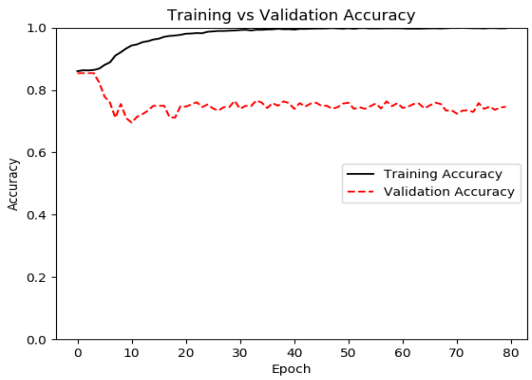

Fig 2: Accuracy Plot

The frequency of words, topics and count of these words are plotted as boxplot as in Fig 3

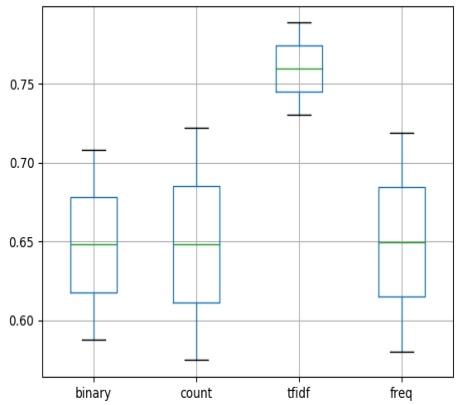

Fig 3: Boxplot showing the frequency of words

REFERENCES

1. Anukarsh G Prasad, Sanjana S, Skanda M Bhat, B S Harish "Sentiment Analysis for Sarcasm Detection on Streaming Short Text Data", 2nd International Conference on Knowledge Engineering and Applications, IEEE, 2017.
2. Sana Parveen, Sachin N. Deshmukh, "Opinion Mining in Twitter – Sarcasm Detection" International Research Journal of Engineering and Technology (IRJET), volume 04, issue 10, pages 201-204, October 2017.
3. ParasDharwal, TanupriyaChoudary, Rajat Mittal, Praveen Kumar, "Automatic Sarcasm Detection using Feature Selection", International Conference on Applied and Theoretical Computing and Communication Technology, IEEE, 2017.
4. Sindhu. C, G. Vaidhu, Mandala Vishal Rao, "A Comprehensive Study on Sarcasm Detection Techniques in Sentiment Analysis", International Journal of Pure and Applied Mathematics, volume 118, pages 433-442, 2018.
5. Tanya Jain, NileshAgrawal, GarimaGoyal, NiyatiAggrawal, "Sarcasm Detection of Tweets: A Comparative Study", Tenth International Conference on Contemporary Computing (IC3), IEEE, August 2017.
6. Sreelakshmi K, Rafeeque P C, "An Effective Approach for Detection of Sarcasm in Tweets", International CET Conference on Control, Communication, and Computing (IC4), pages 377-382, IEEE, July 05-07, 2018.
7. Manoj Y. Manohar, PallaviKulkarni, "Improvement Sarcasm Analysis using NLP and Corpus based Approach", International Conference on Intelligence Computing and Control Systems (ICICCS), IEEE, 2017.