International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 8 Issue 6, June -2019, Impact Factor: 4.059

Sarcasm Detection on Twitter data using Different Machine Learning Classifier

Vidhu Jain

IGDTUW, Dept of CSE, New Delhi, India

ABSTRACT

The Majority of research has been carried out in the area of sentiment analysis. Sentiment analysis is the progressive field of natural language processing. It is a way to detect the attitude, state of mind and emotions of the person towards the product, services, movies, etc. Sarcasm is a special kind of sentiment which infer the opposite meaning of what people convey in the text. posting the sarcasm message on social media like twitter, Facebook become new trend to avoid the direct negativity. In this paper we are trying to analyze the sarcasm, hidden sarcasm present in the online platform like twitter in the form of tweets. Twitter is the online platform where people can read and writes the short message to express their opinion on various things like politics, movies, product and services. In this paper we efficiently conduct the sentiment analysis with correct opinions of sarcastic sentence also predict the accuracy of the developed system.

Keywords: sentiment analysis, sarcasm, twitter, machine learning classifier.

HOW TO CITE THIS ARTICLE

Vidhu Jain, "Sarcasm Detection on Twitter data using Different Machine Learning Classifier", International Journal of Enhanced Research in Science, Technology & Engineering, ISSN: 2319-7463, Vol. 8 Issue 6, June -2019.

1. INTRODUCTION

Sentiment analysis is the case of natural language processing (NLP) is to track the state of people mind about the specific item or subject. Sentiment analysis that is also known as opinion mining, involves in creating a frame work to gather and evaluate opinion regarding a product in review blogs, tweets. Opinion mining and sentiment analysis both the terms are interchangeable. They express a common significance but few researchers expressed that there is a slight difference in sentiment analysis and opinion mining. Opinions of people about an entity are extracted and analyzed in case of opinion mining whereas sentiment analysis recognizes the emotion or sentiment or attitude present in a text and then analyzes of that sentiment or emotion or attitude takes place. Sentiment analysis mainly classify the text in three categories such as 'positive', 'negative' and 'netural'. People share their opinion, views, reviews on the online social media and popularity in topic ranging from entertainment, politics, products and many more.

A large amount of structured and unstructured data is produced as people share information, thoughts and views with the world through social and the analysis of this data is a challenging task for extracting opinion and sentiment. The huge amount of user generated opinion need to be analysed to make better decision by politicians, company and researcher and so on. Out of these topic, Politic is one which reviews much attention. Twitter is a platform for the politician for up-to- date status about the campaign, ralies and schemes for human welfare and than people respond to the particular news. In the field of marketing and company use opinion and sentiment to build up the stratergy to understand the feeling of the customer towards the product and brand.

Sarcasm is a type of sentiment where people express their negative feeling using positive and intensified positive words. While speaking, people often use heavy tonal stress and certain gestural clues like rolling of the eyes, hand movement, etc. to reveal sarcastic. In the textual data, these tonal and gestural clues are missing, making sarcasm detection very difficult for an average human. Sarcasm Sentiments analysis is rapidly growing area of NLP with research ranging from word, phase and sentence level classification to document and concept level classification. Sarcastic sentiment detection

is classified into three categories based on text features used for classification, which are lexical, pragmatic and hyperbolic.

Lexical:

Text properties such as unigram, bigram, n-grams, etc. are classified as lexical features of a text.

Hyperbole:

Hyperbole is another key feature often used in sarcasm detection from textual data. A hyperbolic text contains one of the text properties, such as intensifier, interjection, quotes, punctuation, etc. Previous authors used these hyperbole features and achieved good accuracy in their research to detect sarcasm in tweets.

Pragmatic:

This section describes the overall framework for capturing and analyzing tweets streamed in real time. In addition, the architecture of Hadoop HDFS followed by POS tagging, parsing and sentiment analysis of the given phrase or sentence are elaborated.

The rest of the paper as follows section II represent the literature survey. Section III represent the methodology and dataset used for analysis the purpose. In section IV have Existing techniques and section V represent the result finally section VI show the conclusion of our study

2. LITERATURE SURVEY

In [1]Sarcasm detection in twitter is a very important task as it had helped in the analysis of tweet. The goal of this research is to analysis and predict which party or candidate will win in upcoming Indian Central government election of 2014 based on sarcastic tweets. The most important goal is to find out whether more no of sarcastic tweets for particular party or candidate can result in lesser no of seat in central government election. Collecting the dataset from the twitter and Data pre- processing can be done. Lexical analysis based approach is used. Supervised learning is used to identify the sarcasm. Polarity of sarcastic tweets is calculated. Get the polarity of 0.475. Here we change the polarity of sentence 1 +0.475Sentence1 is having positive polarity that is for NAMO and sentence2 have negative polarity that is for CONG. In [2] a fair amount of work has been done on automatically detecting emotion in human speech, there has been little research on sarcasm detection. Applied 12 classification algorithms (Gradient Boosting, Gaussian Naive Bayes, etc.) on 4 types of datasets (Set1, Set2, Set3, Set4) and varied the split ratio of the datasets to check for the accuracy of every algorithm in different Situations Applied the behavioral approach to sarcasm detection on twitter dataset. In set 4 were, we found gradient boosting to give the best accuracy in all 3 cases of a split ratio-50:50, 25:75, 10:90 (85.14%, 85.71%, and 85.03%). In[3]Sarcasm detection in twitter is a very important task as it had helped in the analysis of tweet. With the help of sarcasm detection, companies could analyze the feelings of user about their products.

The paper provides the polarity of tweets which include whether the tweet is positive, negative or neutral. Polarity confidence and subjectivity confidence are also found. Accuracy of tweets is found using Naïve Bayes and SVM classifiers. In [4] Naive Bayes, one-class SVM and Gaussian kernel are few algorithms commonly used to perform the same task. To improve accuracy we implement the capitalization. In which we focus on the word which are capital .In this we detect the word which are to be focused to be spoken. Through the capitalization technique the feature is to be extracted and than naïve bayes algorithm is to be applied. Naive bayes is one of the most popular in machine algorithm. Naïve bayes is a type of classifier which use bayes theorem. All the feature considered to be independent and making equal contribution to final result. The accuracy of naïve bayes is 65.35% the accuracy of 2ndalgorithm is65.78% which is decision tree. The accuracy of 3rd algorithm is 69.37%which is SVM. Novel nayes bayes method is to detect the sarcasm in amazon alexa dataset .The data set is divided into training and testing set .The feature that extract from the training set is to improve the performance of techniques .Sarcasm is interconnected to language traints.

3. METHODOLOGY

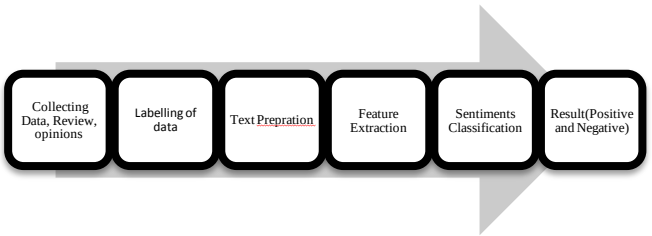

Fig 1: Methodology

Many people share their views and opinions about Sarcasm detection on microblogging sites such as twitter and Facebook. In our proposed approach we will extract the sentiments from the tweets and comments. By analysing these tweets and comments we can classify the polarity of these opinions as positive, negative and neutral. The various steps in the proposed Sarcasm detection approach is as follows:

1. **Collecting data:** The first step of sentiment analysis consists of collecting data from user generated content contained in blogs, forums, and social networks. We had streamed tweets from the twitter using Twitter Streaming API. Twitter is a microbloging website through which we can read and write short message (140 character text message) called tweets. So we collect the political tweets using keywords such as #BJP#Congress etc.

   All the tweets are related to these keywords and are stored in our data base for further processing.

2. **Labeling of data:** After collecting the data from the twitter. We had labeled this as sarcastic or not sarcastic.

3. **Text prepration:-**Data cleaning is done in this step. Tweets collected using the above mentioned keyword were having unwanted feature such as hyperlink, hashtags, @, were removed because it includes noise with data. The retweets and tweets containing English Language were removed. Let us take an example the intruder has no place in this country for the ears to be heard freely. After the formation of our government, one intruder will be chosen and carried out: Shri @Amit Shah #Desh Modi Ke Sath. Now cleaning the data is done with theabove tweets like @Amit shah and #Desh Modi ke Sath are removed**.**

4. **Feature extraction:-**Sentiments are detected from the prepared text. In feature extraction phase, the extracted sentences of the reviews and opinions are examined. Sentences with subjective expressions (opinions, beliefs and views) are retained and sentences with objective communication (facts, factual information) are discarded.

5. **Classification Sentiment:** Extracted features are classified into positive, negative. In Sentiment classification step, subjective sentences are classified in positive, negative, good, bad; like, dislike but classification can be made by using multiple point.

6. **Output:-** The final step of sentiment analysis process is presentation of output.

Table1: Representation of Positive and negative tweets

Positive Tweet percentage	**71.2%**
Negative tweets percentage	**28.8%**

4. EXISTING APPROACHES

Machine learning:-A machine learning approach use accelerated amount of machine learning to resolves text classifier [13]. It may utilize the supervised and unsupervised machine learning approach. The supervised approach is mostly used by machine learning where there is a limited collection of classes. This technique requires labeled for training of classifier. Automatic classifier use training set to learn the distinctive attribute of document and a test set is utilized to approve the performance of automatic classifier in machine learning based classification. When it is hard to find the labeled training document then the unsupervised stratergies are utilized. Unsupervised learning does not need earlier training to mine the information. The machine leaning is applicable to sentiment analysis that is mostly related to supervised classification. The machine learning approach is utilized for anticipating the sentiment polarity depend upon the training and test dataset. Various machine learning method have been used for classification of the review. The fundamental procedure of machine learning approach involves pre processing, Text expression (feature selection, feature weighting), characterization and result.

1. Suppot vector machine learning

Support vector machine learning which works for regression and classification problem. A support vector machine constructs a hyper-plane which can be used for classification, However it is mainly used in classification challenges. In SVM we used each data item as a point in n-dimensional space (where n is the no of feature in our problem set) with the value of each feature being the value of particular coordinate. Then we can perform the classification by calculating the hyper-plane that differentiate the two classes efficiently. The distance between the hyper plane and the nearest data point from either set is known as margin. The goal is to choose a hyper plane with the greatest possible margin between the hyper plane and any point within the training set, giving a greater chance of new data being classified correctly.

2. Decision Tree

A Decision tree is a tree like structure. In this each external node represent an analysis on a feature, each branch nodes denotes the outcome of the analysis and each leaf nodes carries and represent a class label. The top node in the descision tree is called root node. We used decision tree because they can easily handle the high dimensional data and their representation is self –explanatory. Decision tree is a white box type of machine learning. It shares the internal decision making logics, which is not found in the black box type of algorithm such as neural network. The time complexity of the decision tree is the function of the no of records and no of attributes in the given data. The decision tree is a distribution-free or non-paramitric which does not depend on the probability distribution assumption. Decision tree can handle high dimension data with good accuracy. We have used classification and regression tree (CART) algo for construction of decision tree. We can implement this algorithm using python. This algorithm generates binary tree for classification. During tree construction, selection for appropriate attributes is essentials. Proper decision should be made about the attribute which attribute should be placed on the top and which attribute should we be placed at deeper level.

3. Naive Bayes

Naive bayes is the one of the most popular machine learning algorithm .It is fast algorithm for classification problem. It give great result for analyzing the textual data .Naive bayes is a type of the classifier which use the bayes theorem.

Bayes theorem basically calculates the conditional probability of the occurance of an event based on prior knowledge of conditions that might be related to the event. Naive bayes classifier is based on baye's theorem which gives the conditional probability of an event A given B bayes theorem:-

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)}$$

P(A/B) =conditional probability of A given B.
P(B/A) = conditional probability of B given A.
P(A) =Probability of A.
P(B) = Probability of B.

5. RESULT

We collected some tweets from the twitter all the tweets related to political parties and than we check the polarity that is wither the are positive or negative. We have applied such as naive bayes, Support vector machine learning and decision tree to check the accuracy the results are shown in table below

Table2: Accuracy of SVM classifier with variation in C, Gamma, Kernel

C	gamma	Kernel	Accuracy
1	0.005	Linear	74%
100	0.1	Rbf	77.14%
10	0.01	Polynomial	75%

Table3: Accuracy of Naïve classifier with variation in split ratio

Split ratio	Accuracy
0.20	99.64%
0.10	99.68%

Table 4: Accuracy of decision tree with variation in tree, N-fold, Maximum size.

Tree	N –fold	Max depth	Minimum size	Mean accuracy
1	5	10	1	77.39%
5	5	10	1	76.81%
10	5	10	1	76.52%

CONCLUSION

Sarcasm detection on twitter data provides various opinion about public on the various topics like political, company, product. We collect the political tweets and than check the polarity of tweets. We calculate the accuracy of SVM classifier, Naïve bayes classifier and Decision tree classifier .Now we calculate the accuracy of the supervised learning algorithm. The accuracy of naive bayes is 99.68%. Accuracy of SVM is 77.14%. And accuracy of decision tree 77.39%. Thus Naïve bayes provides more accuracy as compared to other classifer.

REFERENCES

[1] Tayal, D. K., Yadav, S., Gupta, K., Rajput, B., & Kumari, K. (2014, March). Polarity detection of sarcastic political tweets. In *2014 International conference on computing for sustainable global development (INDIACom)* (pp. 625-628). IEEE.
[2] Ahuja, R., Bansal, S., Prakash, S., Venkataraman, K., & Banga, A. (2018). Comparative Study of Different Sarcasm Detection Algorithms Based On Behavioral Approach. *Procedia computer science*, *143*, 411-418.
[3] Saha, S., Yadav, J., & Ranjan, P. (2017). Proposed approach for sarcasm detection in twitter. *Indian J. Sci. Technol*, *10*, 25.
[4] Anand kumar D. Dave, Nikita. P. Desia," A Comprehensive Study of Classification Techniques for Sarcasm Detectionon Textual Data" 2016 international conference one lectrical, electronic and optimization technique (pp.1-7).
[5] Loganantharaj, R. (2007, May). Extensions of Naive Bayes and their applications to bioinformatics. In *International Symposium on Bioinformatics Research and Applications* (pp. 282-292). Springer, Berlin, Heidelberg.
[6] Rajadesingan, Ashwin, Reza Zafarani, and Huan Liu. "Sarcasm detection on twitter: A behavioral modeling approach." *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 2015.
[7] Maynard, D. G., and Mark A. Greenwood. "Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis." *LREC 2014 Proceedings*. ELRA, 2014.
[8] Pandey, A. C., Seth, S. R., & Varshney, M. (2019). Sarcasm Detection of Amazon Alexa Sample Set. In *Advances in Signal Processing and Communication* (pp. 559-564). Springer, Singapore.
[9] Mouthami, K., Devi, K. N., & Bhaskaran, V. M. (2013, February). Sentiment analysis and classification based on textual reviews. In *2013 international conference on Information communication and embedded systems (ICICES)*(pp. 271-276). IEEE
[10] Mehndiratta, P., Sachdeva, S., Sachdeva, P., & Sehgal, Y. (2014, December). Elections again, twitter may help!!! a large scale study for predicting election results using twitter. In *International* Conference *on* Big *Data Analytics* (pp. 133-144). Springer, Cham
[11] Bouazizi, M., & Ohtsuki, T. (2015, August). Opinion mining in Twitter: How to make use of sarcasm to enhance sentiment analysis. In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1594-1597). IEEE
[12] Dave, A. D., & Desai, N. P. (2016, March). A comprehensive study of classification techniques for sarcasm detection on textual data. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (pp. 1985-1991). IEEE.
[13] Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. *International Journal*, *2*(6), 282-292.