International Journal of Current Engineering and Technology
©2021 INPRESSCO®, All Rights Reserved

E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
Available at http://inpressco.com/category/ijcet

Research Article

Twitter Sentiment Analysis using Machine Learning

Ms.Swarupa Kulkarni and Prof.Priyanka Kedar

Department of Computer Engineering, Dhole Patil College of Engineering Pune, India.

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

Development of web technology and its growth lead to a large volume of data present in the world wide web for users, also a lot of data is generated via sites present on the web. The Internet has become a platform for various things such as online learning, social networking, exchanging ideas, online marketing and sharing opinions. Social networking sites like Facebook, Google+, Twitter are rapidly becoming popular as they allow people to express their views about different topics, have a discussion within communities, independent of the location post messages across the world. Natural language processing, sentiment analysis these are continuously evolving fields, there is a vast scope of development in these fields. Sentiment analysis of Twitter data is mainly focused in this paper. On Twitter, information is present in form of tweets where opinions are heterogeneous, highly unstructured and are either negative or positive, or neutral in some cases. Various machine learning algorithms are used for sentiment analysis of tweets and their comparison is carried out in this paper.

Keywords: *Twitter, Sentiment Analysis, Machine Learning*

1. Introduction

Sentiment analysis can be explained as an automated process for mining of views, opinions, attitudes and emotions from tweets, speech, text, and database source through Natural Language Processing it is also known as opinion mining. Minor difference between Sentiment analysis and Opinion Mining is, opinion mining refers to extracting and analysing sentiments about particular entity while sentiment analysis refers to analysing sentient of present data. It is used to determine whether a piece of text is positive, negative or neutral. Now days many people use social networking sites to express their views about something. Many companies take polls or surveys to get feedback about newly launched products, sentiment analysis plays very important part to extract sentiment of users from gathered data. [6] Analysis can be done at various levels such as document, phrase and sentence. In document level analysis, entire document is analysed to check whether the sentiment of document is positive, negative or neutral. In phrase level, analysis of phrases present in a sentence is done to check polarity of given data. In sentence level analysis, each sentence is taken into account and later on classified into number of classes depending on its polarity. [7] The main goal of Sentiment Analysis is to make use of data in order to obtain important insights regarding public views, that would assist making smarter business decisions, better product consumption and political campaigns. Sentiment classification has number of application which is helpful in business, marketing and increasing sell of the product.

Significant research has been carried out to analyse informal texts and check their polarity with the help of lexicon approaches and Machine learning based approaches. As no lexicon or dictionary is capable of handling the amount of noisy content within data, they are out of scope for modern day Sentiment analysis. Machine learning based approaches are more suitable as they learn from this unstructured data and build a feature vector of their own, they do not depend on any predefined set of feature. These techniques are helpful in creating model for data with diverse contents. Feature extraction is main task in machine learning approaches as feature vector is built by using these features, which later on used by classifiers. In this paper we are doing sentence level analysis, different machine learning algorithms are used. The rest of this paper is ordered as follows: Section II presents literature survey listing some related work on sentiment analysis. The proposed methodology is explained in Section III. Experimental results and discussion is in Section IV. Finally, Section V concludes the paper.

2. Literature Survey

Alec Go et al.[1] has used machine learning algorithms like SVM, Naive Bayes and Maximum Entropy on Tweet data. Unigrams, bigrams, POS tags etc are the features used. They used emoticon features to build training sets. Using these machine learning algorithms on Twitter data they got maximum accuracy of 82.2% for unigrams.

Wan et al. [2] proposed Ensemble based Twitter classification method for the Airline domain. Majority vote was used as an ensemble. Total 5 base classifier algorithms were used. They were Naïve Bayes, Bayesian Network, SVM, Random Forest and C4.5 Decision Tree. C4.5 Decision Trees performed better than rest of the algorithms.

Li et al.[3] experimented AdaBoost with SVM and shown that it is able to perform better than Decision Trees and Neural Networks by making some adjustments in the kernel parameter. The proposed algorithm also gives good results on imbalanced datasets for classification problems.

Xia et al.[4] proposed an ensemble framework for sentiment classification. They used two types of feature sets with three base classifiers. Two types of feature sets are made using Part-of-speech data and Word-relations. Maximum Entropy , Naive Bayes and Support Vector Machines are selected as base classifiers in this framework. They applied ensemble algorithms such as Metaclassifier combination, Weighted combination and Fixed combination for sentiment classification which gave better accuracy.

Andreas Kanavos et al.[5] proposed a system which was created in Hadoop as well as Spark. Spark converts programs into an MapReduce job, it is an open source platform. Ankita

Gupta et al.[6] proposed an hybrid model using SVM and KNN classifiers. Ali Hasan et al.[7] proposed framework for tweets in Urdu language they pre-processed and classified tweets. For polarity calculations they used frameworks such as SentiWordNet, W-WSD and TextBlob.

3. Proposed Methodology

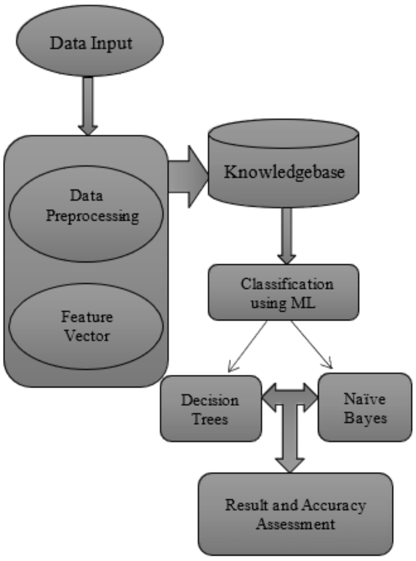

Figure 1 Proposed System for Sentiment Classification

The proposed system accepts the dataset which has tweets from real world, pre-processes it and provides it to machine learning classifiers generating results.

A. Data Gathering Tweets can be extracted from Twitter with help of API's provided by Twitter. TweePy is an python library which can be used to access Tweeter API. Data obtained from Twitter can be stored in CSV files. CSV separate each field with a comma, hence make it very easy to access the particular field which consists of text. CSV files also provide faster read/write operations as compared to others. [8] We have downloaded dataset from popular 'Kaggle' website. [12] Kaggle is an online community of data scientists and machine learners, it is repository of various datasets from number of domains. Dataset which we have used for our experimentation consists of real world tweets, they are labelled as positive or negative (sexist/racist) sentiment tweets.

B. Data Pre Processing: The Preprocessing stage is very crucial in fetching out important information from the informal and noisy text extracted in the form of tweets. Accuracy of model decreases if we provide noisy data to it, hence careful preprocessing is most important part in model designing. Following are steps which we took while preprocessing data: 1. Remove website URL's and links from text 2. Trim the repetitions like 'looool' to 'lol'. 3. Removing hashtags and mentions from text 4.

Converting the text to lower case 5. Remove extra white spaces and punctuation like '@' and '#'. [8]

We created method in Python for preprocessing task and passed our dataset to that method so that all the tweets from the dataset will be preprocessed in same manner.

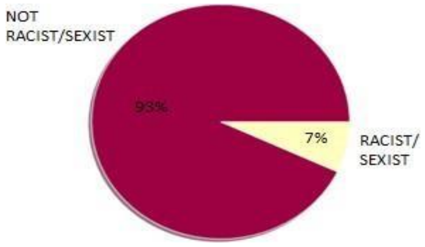

Figure 2 Tweets comparison based on labels

Like real world scenarios, dataset which we downloaded from Kaggle was imbalanced, number of positive tweets was much higher than number of negative tweets thus oversampling is carried out. SMOTE (Synthetic Minority Over-sampling Technique) function of 'imblearn' library is used for oversampling. SMOTE creates new minority instances between existing (real) minority instances.

C. Feature Vector Generation: Feature extraction is the process of extracting the more important feature for the task of Sentiment classification. The accuracy of the machine learning algorithm depends on feature set that is used. Stopwords are the most common words but they do not contribute in polarity of the statement hence they can be omitted. We can use 'tf-idf' scheme to give weights to term as per their importance. [7]

Term frequency TF(wi ,d) is nothing but count of a term (wi) in document (d). Large value of an term frequency denotes the importance of term respect to that particular document. Terms which are present in too many documents are suppressed as they are mostly stopwords. This suppression is handled by Inverse document frequency.

In other words, tf-idf scheme assigns a weight to term t which is present in document d. It has below possibilities:

1. Weight will be 'High' when an term t has multiple occurrences in small number of documents.
2. Weight will be 'Lower' when the term t has fewer occurrences in an document or it has occurrences in many documents.
3. Weight will be 'Lowest' when term t has occurrence in almost all documents.

D. Classification using ML: We have used Python programming language for our experimentation. Python language is very flexible and easy to understand. It provides many packages such as Pandas, Numpy which make data analysis very easy. To classify tweets into categories such as positive or negative, we used machine learning classifiers. 'scikit-learn' is popular library available in Python programming language, scikit-learn provides implementation of many machine learning algorithms for classification, regression and clustering. To install scikit-learn we used 'Python pip', it is an package management system. Below are the algorithms which we used for classification:

1. Logistic Regression

The name includes regression but logistic regression is an classification algorithm. While Linear regression algorithm is used to predict continuous values, Logistic Regression algorithm is used for classification problems such as ours.

Logistic regression provides better result in binary classification problem that is problem where there are only two target classes are present. Our experimentation also has only two target classes as we are classifying tweets as either 'positive' or 'negative'. h(x)= 1/ (1 + e^x) is called as an logistic function which is used as an transformation function in logistic regression algorithm. This function produces an S-shaped curve. The output generated by logistic function lies between 0 and 1. A threshold is applied on probability value generated by logistic function to finally convert it into binary classification.

2. Naïve Bayes

Bayes Theorem is used to calculate probability of an event when condition is that some other even is already occurred. When value of some variable is given we find out probability of outcome. [6] [7] To calculate probability of hypothesis being true when given prior knowledge, Bayes Theorem is given as:

P(h|d)= (P(d|h) * P(h)) / P(d)
Here,
P(h|d) = Probability that hypothesis will be true given data d also called as Posterior probability.
P(d|h) = Probability of (d) data given that hypothesis is true also called as Likelihood.
P(h) = Probability that hypothesis is true also called as Class prior probability.
P(d) = Probability of the (d) data also called as Predictor prior probability.
Algorithm is referred as 'naive' as it assumes that all the variables from problem are not dependant of each other, which is in turn a naive assumption in case of real-world examples.

3. Decision Tree

Decision trees are widely used in many fields such as operations research, to find a strategy to reach an specific goal, specially in decision analysis, they are also a common tool in machine learning. It has an structure like flow charts. In decision trees every internal node denotes a "test" on an attribute, every branch shows the outcome of that test, and leaf node shows a class label that is decision taken after computing all the attributes. Classification rules are represented by an path taken from root to leaf node. It is one of the widely used classification technique.

4. Result and Discussions

In this section we report the results obtained by using various machine learning algorithms for classification of tweets.

A. Dataset Used

As mentioned earlier we downloaded publicly available dataset from 'Kaggle'. Training dataset has total 31962 number of tweets. Out of total tweets 2242 number of tweets were labeled as racist/sexist tweets, which constituted only 7% of total tweets. Hence we carried out process of oversampling to make data normally distributed and eliminate bias. Tweets are classified into two classes positive and negative, we do not considered tweets with neutral sentiments. We studied that more cleaner the data more accurate results were obtained.

B. Accuracy Evaluation

In the field of machine learning mainly for the problem of statistical classification confusion matrix is used, also known as an error matrix. [11] A confusion matrix/error matrix is nothing but an table that is commonly used to describe the performance of a classification model on test data for which we know the true values.

Table 1 Confusion Matrix

	Predicted Positives	Predicted Negatives
Actual Positive	TP	FN
Actual Negative	FP	TN

Accuracy = (TP+TN)/(TP+TN+FP+FN)
Precision = TP/(TP+FP) Recall = TP/(TP+FN)
F1 Score = (2* Precision*Recall)/(Precision+ Recall) [11]

We have used two performance metrics, Accuracy and F1 Score out of these four.

Accuracy: Accuracy is one of the metric for evaluating performance of classification models. [10] Informally, accuracy is nothing but the fraction of predictions our classification model got right. Formally, we can define accuracy as following definition:

Accuracy = (Number of correct predictions) / (Number of total predictions) [10]

For binary classification problems such as ours accuracy can be calculated in terms of Positives and Negatives as shown in formula above. Accuracy alone is not sufficient to evaluate performance of model when we're working with a imbalanced data set, where there is a large difference between the number of negative and positive labels, hence we have also used F1 Score metrics.

F1 Score: F1 score is used to measure the accuracy of test. It is also called as F-score or F-measure. F1 score is harmonic average of both precision and recall. Best value for F1 score is 1 and worst value is 0.

Table 2 Algorithm performance for proposed system

Name of the Algorithm	Accuracy	F1 Score
Naive Bayes	83%	85%
Logistic Regression	89%	90%
Decision Tree Classifier	92%	92%

Conclusions

The proposed work is focused on sentiment analysis of tweets by using several machine learning algorithms. We have observed that clean data has large impact on results obtained. Decision Tree classifier outperformed Logistic Regression and Naive Bayes algorithms.

We can enhance proposed system by taking live data input from Twitter with help of Tweepy API. We can apply ensemble algorithms such as boosting to further improve performance of system. Also we can work in direction of making our system multilingual, so that it can handle tweets from different languages such as Hindi, Urdu etc.

References

[1]. Alec Go, Richa Bhayani, and Lei Huang, “Twitter sentiment classification using distant supervision”, Technical report, Stanford Digital Library Technologies Project, 2009.

[2]. Wan, Yun, and Qigang Gao. ”An ensemble sentiment classification system of twitter data for airline services analysis.” 2015 IEEE

[3]. International Conference on Data Mining Workshop (ICDMW). IEEE, 2015.

[4]. Li, Xuchun, Lei Wang, and Eric Sung. ”AdaBoost with SVM-based component classifiers.” Engineering Applications of Artificial Intelligence 21.5 (2008): 785-795.

[5]. R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classificationalgorithms for sentiment classification”, Information Sciences: an International Journal, vol. 181, no. 6, pp. 11381152, 2011.

[6]. Andreas Kanavos, Nikolaos Nodarakis, Spyros Sioutas, Athanasios

[7]. Tsakalidis, Dimitrios Tsolis and Giannis Tzimas,” Large Scale Implementations for Twitter Sentiment Classification”,

[8]. Multidisciplinary Digital Publishing Institute,2017.

[9]. Ankita Gupta, Jyotika Pruthi and Neha Sahu, ” Sentiment Analysis of Tweets using Machine Learning Approach”, International Journal of Computer Science and Mobile Computing, 2017.

[10]. Ali Hasan, Sana Moin, Ahmad Karim and Shahaboddin Shamshirband,

[11]. “Machine Learning-Based Sentiment Analysis for Twitter Accounts”, Multidisciplinary Digital Publishing Institute, 2018

[12]. Pulkit Garg, HimanshuGarg, VirenderRanga, “Sentiment Analysis of the Uri Terror Attack Using Twitter”, International Conference on Computing, Communication and Automation, 2017

[13]. https://scikit-learn.org/

[14]. https://developers.google.com/machine-learning/

[15]. https://machinelearningmastery.com/

[16]. https://kaggle.com/