ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Sentimental Classification of Social Media using Data Mining

Farhan Laeeq
Department of CSE, SEST
Jamia Hamdard,
New Delhi, India

Md. Tabrez Nafis
Department of CSE, SEST
Jamia Hamdard,
New Delhi, India

Mirza Rahil Beg
Computer Centre,
Jamia Hamdard,
New Delhi, India

Abstract: In Today's life, social networking sites provides a great source of communication. So it provides the important source for understanding the emotions of people. In this paper, we use data mining techniques for the purpose of classification to perform sentiment analysis. We collect the facebook dataset and apply the optimized selection(Brute Force) and then we use three different classifiers and also compare their results in order to find which one gives better results.Rapid Miner tool is being used, which helps in building the classifier as well as able to apply it to the testing dataset.

Keywords: Sentiment Analysis, Data Mining, Facebook, Naive Bayes, Support Vector Machine, Decision Tree.

I. INTRODUCTION

Sentimental Analysis refers to the application of natural language processing, computational linguistics, and text analytics to identify and extract subjective information from the source materials. It aims to determine the attitude of a speaker or a writer towards any topic or incident. It is the computational study of people's opinion, appraisals, attitudes and emotions towards entities.

Data mining also called knowledge discovery in databases, or KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step. Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process[1]. Data classification is the process of arranging data into categories for it's most effective and efficient use[2]. Classification is a data mining function that assigns items in a collection to target categorises or classes. The goal of classification is to accurately predict the target class for each case in the data. An algorithm that implements classification especially in a concern implementation, is known as a classifer. The term classifier sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category[3].

In this work, we are using three different classifiers to find out which classifier gives the best result in terms of better precision and recall ratios and accuracy.

II. RELATED WORK

In the field of Sentiment Analysis, a lot of work has been already carried out. Pragya Tripathi, Santosh Kr Vishwakarma, Ajay Lala[4] have worked on the sentiment analysis of English Tweets using Rapidminer. They collect the dataset that are in natural language and applies text mining techniques and then use it to build sentiment classifier that is able to predict the tweet is happy, sad and neutral. Mnahel Ahmed Ibrahim, Naomie Salim[5] undertakes the sentiment analysis of Arabic tweets extracted through Twitter microalgae, and then various classifiers like Naive Bayes, SVM, K-Nearest Neighbour. O'Keefe et al.[6] proposed a new technique to select features attributes weight and applied Naïve Bayes and SVM classifier on it. In this, the author obtained classification accuracy of 87.15% by using only 29% of the selected attributes.

Osaimi and Badruddin[7] have proposed the work, on sentiment analysis of the tweets in Arabic Language. In this work they build different classifiers by training them with the proper dataset and then analyzed the accuracy of these classifiers in order to predict the correct sentiments. K.Bhuvaneswari and R. Parimala[8] proposed a method for sentiment classification using correlation based feature selection. They applied different levels of data pre-processing techniques, then correlation attribute method is used for feature selection, finally two popular classifiers namely Naïve Bayes and Support Vector Machine are implemented and performance measures were evaluated.

Syed Taha Owais, Md.Tabrez Nafis and Seema Khanna[9] proposed an improved method for detection of satire from User-Generated content. In this paper, author performed featured extraction of all articles in the training set and assigned weights to all the words that remain after pre-processing task.They found that the combination of SVM's with BNS feature scaling gives high precision. Mangal Singh, Md.Tabrez Nafis and Neel Mani[10] have proposed the worked on Sentiment Analysis and Similarity Evaluation for Heterogeneous-Domain product Reviews. They demonstrated sentiment classification and scaling with similarity evaluation among reviews. Review data is pre-processed and cleaned for

data preprocessing. Afshan Shujat, Md.Tabrez Nafis and Vishal Sharma[11], provides a solution to CoCos problem in Recommender System based on SNA. In this paper, they proposed a solution to address CoCos problem by exploiting the user social network information that was obtained by tracking the activities of the users.

Alok K Pathak and Md. Tabrez Nafis[12] examined the identification of influentials in the popular online social network, Twitter using Google‟s PageRank Algorithm [13] and optimized the calculation of ranking score by considering the dynamism of Retweet or Mention functionality.More the number of retweets by multiple users in the follower graph better is the user's influencing capability.

III. METHODOLOGY

The tool that is used in this work is Rapid Miner 7.4[14].

RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for machine learning, text mining and predictive analysis. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including data preparation, results visualization, validation, and optimization. In our work, we will use the operators of feature selection, classification, and validation etc. For reading an object from data repository we use the data access operators(In this case we use Retrieve operator).For Feature Selection, we use Optimize Selection(Brute Force) which gives the most relevant attributes of the given example set by trying all possible combinations. We use three most popular classifiers for the purpose of classification namely as- K-NN, Naïve Bayes and Decision Tree. A K-NN classifier stores all available cases and classifies new cases based on a similarity measure. K-NN has been used in statistical estimation and pattern recognition. K-NN are selected based a on distance metric.There are three choices to measure the distance, but the most popular choice to measure the distance is Euclidean.

$$d(x,y) = \sqrt{(x-y)^2}$$

Where x and y are the query point and a case of the example sample, respectively.

A Naive Bayes Classifier belongs to the class of simple probabilistic classifiers based on applying the Bayes' Theorem with strong (naive) independence assumptions between the features. A Naive Bayes classifier is highly scalable and requires a number of linear parameters in the number of variables(features/predictors) in a learning problem.Naive Bayes is a conditional probability model.

A Decision Tree Classifier uses a tree structure to organize a series of test questions and conditions.Decision Tree Classifier is widely used classification technique. To solve the classification problem it applies a straightforward idea. In Decision Tree, internal nodes and roots contain the condition to be tested while all the terminal nodes are assigned a label Yes or No.

Figure 1 shows the flow of the main process.Retrieve operator is used to reading an object from the data repository.Optimize Selection(Brute Force) operator is used to select most relevant attributes by trying all possible combinations.

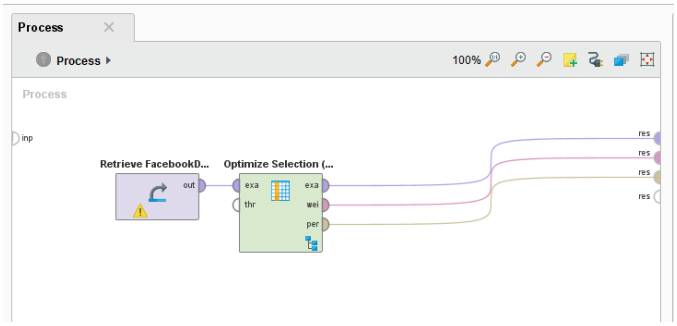

Figure 1: Main Process

Figure 2 shows the cross-validation operator is used in order to estimate the statistical performance of a learning operator. It is mainly used to estimate how accurately a model will perform in practice[15]. It will also return the labeled data if desired.

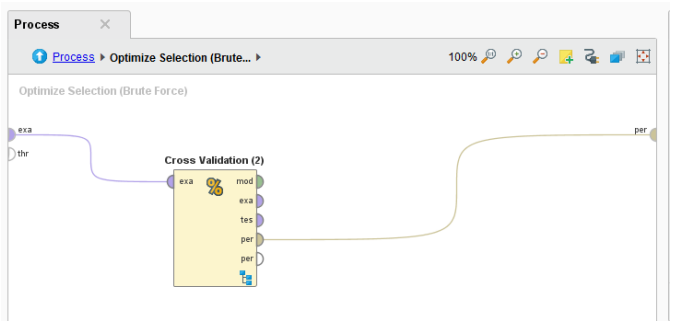

Figure 2: Optimize feature selection operator.

Figure 3, figure 4 and figure 5 shows the sub-processes within the Cross Validation operator. K-NN classifier, Naive Bayes classifier, and Decision Tree classifier are being used respectively.

Cross Validation operator consist of two sub-processes namely training and testing. The training set applies as an input to the classifiers which gives the model. Then the model can be applied on already learned or trained model on an Example Sets which gives labeled data as a output. Then this labeled data is passed as a input to the Performance Operator. The Performance operator is used for statistical performance evaluation of classification tasks[16].

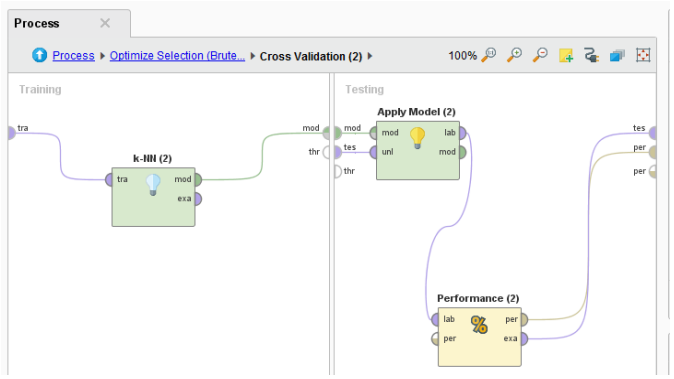

Figure 3: K-NN Classifier

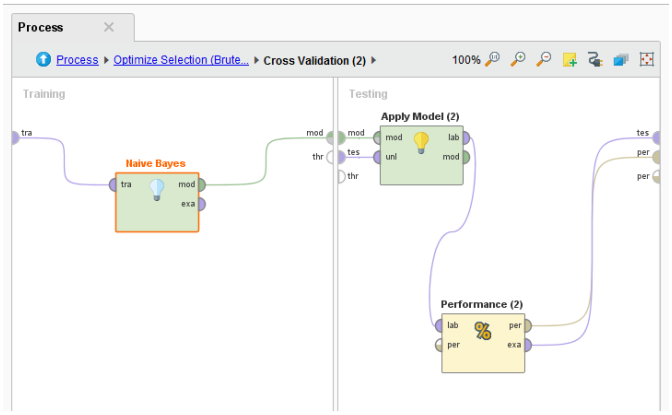

Figure 4: Naïve Bayes Classifier

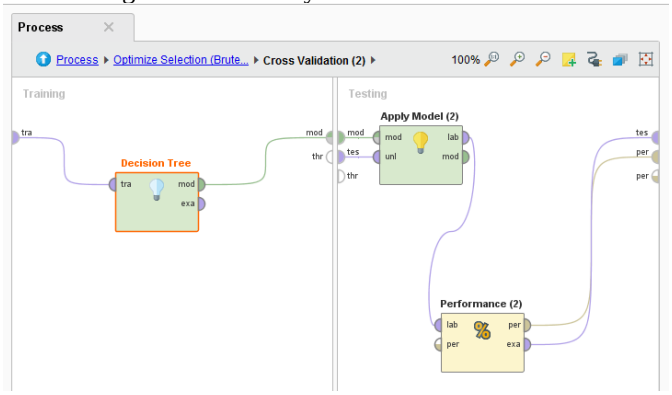

Figure 5: Decision Tree Classifier

IV. EXPERIMENTS AND PERFORMANCE ANALYSIS

In this work we use the facebook[17] dataset, this dataset can be generated manually or by using some data analytics tools.They are processed with the feature selection operators, like forward selection, backward elimination, optimize selection, etc. available in Rapid Miner.For this dataset, optimize selection feature select the most relevant attributes and try all combination, after that various classifiers are applied on the reduced dataset and their precision and recall values are compared.

Figure 6, figure 7 and figure 8 shows the attributes weightsin case of K-NN, Naïve Baye's and Decision Tree for the optimize selection (Brute Force) approach.

AttributeWeights (Optimize Selection (Brute Force))

attribute	weight
ID	0
Name	1
Date Joi...	1
Activity	0
No of Fol...	1
Profile Pi...	0
No of Fri...	1
Likes Co...	0
Comme...	0
Shares ...	0
Tag Count	0
Pokes C...	1

Figure 6: Attributes weight for K-NN Classifier

AttributeWeights (Optimize Selection (Brute Force))

attribute	weight
ID	0
Name	0
Date Joi...	1
Activity	0
No of Fol...	0
Profile Pi...	0
No of Fri...	1
Likes Co...	1
Comme...	1
Shares ...	1
Tag Count	0
Pokes C...	0

Figure 7: Attributes weight for Naive Bayes Classifier

AttributeWeights (Optimize Selection (Brute Force))

attribute	weight
ID	0
Name	0
Date Joi...	1
Activity	1
No of Fol...	0
Profile Pi...	0
No of Fri...	0
Likes Co...	1
Comme...	0
Shares ...	1
Tag Count	0
Pokes C...	1

Figure 8: Attributes weight for Decision Tree

Figure 9, figure 10 and figure 11 shows the confusion matrix for the K-NN, Naïve Bayes and Decision Tree classifier-the template will do that for you.

PerformanceVector

```
PerformanceVector:
accuracy: 77.50% +/- 18.87% (mi
ConfusionMatrix:
True:   Yes     No
Yes:    29      7
No:     4       9
```

Figure 9: Performance Vector

PerformanceVector

```
PerformanceVector:
accuracy: 80.00% +/- 17.89% (mikro: 79.59%)
ConfusionMatrix:
True:   Yes     No
Yes:    31      8
No:     2       8
```

Figure 10: Performance Vector

PerformanceVector

```
PerformanceVector:
accuracy: 78.00% +/- 10.77% (mikro: 77.55%)
ConfusionMatrix:
True:   Yes     No
Yes:    30      8
No:     3       8
```

Figure 11: Performance Vector

V. CONCLUSION

In this paper, sentiment classification of social media is done with the help of data mining techniques. We used three classifiers - K-NN, Naive Bayes and Decision Tree. The result shows that the accuracy of K-NN, Naive Bayes and Decision Tree that is 77.50%, 80% and 78%. We found that the best classifier to be used with social media dataset is Naive Bayes although Decision Tree also perform well.

We will extend this work by increasing the size of dataset. We also increase the number of labels and change the labels based on the needs.

VI. REFERENCES

[1] Fayyad, Piatetsky-Shapiro, and Smyth, "From Data Mining to Knowledge Discovery: An Overview," in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34.

[2] Data Classification, http://searchdatamanagement.techtarget.com/definition/data-classification.

[3] Statistical Classfication, https://en.m.wikipedia.org/wiki/Statistical_classification.

[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] 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.

[6] 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.

[7] 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."

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

[9] Syed Taha Owais, Md. Tabrez Nafis, and Seema Khanna,"An Improved Method for Detection of Satire from User-Generated Content," in IJCSIT, vol. 6, no. 3, 2015.

[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] Afshan Shujat, Md. Tabrez Nafis, and Vishal Sharma, "A Solution to CoCos Problem in Recommender System based on SNA," in IJCA, vol. 144, no. 3, June 2016.

[12] Alok K Pathak and Md. Tabrez Nafis, "To Find Influential's in Twitter based Information Propagation," in IJCA, vol. 118, no. 13, May 2015.

[13] The original PageRank paper by Google founders Sergey Brin and Lawrence Page - http://wwwdb.stanford.edu/~backrub/google.html.

[14] RapidMiner, https://rapidminer.com.

[15] Wikipedia, Cross Validation, https://en.wikipedia.org/wiki/Cross-validation_(statistics).

[16] Performane Operator, http://docs.rapidminer.com/studio/operators/validation/performance/predictive/performance_classification.html.

[17] Facebook, https://www.facebook.com/flaeeq.