Comparative Study of different algorithms in Sentiment Analysis on Twitter

Deep Ranjan Kumar[1], Dr. Mohan Aradhya[2]

[1]PG Student, Master of Computer Applications, RV College of Engineering, Karnataka, India
[2]Assistant Professor, Master of Computer Applications, RV College of Engineering, Karnataka, India

ABSTRACT

Today, understanding people's concern is important in various domains. Lots of people express their feelings in various way on social media. Among people's sentiment is very important because it has great impact on our society. For various purposes understanding the sentiment of people is important like business, elections and so on. For any business knowing the sentiment of people is important and it helps to grow it by taking some decisions on it. Organizations can make decisions about their many products when they get to know that what people think about their product that they are in support or in against. Social media is a very big platform where data can be collected. Twitter is a social media platform and millions of people uses it and write something about any subject that shows their feelings. It is the best platform to get text data because most of the people writes here nearly 500 million tweets are sent each day. Sentiment analysis is an opinion mining where sentiments of people can be categorized according to their tweets. The Sentiment Analysis proves its efficiency in analyzing the emotions of the user about a particular subject. There are various algorithms used in this project every algorithm predicts the result but which algorithm is the best suitable that it can predict the best output with same types of data. It can help in categorizing the tweets that how many persons is supporting the subject and how many are not. It can categorize the positive and negative sentence and that can be used for making some important many decisions.

Keywords: - *Sentiment analysis, Twitter, Data pre-processing, Feature extraction, training, testing, accuracy rate, Naïve Bayes, Support Vector Machine.*

1. INTRODUCTION

Sentiment analysis is used to analyze the sentiment of people. It is an opinion mining which extracts the sentiments of people from tweets they have posted on twitter. In any social website huge amount of data is gathered and that data can be used. Twitter is the best platform to extract text data. More than 300 people uses twitter in a month and most of them express their feelings by writing tweets. Using the sentiment analysis technique data can be collected on specific subject and that data can be used to train the model and that model is used to predict the output. The most important step in sentiment analysis machine learning approach is to find the best suitable algorithm to train the model. There are various classification and clustering algorithms are available that can be used in sentiment analysis. It is the comparative study of sentiment analysis accuracy should be better with their data type that which is better algorithm for sentiment analysis.

1.1 Objective

- To find the percentage of positive, negative and neutral sentence.
- To identify the various algorithm applied to solve the algorithm.
- To find the optimized way to reach the solution**.**

2. EXISTING WORK

2.1 Literature Survey

This paper explains about the sentiment analysis on twitter that how it can be started. Author explains about the steps involved in machine learning. Data gets collect in the first step and then that data gets cleaned according to the need because purpose of the survey defines the type of data. After cleaning process of data feature extraction is important step then training of model starts by providing source data and target data then model is used to predict the output [1]. This paper explains about the process that how sentiments can be extracted using social media. Autor proposed a method for assessing the semantic orientation of text that how can be positive and negative sentences extracted from the sentences. This paper explains about the method used for machine learning technique to categorize the positive and negative sentences from the documents [2]. This paper explains about the algorithm like Naïve Bayes and maximum entrophy is used for classification that how to categorize the text in positive and negative in better way. This paper also explains about how algorithm is used all the steps in algorithm is explained to understand what type of data format should be there to use with algorithms [3]. This paper explains about how twitter can be helpful for sentiment analysis and opinion mining and how it can be a corpus for this type of project. It explains that there are many social networking websites are available but how Twitter can be the best because it has huge collection of text data and millions of tweets are posted daily [4]. This paper explains that how news and blogs can be analyzed. They are very useful for sentiment analysis because if sentiment analysis is done on news and blogs then it can be analyzed the condition of subject by extracting the positive and negative news. It can also be used to analyze the emotion of people about any subject that they are sad, angry and so on [5]. This paper explains that how sentiment analysis can be done using support vector machine after collecting the data from different sources. Support vector machine is useful classification algorithm and it is used widely in sentiment analysis. This paper also explains that how the data can be arranged after collecting it different sources [6]. This paper explains about the naïve Bayes and KNN algorithm that how it works. How machine learning problems can be solved using these algorithms from scratch is explained [7]. The Author also explains about the customer views which is happening on the Online Market using different classifiers like Decision Tree, K-NN, Naïve Bayes. The result displays the accuracy in different classifiers like for Decision tree it is 80 %, for K-NN it is 78%, for Naïve Bayes it is 77% [8].

2.2 Related Work

Sentiment analysis is uses in various area to know the sentiment of people that what they are thinking about any subject to make a better decision. Any organization can make a good decision when they know about the people emotion and sentiment that what they are thinking about their product. Data collection is the first step for this project source can be anything in this project Twitter is the source and tweets will be collected for data. Data pre-processing is the important step after data collection. In this step various work will be done like removal of stop words, stemming which cuts the words into its root form, Lemmatization which is used to convert the word in root form of verb. After pre-processing of data feature extraction of data is very important to make the data appropriate format to use in algorithm to train the model which uses Bag of Words, TF-IDF. When feature extraction is done then training of model will be started using various algorithms like Naïve Bayes and Support Vector machine and then accuracy will be tested. To train the model one source and one target dataset is needed and then provide the data to predict the output. These all works are related to the steps involved in machine learning technique to solve the sentiment analysis problem and how can be found better output in efficient way. Data pre-processing is very important technique because if the collected data is not relevant to the model then model can be over trained or under trained which can produce wrong output. Neither any same type of data should be at the time of model training nor any unrelated data to the subject like in tweets URLs, hashtags, retweets. In feature extraction choosing the algorithm to format the data is very important because every algorithm work in different way so it needs the different data format. Selection of classification algorithm is the most important step in this project and it depends on many criteria like what type of data is going to be used, what type of output is going to be produced, size of data. Then the appropriate algorithm will be used to analyze the sentiments. Appropriate algorithm is used to train the model with appropriate data then accuracy of model will be good and predicts the result.

3. TOOLS AND TECHNOLOGIES REQUIRED

- **Python3:** Python supports libraries like matplotlib, pandas, numpy, scikit-learn which helps in storing and performing various operations on it. It is a language which has huge collection of libraries that will be used in this project. For development perspective in machine learning it provides many tools and libraries that is used in machine learning with better performance.
- **Tweepy:** It is a library supported in python which helps in collecting the data from twitter. Using this library developer can connect to the twitter API and can access the data according to the specific subject. It helps in collecting the tweets from twitter.
- **Scikit-learn:** It is a free software machine learning library for python. It supports various clustering and classification algorithms. It helps in many data preprocessing techniques also.

4. ARCHITECTURE DIAGRAM

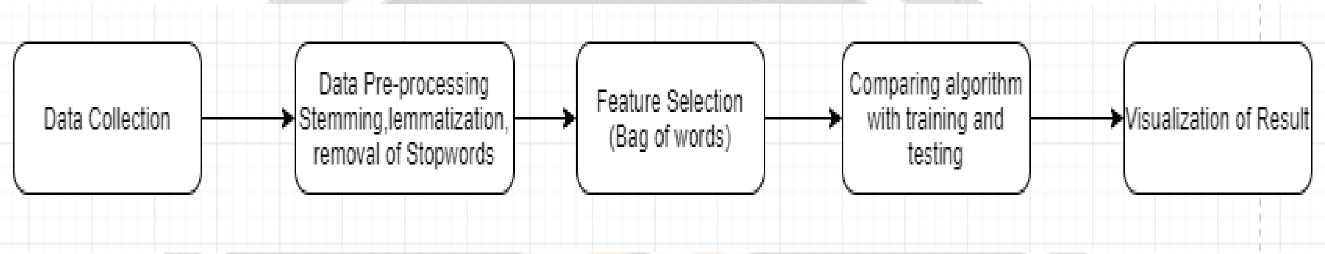
Data Collection
Data Pre-processing Stemming,lemmatization, removal of Stopwords
Feature Selection (Bag of words)
Comparing algorithm with training and testing
Visualization of Result

Fig 1 : Architecture Diagram

Fig 1 explains the architectural flow of the system in which it collects the data from twitter and pass it to the trained model then it predicts the result. This diagram explains the steps involved in it in which the first step is collection of data from the twitter. Data pre-processing is performed after data collection to make it able to process. Feature extraction is performed after cleaning in which pre-processed data will be passed to convert the data through bag-of-words. Then comparing the algorithm through training and testing process then result will be visualized.

5. COMPARITIVE STUDY

Table 1: Comparison of Learning Algorithm based on parameters (Accuracy, Type of data)

Learning algorithm	Parameter	Description
Multinomial Naïve Bayes classifier	Accuracy, types of data	This algorithm works on Bayesian theorem and the data will be independent in dataset and accuracy rate will be nearly 80%
Support Vector Machine	Accuracy, types of data	It is a linear regression model can solve linear and non-linear problem, it works with independent data in data is mapped one to one and accuracy rate will be to 80%

Decision tree classifier	Accuracy, types of data	It builds the classification model in the form of a tree like structure and make association rules from it. It works well with related data and the accuracy rate will be 75 % to 80%

Table 1 explains about the different Learning Algorithms like Decision Tree classifier, Support Vector Machine and Multinomial Naïve Bayes Classifier depending upon different parameters like Accuracy, type of data. The Description column explains briefly about the percentage of accuracy and also deeps down explains about the structure of how the data need to be, if these above-mentioned classifiers need to be applied in order to reach to the proper conclusion about the sentiment analysis of any subject.

6. CONCLUSION

Sentiment analysis is used to know the sentiment of people to make better decision. It categorizes the sentence in positive and negative. With the best suitable algorithm model will be trained with better accuracy so that it can predict the result. Accuracy rate is the most important aspects to analyze which is the best suitable algorithm for the sentiment analysis according to the dataset and types of result are needed.

7. REFERENCES

[1]. Akshi Kumar and Teeja Mary Sebastian," Sentiment Analysis on Twitter", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012, ISSN (Online): 1694-0814.

[2]. Tomohiro Fukuhara, Hiroshi Nakagawa, Toyoaki Nishida," Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events", ICWSM'2007 Boulder, Colorado, USA.

[3]. Songbo Tan, Xueqi Cheng, Yuefen Wang, Hongbo Xu, "Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis", ECIR 2009, LNCS 5478, pp. 337–349, 2009.

[4]. Alexander Pak, Patrick Paroubek," Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta.

[5]. Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena," Large Scale Sentiment Analysis for News and Blogs", ICWSM'2007 Boulder, Colorado, USA.

[6]. Tony Mullen and Nigel Collier," Sentiment analysis using support vector machines with diverse information sources", Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, July,2004.

[7]. Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas Beepa Bose Sweta Tiwari, "Sentiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier", International Journal of Information Engineering and Electronic Business 8(4):54-62 · July 2016.

[8]. Achmad Bayhaqy, Sfenrianto, Kaman Nainggolan, Emil R. Kaburuan," Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor and Naïve Bayes", 2018 International Conference on Orange Technologies (ICOT), 23-26 Oct. 2018.