International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 12 | Dec-2017 www.irjet.net p-ISSN: 2395-0072

Sentiment Analysis and Classification of Tweets Using Data Mining

Md Shoeb[1], Jawed Ahmed[2]

[1]*Student, Department of computer science, Hamdard University, New Delhi, India*
[2]*Assistant Professor, Department of computer science, Hamdard University, New Delhi, India*

Abstract - *These days, Social networking sites like twitter, facebook, etc. are the great source of communication for internet users. So these become an important source for understanding the opinions, views or emotions of people. In this paper, we use data mining techniques for the purpose of classification to perform sentiment analysis on the views people have shared on Twitter, which is one of the most used social networking sites nowadays. We collect dataset, i.e. tweets from Twitter and apply text mining techniques – transformation, tokenization, stemming etc to convert them into a useful form and then use it for building sentiment classifier. Rapid Miner tool is being used, that helps in building the classifier. Here, we are using three different classifiers on the data and then compare the results to find which one gives better accuracy and better results.*

***Key Words*: Rapid Miner, Classification, data mining, sentiment analysis**

1. INTRODUCTION

In recent times, people are using social networking sites like twitter, facebook, blogs for expressing their sentiments, views, feedbacks, opinions etc. and the opinions of other people have always been important to us in many ways. So, there comes a need to analyze their views and sentiments. Sentiment Analysis is the implementation of natural language processing, text analytics, and computational linguistics that assists in recognizing and extracting the useful information from the source matter[1]. It aims to ascertain the point of view of a speaker or a writer towards any topic or incident by analyzing their comments on social networking sites. Data mining also called knowledge discovery in databases that means the complete process of discovering the beneficial knowledge from data. It is the process of obtaining attractive and serviceable designs and relationships in large volumes of data[2].

Data classification is the process of classifying the data into some categories for its most efficacious and productive use. The goal of the classification is to predict the target class accurately for each and every case in the data. An algorithm that specially used to implements classification is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, that is implemented by a classification algorithm.

Text mining is the analysis of the data being used to extract the useful data from. It is used to process textual information and extract meaningful data from the text. Generally, some natural language processing or information retrieval methods or some pre-processing of text is done in order to make it useful for applying data mining algorithms.

In this work, we are using three different classifiers to extract the thoughts and sentiments of the people, they share on twitter through their tweets and classify them into different categories. And compare the results to find out which classifier gives the best result in terms of better precision and recall ratios and accuracy.

2. RELATED WORKS

This section contains a review of the work previously done in the field of sentiment analysis for the live data. A lot of work has been carried out till date in this field for the data from the users on social media in order to extract the sentiments of people towards any topic, products, trend etc. The studies focus on extracting useful information from the natural language of users and process it to get the real sentiments from the language. Osaimi and Badruddin[3] have done a lot of work on the sentiment analysis of the tweets on the twitter in the Arabic language. In this, they build different classifiers by training them with a proper dataset and then analyzed the accuracy and result of these classifiers in order to predict the correct sentiments. Pragya Tripathi, Santosh Kr Vishwakarma, Ajay Lala[4] have proposed the work on the sentiment analysis of English tweets using rapid minor. They collect the dataset from the twitter that is in natural language and applies the techniques of text mining and use it to build the sentiment classifier. O'Keefe et al.[5] have proposed a technique to select the features attributes weight and applied two classifiers on it i.e. Naïve Bayes and SVM. In this work, the author obtained classification accuracy of 87.15% by using only 29% of the selected attributes. Pak and Paroubek [6] have also worked in this field. The author used the data of Twitter to perform linguistic analysis and then build a classifier that is highly efficient. Pang and Lee[7] presented the broad overview of the existing work done by Pak and Paroubek. The authors describe the existing techniques and approaches for an information retrieval, in their survey.

K.Bhuvaneswari and R. Parimala[8] have proposed in their work, a method for sentiment classification using correlation-based feature selection. They applied different data pre-processing techniques, then used a correlation attribute method for feature selection, and then finally two classifiers namely Naïve Bayes and Support Vector Machine are implemented and results were evaluated. Farhan Laeeq, Md. Tabrez Nafis and Mirza Rahil Beg[9] have proposed a work on sentiment classification of social media. In their

work, they used three classifiers namely K-NN, Naive Bayes and Decision Tree Classifier for sentiment classification and obtained a result that shows the accuracy of K-NN, Naive Bayes, and Decision Tree classifier is 77.50%, 80%, and 78% respectively. Mangal Singh, Md.Tabrez Nafis and Neel Mani[10] have worked on Similarity Evaluation and Sentiment Analysis on the Reviews for Heterogeneous-Domain product. They demonstrated scaling and sentiment classification with similarity evaluation among the reviews on the product. And the Review data is pre-processed and cleaned for the data preprocessing. Mnahel Ahmed Ibrahim and Naomie Salim[11] have worked on the sentiment analysis of Arabic tweets extracted through Twitter, and then various classifiers like Naive Bayes, SVM, K-Nearest Neighbour are applied on data to find the best result. Eibe Frank and Albert Bifet[12] proposed challenges that Twitter data define, focused on classification problems, and consider for sentiment analysis. Isabella et al., [13] have proposed movie reviews for sentiment analysis and evaluated feature selectors to improve the performance of the classifiers.

3. METHODOLOGY

Here, the tool we used in this work is Rapid Miner[14]. Rapid Miner is an open source platform that used in the data science and developed by the company of the same name that provides an integrated environment for machine learning, data prep, text mining, model deployment, business analytics and predictive analytics. This tool offers more than 1000 drag-and-drop operators which can be used to perform data mining operations, easily and quickly. It is used for business and commercial purpose and also for education, research, training, rapid prototyping, and application development and model deployment. In this work, we will use some of the operators like text mining, classification, validations etc. For converting a natural language text into a useful form to obtain precise results we used operations on data like text preprocessing, data processing etc. Text Preprocessing People use asymmetry in their language while writing their opinion or reviews about anything on the social sites. Text in their tweets may contain uppercase words, emotion symbols, sometimes last letter repeating to add the degree of emotions in the word and some words that do not add any sentiment. So preprocessing is done to clean the irregularities from the text for further processing. Sentiment Vector Each word in the sentence has some emotion and that could be either positive or negative. Thus some sentiment vector is assigned to each word in training data having two attributes as strength and polarity. The polarity has two values positive and negative. Strength has some scale that has two values: 0, 1. Data processing. In this step, transformation and tokenization are done. Reviews may have more than one sentences and each and every sentence may express a different kind of emotion. Here each review is taken as an array of sentences and with the help of lexical analysis, each review is converted to the intermediate form. Tokenization is done to split the text data into tokens. Splitting points are defined using all special characters or non-letter characters. Then some filtration applied to reduce the length of token sets.

We use three most popular classifiers K-NN, Naïve Bayes, and Decision Tree for the purpose of classification. K-Nearest Neighbor (KNN) takes all the cases in the data and classifies that in new cases on the basis of similarity measures. It works on a distance metric, hence we need to define a metric point for measuring the distance between the query point and cases from the sample. A Naive Bayes Classifier belongs is the classifier based on Bayes' Theorem with strong independence assumptions between the features. A Naive Bayes classifier is highly scalable and requires linear parameters in the number of variables (predictors) in a training set. It is a conditional probability model. A Decision Tree Classifier uses a tree structure to classify the data. In Decision Tree, roots and internal nodes are assigned to a condition to be tested and all the terminal nodes contain a label Yes or No.

Figure 1 shows the fetching of data from the twitter by using search twitter operator and save the data in excel file by using write excel operator.

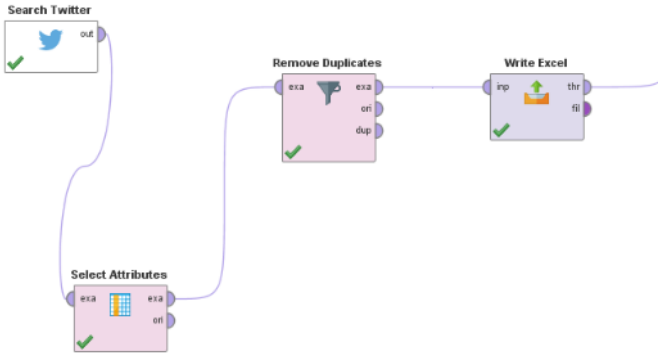

Figure 1 : Data Fetching

Figure 2 shows the main process. Read Excel operator used to read the data present in excel file. Process Documents operator is used to filter and transform the data. Validation operator used to provide training and to applying different data mining algorithms. In process document operator, we use transform, tokenize, filter and stem operators for preprocessing the data.

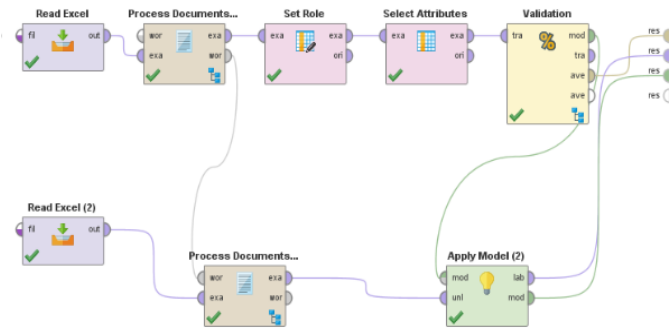

Figure 2 : Main Process

Figure 3 shows the sub-processes within the validation operator. In this, we used three different classifiers operators i.e. Decision Tree classifier operator, K-NN classifier operator and Naïve Bayes classifier operator.

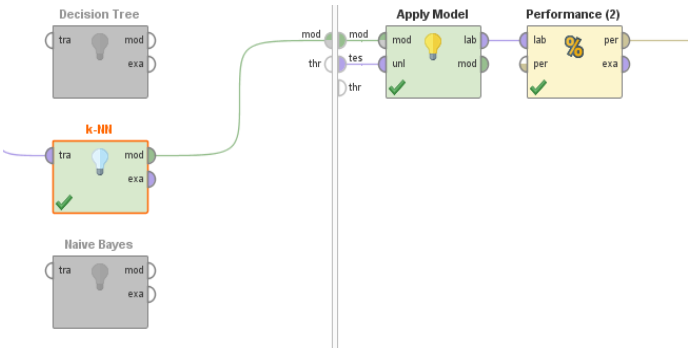

Figure 3 : Sub-Process validation operator

4. EXPERIMENTS AND PERFORMANCE ANALYSIS

In this section we describe the data that is used in this work, the tweets are collected from the twitter.com[15] social networking site which is in text natural language. We collected the tweets from Twitter. Then processed the text mining operators available in Rapid Miner on the data before applying to the classifiers for training and testing. For training purpose, we classify the tweets in two types of labels- positive and negative. These labels will be then used to train the classifier and based on this learning predict the label of the testing dataset. The dataset provided for the testing and based on the learning provided to the classifier. We are using here three different classifiers to predict the label. The results of prediction finally done by the classifiers are shown in below figures.

accuracy: 84.66%

	true positive	true negative	class precision
pred. positive	224	67	76.98%
pred. negative	8	190	95.96%
class recall	96.55%	73.93%	

Figure 4 : Decision Tree Accuracy

accuracy: 50.72%

	true positive	true negative	class precision
pred. positive	230	239	49.04%
pred. negative	2	18	90.00%
class recall	99.14%	7.00%	

Figure 5 : K-NN Accuracy

accuracy: 64.42%

	true positive	true negative	class precision
pred. positive	152	94	61.79%
pred. negative	80	163	67.08%
class recall	65.52%	63.42%	

Figure 6 : Naïve Bayes Accuracy

precision: 95.96% (positive class: negative)

	true positive	true negative	class precision
pred. positive	224	67	76.98%
pred. negative	8	190	95.96%
class recall	96.55%	73.93%	

Figure 7 : Decision Tree Precision

precision: 90.00% (positive class: negative)

	true positive	true negative	class precision
pred. positive	230	239	49.04%
pred. negative	2	18	90.00%
class recall	99.14%	7.00%	

Figure 8 : K-NN Precision

precision: 67.08% (positive class: negative)

	true positive	true negative	class precision
pred. positive	152	94	61.79%
pred. negative	80	163	67.08%
class recall	65.52%	63.42%	

Figure 9 : Naïve Bayes Precision

5. CONCLUSIONS

In this study, an attempt is made to classify sentiment analysis for tweets with the help of text mining and data mining techniques. We use three different classifiers – Decision Tree, K-NN, and NaïveBayes. All the three classifiers predicted the labels for a dataset. The result shows that the accuracy of Decision Tree, K-NN and NaïveBayes is 84.66%, 50.72%, and 64.42% respectively. The result also shows that the precision of Decision Tree, K-NN and NaïveBayes is 95.96%, 90.00%, and 67.08% respectively. We can see that Decision Tree classifier is the best classifier to be used with social media dataset as it gives the more accurate and precise prediction.

FUTURE WORK

In future, we aim to use the large and complex dataset and the number of labels can also be increased. We can include other languages also and use special characters and non-letter characters as well. It would be valuable to include the Emoticons as it widely used in social media to represent the expressions.

ACKNOWLEDGEMENT

I would like to thank my supervisor Mr. Jawed Ahmed (Assistant professor, department of Compute Science, Jamia Hamdard), for their guidance and inputs throughout this work. Last but not the least, I also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of the project, without which it would have been difficult to carry out this work.

REFERENCES

[1] Sentiment Analysis https://en.wikipedia.org/wiki/Sentiment

[2] Priyadharsini.C and Dr. Antony Selvadoss Thanamani, "An Overview of Knowledge Discovery Database and Data mining Techniques".

[3] Salha al Osaimi and Khan Muhammad Badruddin, Dept of Information System, Imam Muhammad ibn Saud Islamic University, KSA. "Sentiment Analysis of Arabic tweets Using RapidMiner."

[4] Pragya Tripathi, Santosh Kr Vishwakarma, and Ajay Lala, "Sentiment Analysis of English Tweet Using Rapidminer," in International Conference on Computational Intelligence and Communication Networks, 2015, pp. 668-672.

[5] O'Keefe. T and Koprinska I, "Feature Selection and Weighting in Sentiment Analysis," in Proceeding of 14th Australasian Document Computing Symposium, Dec 2009, Sydney, Australia

[6] Alexander Pak, Patrick Paroubek from Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS, Bˆatiment 508,F-91405 Orsay Cedex, France, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining."

[7] Bo Pang and Lillian Lee from Yahoo! Research, 701 First Avenue, Sunnyvale, CA 94089, USA, Computer Science Department, Cornell University, Ithaca, NY 14853, USA. "Opinion Mining and Sentiment Analysis."

[8] K. Bhuvaneshwari and R. Parimala, "Correlation Base Feature Selection for Movie Review Sentiment Classification," in IJARCCE, vol. 5, no. 7, July 2016.

[9] Farhan Laeeq, Md. Tabrez Nafis and Mirza Rahil Beg "Sentimental Classification of Social Media using Data Mining," in IJARCS.

[10] Mangal Singh, Md. Tabrez Nafis, and Neel Mani, "Sentiment Analysis and Similarity Evaluation for Heterogeneous-Domain Product Reviews," in IJCA, vol. 144, no. 2, June 2016.

[11] Mnahel Ahmed Ibrahim and Naomie Salim, "Sentiment Analysis of Arabic Tweets: With Special Reference Restaurant Tweets," in IJCST, vol. 4, no. 3, May – June 2016, pp. 173–179.

[12] Albert Bifet and Eibe Frank from University of Waikato, Hamilton, New Zealand, "Sentiment Knowledge Discovery in Twitter Streaming Data."

[13] J. Isabella & Dr. R.M.Suresh, " Analysis and Evaluation of Feature Selectors in Opinion Mining", Indian Journal of Computer Science and Engineering, (ISSN: 0976-5166), Dec 2012-Jan 2013, Vol. 3 No.

[14] https://rapidminer.com

[15] https://twitter.com