Sentiment Recognition in Tweets

Nisha Mansuri (30130025)
Schulich School of Engineering
University of Calgary
Calgary, AB
nishasajidahmed.mans@ucalgary.ca

Mitalee Khanna (30123162)
Schulich School of Engineering
University of Calgary
Calgary, AB
mitalee.khanna@ucalgary.ca

***Abstract*—Our primary goal is to classify the sentiment behind the tweet using a powerful classification algorithm. Sentiment analysis is a part of Machine learning and data mining which takes data and converts it into useful information. The fact that we can make estimations, predictions and give the ability to machines to learn by themselves is both powerful and limitless in terms of application possibilities.**

Keywords—Sentimental Analysis, Natural Language Processing, data preprocessing, Classification Algorithms

I. Introduction

Twitter is the most known microblogging website in the world, which allows anyone to express their thoughts, opinions, and more in just 140 characters. Due to the character limit of tweets and the time it takes to send one, tweets tend to reflect what people feel in real-time. Every event, news, or activity around the world is shared, discussed, and posted on social media by millions of people. E.g., “Friends is the best comedy show ever” or “Delicious dinner at Copper Chimney! :D” or “OMG! That is so scary!”

There is an enormous number of user data generated on social media platforms that can be used to analyze user behavior in terms of their emotions, opinions, and attitudes toward certain events. This information can be helpful for brands while developing new products.

Even though we have a large amount of user data, converting this data into useful information is a critical task. One such example is finding the sentiment of the user behind his/her tweet. There are many classification algorithms like Naive Bayes, Random Forest, Logistic Regression, support vector machine, Decision Tree Classifier, Gradient Boosting Machine, etc., available that use machine learning for recognizing the sentiment behind sentences.

II. Related Work

In general, the sentimental analysis of the data contains two steps: preprocessing and classification. The subjective information can be extracted from a document, preprocessed it using NLP (Natural Language Processing), and try to classify it as positive, neutral, or negative according to its polarity. There are various kinds of classification models used in sentiment analysis to classify a sentence as positive or negative: SVM, Naive Bayes, Maximum Entropy, Decision Tree, Random Forest, XGBoost, Convolutional Neural Networks.

Some of the Research mentions the hybrid approach to recognize the sentiments behind the tweets. Some research mentions the combination of machine learning and deep learning.[3]. While other mentions the combined approach between the classification algorithms which produces better accuracy than the individual ones.[1] The reference [3] outlines the approach of using Logistic Regression and Stochastic Gradient Descent classifier to recognize the sentiments behind the tweets. The “Real-time emotion Recognition of Twitter posts using a hybrid approach” uses the machine learning approach and deep learning approach together to classify the Twitter data.

III. Approach

The main purpose of the project is to compare the effectiveness of different classification algorithms with the assembled voting classification algorithm and check if the assembled classification works better than the individual ones on Twitter data to recognize the Sentiment behind it. The data preprocessing and converting them into a useful form is a very important part while applying classifying algorithms. As the accuracy of the data directly affects the accuracy of the algorithm. So, this project basically contains three steps:

1. Data preprocessing

2. Classification using various algorithms

3. Comparison between ensemble algorithm and the individual ones.

Contribution

This document contains a detailed solution for creating a fully annotated dataset. The suggested method may be utilised as the basis for training in a wide range of sentiment analysis tasks. The provided solution includes the following functions, which are detailed below:

- Data pre-processing using the natural language processing, unique frequency distance matrix and positive, negative word directory.
- A trained dataset which is classified into positive and negative polarity tweet using various classification algorithms.
- A comparison between the stand-alone classification algorithms - Naive bayes, Random Forest algorithm and Logistic Regression with the ensemble voting classifier using the accuracy rate.

A. Data preprocessing

The Sentiment140 dataset is used for training and testing purposes. The preprocessing tasks include

1. Reform the data of Sentiment140 into the format that can be easily fed to various classifiers.

2. Convert various emojis into negative and positive emojis

3. Identify user mentioned and URL string present in the tweets and convert them into the string data.

4. Strip space, dots, and other irrelevant symbols from the tweet.

5. Identifying words from the tweet.

We are generating statistical information about the preprocessed dataset. We are creating the frequency distance files for recognizing the frequency of URLs, emojis, words, etc. Sometimes in Natural Language processing, it is hard for a machine to recognize the sentiment behind the two words that contain the positive word but have a negative sentiment. For example, “I am not happy with today’s weather.” Here “happy” can be classified as a positive sentiment. Meanwhile, because of the “not” word, this statement should be classified as a negative statement. To overcome this problem, we are also using biagram as a part of Natural language processing. The frequency distance vector has been also created for the biagram words. Finally, The various classification models are applied to the dataset.

B. Classification Algorithms

In this project, the main three classification algorithms have been applied to classify the tweets into positive and negative. And after that the comparison between assembled classifier and the stand-alone classifier is done using the validation and the testing accuracy. Assembled classifier is basically the combination of the three classifiers- Naïve Bayes, Random Forest, and the logistic Regression. All the classification algorithms are mentioned below in the details

- Naive Byes

  The Naive Bayes algorithm is used to perform the classification of the feature vector from the dataset. The algorithm performs under the ‘naive’ assumption that features that contribute towards the labels of the dataset are independent of other features.

  The Gaussian Naïve Byes is used for producing the output. The equation used for the prediction is mentioned below

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Figure 1 Gaussian Naive Byes Equation [8]

  The testing accuracy and validation accuracy using this algorithm is 66.99% and 61.54% respectively with the Sentiment140 dataset. According to the observation, it can be seen that the Naïve Byes algorithm is not generalizing properly and overfitting the dataset, which is evident from the differences between testing and validation accuracy.

- Random Forest

  Random forest is a supervised learning algorithm that makes use of multiple decision trees to perform the classification. This collection of multiple decision trees or decision forest works in an ensemble by producing classification labels based on outputs of each decision tree. The label produced by the greatest number of trees is used as a model's prediction.

  This algorithm’s testing accuracy is 74.08% and the validation accuracy is 72.04% with the Sentiment140 dataset.

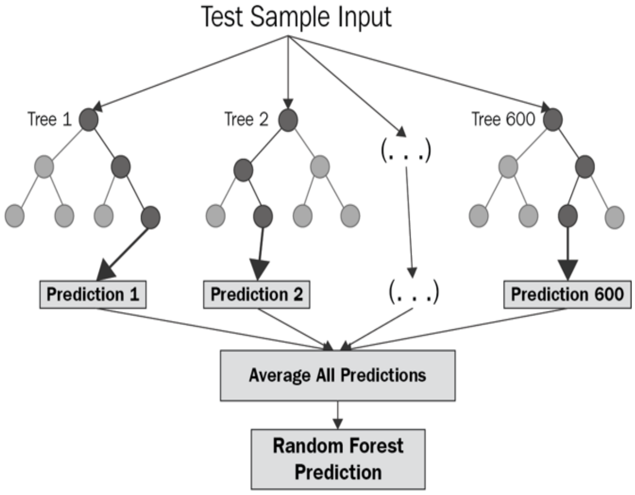


Figure 2 Visualization of Random Forest algorithm [6]

- Logistic Regression

  Logistic Regression is used to predict the probability of the targeted label. It uses a logistic function for prediction, also known as a sigmoid function. It is somehow similar to the Linear Regression. But it uses a more complex cost function to predict the output.

  The testing and validation accuracy using this algorithm is 74.28% and 76.62% respectively with the Sentiment140 dataset, which is quite close to the accuracy of the Random Forest algorithm.

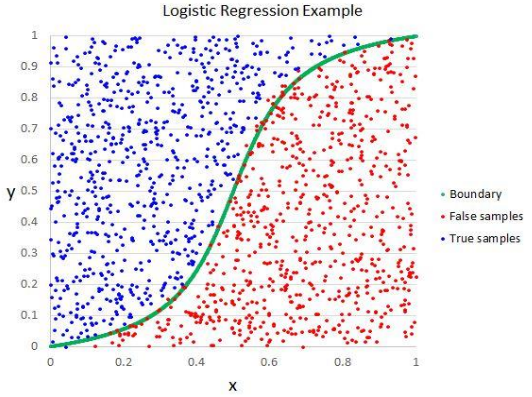


Figure 3 Visualization of Logistic regression [7]

- Sentimental Recognition using voting classifier

  A voting Classifier is an ensemble learning algorithm which combines outputs of various machine learning models and generates a prediction having the most probability among all. To address the problem of sentiment identification of tweet we have employed hard voting ensemble approach containing naïve Bayes, random forest, and logistic regression algorithms.

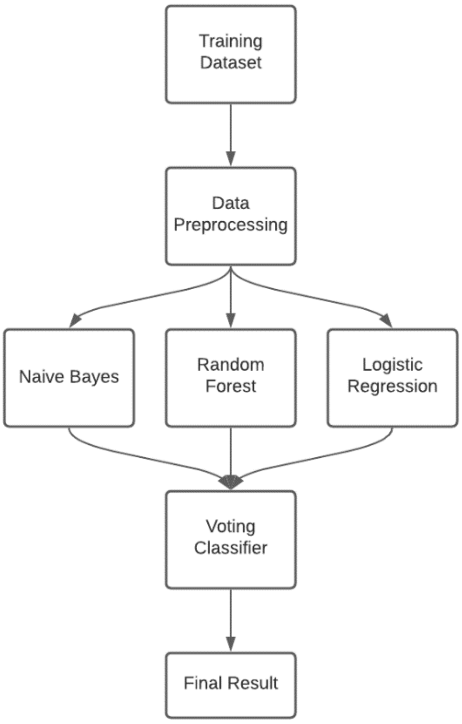


Figure 4 Visualization of voting classifier

In the hard voting approach voting classifier looks at the number of models that gave positive sentiment and number of models that predicted negative sentiment. After that the sentiment having the greatest number of votes is produced as the output of the voting classifier. Using this approach allows us to take advantage of generalizing the predictions by using various models and avoid false positives produced by a single model.

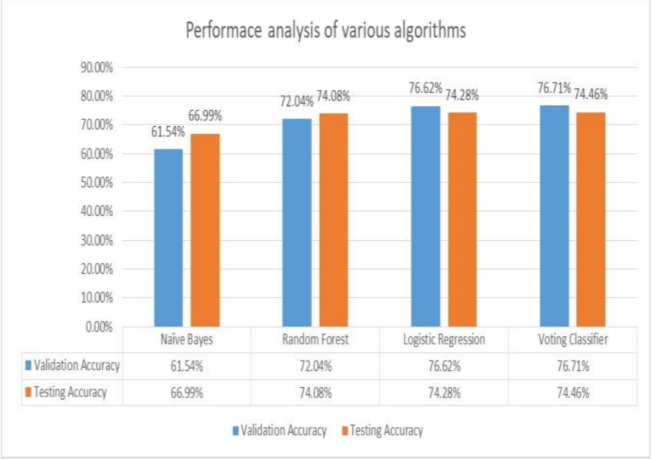
C. *Figures and Tables*

TABLE I. RESULTS (TESTING ACCURACY)

Algorithm	***Testing Accuracy***
Naïve Bayes	66.99%
Random Forest	74.08%
Logistic Regression	74.28%
Voting Classifier	**74.46%**

TABLE II. RESULTS (VALIDATION ACCURACY)

Algorithm	***Validation Accuracy***
Naïve Bayes	61.54%
Random Forest	72.04%
Logistic Regression	76.62%
Voting Classifier	**76.71%**



	Naive Bayes	Random Forest	Logistic Regression	Voting Classifier
Validation Accuracy	61.54%	72.04%	76.62%	76.71%
Testing Accuracy	66.99%	74.08%	74.28%	74.46%

Figure 5 Testing and Validation Accuracy

CONCLUSION

In the area of Sentimental Recognition of tweets, the major approach is to compare the accuracy of the voting classifier with the individual ones and check if the assembling improves the accuracy. In this paper, the ensemble classification model has been proposed to improve the classification prediction. From the results and graphs, it is evident that the voting classifier of Naïve Byes, Random Forest and Logistic Regression produces the better predictions than the stand-alone classifiers.

As a future work, the accuracy of the voting classifier can be compared with CNN based classifier. Even the other combination of the classification algorithm can be tested to generate the better predictions.

REFERENCES

[1] Suhasini, M., & Badugu, S. (2018). Two Step Approach for Emotion Detection on Twitter Data. International Journal of Computer Applications, 179(53), 12–19. https://doi.org/10.5120/ijca2018917350

[2] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] A. Yousaf et al., "Emotion Recognition by Textual Tweets Classification Using Voting Classifier (LR-SGD)," in IEEE Access, vol. 9, pp. 6286-6295, 2021, doi: 10.1109/ACCESS.2020.3047831

[4] Ankit, & Saleena, N. (2018a). An Ensemble Classification System for Twitter Sentiment Analysis. Procedia Computer Science, 132, 937–946. https://doi.org/10.1016/j.procs.2018.05.109

[5] Visualization of random forest algorithm. (2019). [Graph]. Https://Medium.Com/Swlh/Random-Forest-and-Its-Implementation-71824ced454f. https://miro.medium.com/max/788/0*f_qQPFpdofWGLQqc.png

[6] Visualization of logistic regression algorithm. (2016). [Graph]. Https://Helloacm.Com/a-Short-Introduction-Logistic-Regression-Algorithm/. https://helloacm.com/wp-content/uploads/2016/03/logistic-regression-example.jpg

[7] Priyanka, V. (2021). Twitter Sentiment Analysis using Deep Learning. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3885137 https://github.com/Vedurumudi-Priyanka/Twitter-Sentiment-Analysis

[8] Majumder, P. (2020, February 23). Gaussian Naive Bayes. OpenGenus IQ: Computing Expertise & Legacy. https://iq.opengenus.org/gaussian-naive-bayes/