ISSN No. 0976-5697

Volume 11, Special Issue I, May 2020

International Journal of Advanced Research in Computer Science

CONFERENCE PAPER

Available Online at www.ijarcs.info

CONTENT BASED TWEET CLASSIFICATION ON TWITTER

Prof. L A Lalitha
School of Computing and Information Technology
Reva University
Bengaluru, India
lalithala@reva.edu.in

Sumitha P
School of Computing and Information Technology
Reva University
Bengaluru, India
R16CS423@cit.reva.edu.in

P. Snavaja Krishnan
School of Computing and Information Technology
Reva University
Bengaluru, India
R16CS402@cit.reva.edu.in

Sushmita S
School of Computing and Information Technology
Reva University
Bengaluru, India
R16CS425@cit.reva.edu.in

Vinaya R M
School of Computing and Information Technology
Reva University
Bengaluru, India
R16CS538@cit.reva.edu.in

Abstract: Today, Social Media Networks are more powerful and popular than any other forms of media that exist and due to this global nature of social media, the amount of information available and being shared online by the users is tremendous. This large data that is available can be used for different purposes like marketing, data analysis, community detection, fraud detection, sentiment analysis, etc. In this work, we present a model to classify tweets in Twitter and therefore offer a solution to process large amounts of data and derive meaningful conclusions from the same. Here, we first collect tweets from different communities on twitter and process this raw dataset. This processed data is then converted into a vector form so that the textual information is converted to a numeric form for the machine to implement and then a text classification algorithm is applied to this dataset. Finally, after training the machine using this dataset, the working of the model and its accuracy is evaluated by using a dataset of test tweets where the machine predicts the category to which the test tweet belongs. With this model, we have been able to classify tweets into different categories and have achieved satisfactory results.

Keywords: Content-Based Classification, social network, Feature Extraction, text processing, Random Forest

I. INTRODUCTION

Ever since the creation of the first social media site in 1997, the number and popularity of social media has been increasing exponentially. Today, in the 21st century, we have a staggering 2.65 billion people all over the world who are active social media users. With this huge number we can very well imagine the massive amount of data being shared online in the form of images, texts, audios, videos and other activity logs. Twitter, among other popular Online Social Media (OSM) sites like Facebook, Instagram, WhatsApp, etc. [11] is a social networking service that is used for microblogging purposes. Here, the messages are called “tweets” and the Twitter users can post and interact with these tweets. As of 2019, Twitter has about [13] 330 million monthly active users of which more than 40 percent use the service on a daily basis. With a tweet length of maximum 280 characters, [11] the other features of Twitter include hashtags- to let users generate a tag to help other users easily find messages with a specific theme or content, retweets- to repost a tweet, usernames- prefixed with a @ symbol, followers and following count, etc.

With the data available from Twitt**er** consisting of the above-mentioned features, it is possible to conduct a variety of studies and analysis to find patterns and predictions in a OSM network. In large social networks like these where a huge number of tweets regarding various topics are posted every second, it would be very useful to classify these tweets into common categories. Even in the case of trending tweets, if the tweet is recognized as belonging to a category, it would be easier for other twitter users and data analysts to follow or study the tweet. This is the area of our interest. Several works in the past have proposed methodologies to conduct tweet classification in Twitter for purposes like sentiment analysis, spam detection etc. Unlike the aforementioned studies we propose to classify tweets into different categories – religion, sports, climate, politics, music, etc.- such that the study can be further enhanced in the future to classify any tweet on Twitter into common popular categories. We have used the Bag of Words algorithm in our model to extract textual features from the dataset and to then train the machine to classify them into different categories.

In the next section we review the literature survey conducted for this paper, which is followed by the objective of our work in Section III. The methodology used and the modules identified for the same are presented in the following

Sections IV and V. Finally, we discuss our findings and conclude the paper in Section VII.

II. LITERATURE SURVEY

Mohammad Mohaiminul Islam and Naznin Sultana in [1] have investigated the different ways of sentiment analysis by tweet classification from customers' review using 6 different machine learning algorithms. The dataset they have used are two public datasets from IMDB movie review site and Amazon book review site. In the results, the Linear SVM approach was found to be better for sentiment classification than the others with the selection parameter value c= 0.25. They also propose to implement the same technique for datasets from other domains like finance, politics etc. and observe the variation of accuracy to the variations of datasets.

In [2], Nicolas Tsapatsoulis and Constantinos Djouvas, have proposed a system where tokens used by humans are considered the best feature set when compared to automatically extracted features that can be used in tweet classification. They have used a manually collected dataset of 507263 tweets tagged with the hashtags #Greferendum, #dimopsifisma, #OXI, #NAI etc. A result of 88.74% classification performance was achieved using Decision Tree classifier algorithm and propose to improve on the classification performance using a combination of feature sets by a selected weighting scheme.

In [3], Hatma Suryotrisongko, Oky Suryadi, Achmad Farhan Mustaqim and Aris Tjahyanto propose text mining using Naïve Bayes classifier for classifying tweet automatically so that public feedback to the government happens much faster. They have used a dataset of 3000 tweets regarding feedback for Pemerintah Kota Surabaya (Surabaya's City Government) that contained both complaint or non-complaint tweets. Using the Naïve Bayes algorithm, they could classify tweets with an average accuracy of 82.5%.

Syed Muzamil Basha, K. Bagyalakshmi, C. Ramesh, Robbi Rahim, R. Manikandan and Ambeshwar Kumar, in [4] have discussed the methods to discover the best machine learning algorithm for text classification using Document Term Vector representation method. Two data sets - HumanAids and Mouse Cancer- with 150 instances each were considered for the study. Using evaluation metrics like recall, precision and fscore, the Support Vector Machine algorithm was found to perform better than other algorithms. In future they propose to work on different datasets using designed workflow for validation [4].

In [5] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, et al. have discussed the method to classify trending topics on Twitter using the Bag of Words model followed by a network-based model. This classification is done into 18 different categories. They have used a database of 768 randomly selected trending topic from Twitter. They have achieved a 65% accuracy for the text-based model and a 70% accuracy for the network-based model. In future, they propose to integrate text-based classification using Naive Bayes Multinomial and network-based classification to achieve better results.

Heba M. Ismail, Saad Harous and Boumediene Belkhouche, in [6] have proposed a system where four different variations of an input dataset are trained by different classification algorithms like Multinomial NB, Bernouli NB, and Support Vector Machine and their performances are compared. The dataset used consisted of manually collected tweets from the Stanford Twitter Sentiment Data dataset and Multinomial NB was found to perform better than other classifiers for sentiment analysis in this case. As a future enhancement, they propose to implement the model on larger and more representative datasets.

III. METHODOLOGY

In this section we describe the methodology followed to implement the proposed system. The programming language used here is Python. As shown in Figure. 1 the process mainly consists of 3 main steps: dataset creation, implementation and training the model, and testing.

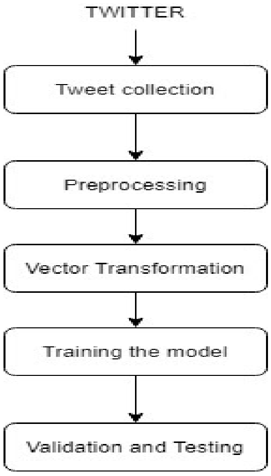


Figure 1. Implementation flowchart

A. *Dataset Creation and Preprocessing*

The first step of dataset creation involves the retrieval of tweets from Twitter. To do this, we need to have a valid Twitter account and create a new application in the Twitter developer section to obtain a consumer and access token. Using these tokens, we can now authorize our app to obtain tweets from Twitter with the help of OAuth Interface from the tweepy library. We have retrieved 5000 tweets each from the categories of politics, religion, sports, climate change and music. These tweets are then preprocessed by removing special characters, symbols, numbers, stop words, uppercase characters etc. and stored into a csv file. The final dataset in the csv file contains the list of tweets along with the category to which each tweet belongs.

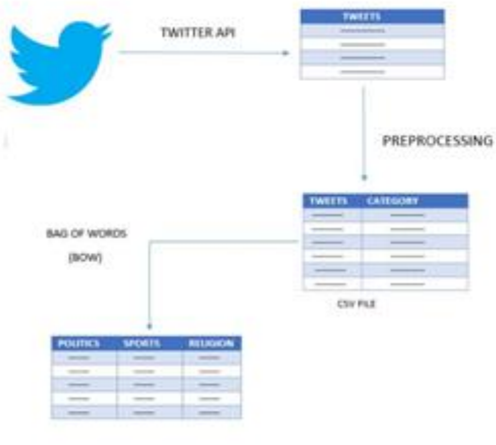


Figure 2. Dataset creation

B. *Training and Building the model*

Text documents cannot be processed by learning algorithms and classifiers in the original raw text format. These learning algorithms expect the data to be in a numerical format. Hence, we have used the bag of words model to implement this transformation. In this model, the text is represented as a bag/collection of words, disregarding grammar and word order but retaining the frequency of each word. Using the sklearn. feature_extraction library we calculate the term frequency vector for each tweet, thereby extracting its features. After having this vector representation of the text, the dataset is divided into training data (70%) and testing data (30%). We can now train the data using different classification algorithms [11]. After fitting the training dataset, the model can now be used to predict the category for test tweets. Here, we have used the Naive Byers, Decision Tree, Random Forest and KNN algorithms to compare and choose the most efficient algorithm.

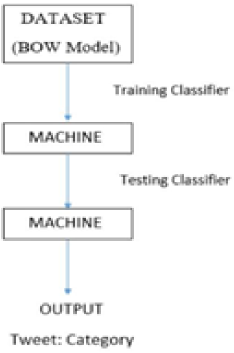


Figure 3. Training and Testing

C. *Validation and Testing*

After the training and building the model, now the performance of the model is tested using precision and accuracy scores. The comparative study of these scores including fscore and recall for each category or the whole dataset will help determine the most efficient algorithm and technique that can be used. Since we have used a 5-class classification, we get a report for the classes politics, religion, sports, climate change and music. In this way, the machine learning model to implement tweet classification in Twitter is implemented.

IV. MODULES IDENTIFIED

The following modules were identified in the implementation of this system. The corresponding packages and functions used in each module is given below:

a) ***Tweet retreival****: tweepy package, Twitter API, OuthHandler, Cursor*

b) ***Preprocessing****: pandas and nltk package, sentence and word tokenizers, re package to remove symbols and stopwords*

c) ***Feature Extraction****: scikit-learn, sklearn and numpy packages, TfidfVectorizer, chi2*

d) ***Training and Testing:*** *train_test_split, CountVectoriser, TfidTransformer, MultinomialN*

V. RESULT DISCUSSION

The performance of a classification algorithm can be measured in different ways using different performance evaluation metrics like confusion matrix, accuracy, AUC, etc. The metrics we use for this testing and comparison of the performance of different machine learning algorithms should be chosen very wisely. Here, the algorithms- Random forest, Decision Tree, KNN, AdaBoost, SGD and Naïve Bayes are compared using the following metrics. Accuracy, precision- the ratio of true positives to total positives, recall- the ratio of true postives to total actual positives and f1 score- a function of precision and recall

From Table I. it can be observed that the Random Forest classifier has the highest accuracy, precision, f1 score and recall value. Decision Tree algorithm has the second highest value for all the metrics, followed by other algorithms. The graph for all the algorithms represented with their corresponding performance metric scores are shown in Fig .4-7. From the results it can be inferred that Random Forest algorithm is best suited for tweet classification into 5 categories with an average accuracy of 85%.

Table I. Performance Metrics of Algorithms

Algorithm	***Accuracy score***	***Precision score***	***Recall score***	***F1 score***
Stochastic Gradient Descent	65.03%	74.33%	60.22%	60.99%
Random Forest	85.74%	87.47%	84.22%	84.35%
Decision Tree	80.76%	79.41%	79.19%	79.18%
Adaboost	70.25%	69.95%	69.18%	69.44%
Gaussian Naïve Bayes	46.80%	66.29%	50.34%	48.84%
K Nearest Neighbour	76.94%	75.94%	76.20%	75.89%

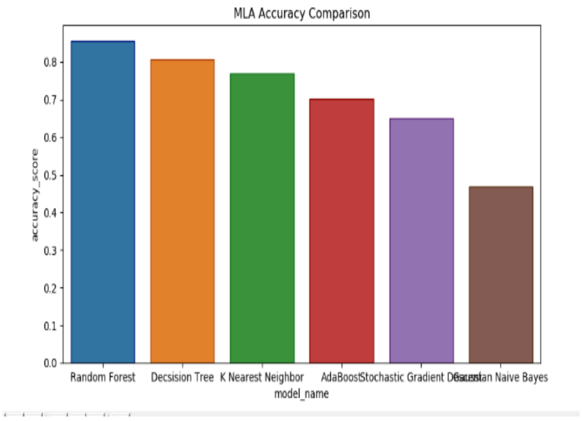


Fig 4. Accuracy Comparison

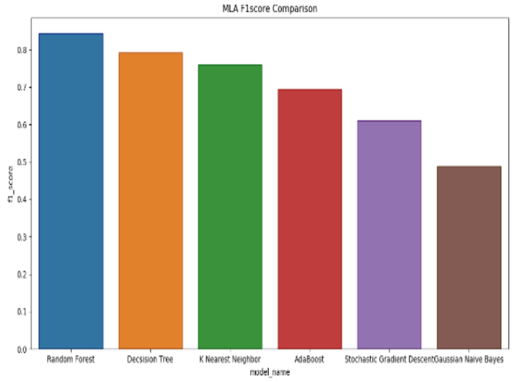


Fig 5. F1 score comparison

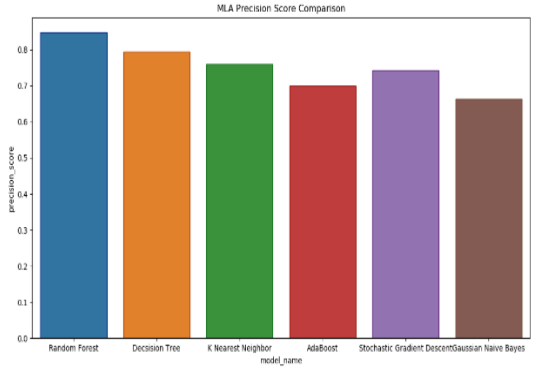


Fig 6. Precision score comparison

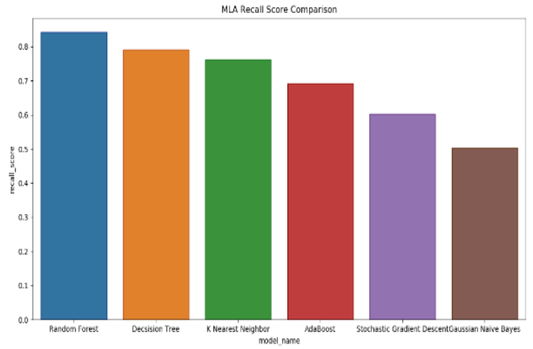
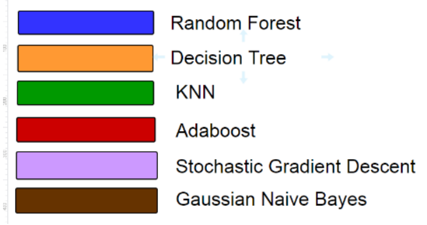


Fig 7. Recall score comparison



VI. CONCLUSION AND FUTURE ENHANCEMENT

In this paper, we have discussed the proposal of using a supervised machine learning technique to classify tweets into 5 categories-sports, politics, religion, music and climate change- in Twitter. We have used the Bag of Words algorithm and with the help of different performance evaluation techniques have concluded that Random Forest classifier is best suited for this model. We also noticed that the accuracy increases as the training dataset size increases. In the future, enhancement of this model can be done by using a bigger dataset for more accurate results. More classes like movies, literature, health, news, etc. can be added to the model so that any tweet in Twitter can be classified as belonging to a particular category. This will also help categorize trending tweets, celebrity tweets, etc. and help users decide the content they want to follow. The accuracy and efficiency of these different algorithms can again be measured using different performance metrics. The results can then be compared and the most suitable one can be implemented as an enhancement.

VII. REFERENCES

[1] Mohammad Mohaiminul Islam, Naznin Sultana, "Comparitive Study on Machine Learning Algorithms for Sentiment Classification", International Journal of Computer Applications, Volume 182-No.21, October 2018

[2] Nicolas Tsapatsoulis, Constantinos Djouvas,"Feature Extraction for Tweet Classification: Do the Humans Perform Better",2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization

[3] Hatma Suryotrisongko, Oky Suryadi, Achmad Farhan Mustaqim, Aris Tjahyanto,"Classification of Citizen Tweets Using Naive Bayes Classifier for Predictive Public Complaints",2018, IEEE 3rd International Conference on Communication and Information Systems

[4] Syed Muzamil Basha, K Bhagyalakshmi, C Ramesh, Robbi Rahim, R Manikandan, Ambeshwar Kumar,"Comparative Study on Performance of Document Classification Using Supervised Machine Learning Algorithms: KNIME",2019,International Journal on Emerging Technologies 10(1): 148-153

[5] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary, "Twitter Trending Topic Classification", 2011,11th IEEE International Conference on Data Mining Workshops

[6] Heba M.Ismail, Saad Harous, Boumediene Belkhouche,"A Comparative Analysis of Machine Learning Classifiers for Twitter Sentiment Analysis",2016,17th International Conference on Intelligent Text Processing and Computational Linguistics

[7] ZHAO Xianghui, PENG yong, YAO Yuangang, WANG Xiaoyi, ZHENG Zhan, "A Classification Method to Detect if a Tweet Will be Popular in a Very Early Stage", 2015, International Conference on Computing, Communication and Security

[8] Jaka E. Sembodo, Erwin B. Setiawan, Moch Arif Bijaksana,"Automatic Tweet Classification based on News Category in Indonesian Language",2018, 6th International Conference on Information and Communication Technology

[9] Hasnat Ahmed,Muhammad Asif Razzaq, Ali Mustafa Qamar, "Prediction of Popular Tweets using Similarity Learning", 2013, IEEE 9th International Conference on Emerging Technologies

[10] Liza Wikarsa, Sherly Novianti Thahir, "A Text Mining Application of Emotion Classifications of Twitter's users using Naive Bayes Method",2015,1st International Conference on Wireless and Telematics

[11] Wikipedia www.wikipedia.org/Twitter

[12] Twitter, www.twitter.com

[13] Geekforgeeks,www.geeksforgeeks.org/bag-of-words-bow-model-in-nlp/

[14] https://www.oberlo.in/blog/twitter-statistics