Social Media Sentiment Analysis

1st Md. Shamiul Haque Khan Shamya
dept. of Computer Science and Engineering(CSE)
School of Data and Sciences(SDS)
Brac University
Dhaka,Bangladesh
md.shamiul.haque.khan.shamya@g.bracu.ac.bd

2nd Niaz Makhdum Khan
dept. of Computer Science and Engineering(CSE)
School of Data and Sciences(SDS)
Brac University
Dhaka,Bangladesh
niaz.makhdum.khan@g.bracu.ac.bd

3rdMd Shariar Hossain Fahim
dept. of Computer Science and Engineering(CSE)
School of Data and Sciences(SDS)
Brac University
Dhaka,Bangladesh
md.shariar.hossain.fahim@g.bracu.ac.bd

4thMd. Farhadul Islam
dept. of Computer Science and Engineering(CSE)
School of Data and Sciences(SDS)
Brac University
Dhaka,Bangladesh
md.farhadul.islam@g.bracu.ac.bd

5thMd. Sabbir Hossain
dept. of Computer Science and Engineering(CSE)
School of Data and Sciences(SDS)
Brac University
Dhaka,Bangladesh
md.sabbir.hossain1@g.bracu.ac.bd

6th Annajiat Alim Rasel
dept. of Computer Science and Engineering(CSE)
School of Data and Sciences(SDS)
Brac University
Dhaka,Bangladesh
annajiat@gmail.com

***Abstract*—Sentiment is an essential part of human life. Social media is the most popular medium today to get to know about other's lives and also to share our own status. Twitter is one such popular platform. This study aims to collect sentimental status from twitter and analyze them through various machine learning algorithms to evaluate their thoughts. Random Forest Classifier, Logistic Regression, KNeighbor Classifier and CNN classifier have been used in the study as they are well known for sentiment classification.**

***Index Terms*—Sentiment analysis, twitter, Random Forest Classifier, Logistic Regression, KNeighbor, CNN**

I. Introduction

Sentiment analysis is one of the major subfields of natural language processing. It is the computational modeling of opinions, sentiments and subjectivity of text [8]. With the application of machine learning, sentiment analysis has been playing a significant role in identifying human sentiment for a variety of purposes. Our objective in the study is to find out the majority of people's emotions on social platform twitter and compare among them. A bunch of sophisticated algorithms of NLP are available today to work with. Random Forest Classifier is prolific in prediction and Logistic Regression is also a well qualified supervised algorithm that can predict categorical value. K-Nearest Neighbor classifier (KNN) processes data based on similarity and proximity to predict data. CNN or Convolutional Neural Network identifies patterns in data to train and process. These algorithms have been used for a better understanding of sentiment classification in real world scenarios.

II. Research Objective

Everyone is highly dependent on their internal emotional world. On a daily basis, we all deal with a variety of challenges, some of which may have an impact on our psyche. Social media posts are the most common way for people to express their present feelings and thoughts to the rest of the globe. It's reasonable to expect both positive and negative outcomes. There are periods when a lot of posts encourage illegal behavior, and other times when a lot of posts make people happy and uplift them. This information may mislead readers on a social issue. It's unusual to read about someone becoming so depressed that they commit suicide and then to find their suicide note online. Identifying such mentalities and responding appropriately is a goal of ours. Fundamental to our approach is the use of NLP to analyze and comprehend user-provided text.

III. Literature Review

Twitter is a dynamic social networking platform for exchanging information. Since, all the data here is produced by different users, it would be useful in understanding the emotion of a user toward a particular context or situation. Using sentiment analysis on twitter messages, we would be able to extract and evaluate the innate sentiment within a sentence [5].

Random Forest Classifier is known to be one of the best classifier algorithms in the field of natural language processing. It creates multiple decision trees during training time and takes them as individual predictors [1]. Afterward, it uses those trees to perform classification and identifies the class of test instances based on the majority of similar classes [2]. In the case of large amounts of data, random forest has been proved to deliver results with better accuracy.

Logistic regression is another analytical algorithm for prediction of data whose working principle follows the method of statistical analysis. It can be of two types- binary logistic regression and multinomial logistic regression. For sentiment analysis, multinomial regression would be our preferred methodology to apply. It is also known as Softmax Regression [7]. It uses vector property to process the data and result output. To illustrate, it creates a vector of variables and then calculates the weight of all variables to estimate the class [9]. The independent variables in this case have to be linearly related.

The K-Nearest Neighbor or shortly known as KNN is another non parametric machine learning algorithm inside the category of supervised learning approach. It performs the classification by comparing the test data points with the similar trained data points [6]. To recognize similarity, it uses the Euclidean method to calculate distance of one test data with all the existing training data. Then according to the distance, a class is labeled on the test data [4].

The Convolutional Neural Network or CNN is a prominent deep learning algorithm mostly used in the field of image and speech recognition. However, its usage in natural language processing is on the rise as well. Being a multilayer network, the CNN algorithm gives the output of one layer as the input for the next layer [11]. Unlike other algorithms, it's a connectionism model for classification. It's working principle was designed based on environmental stimuli and storing information in a form of connections among neurons [3]. The weights in the networks are assigned in accordance with the training data. To reduce the complexity in computation, it uses the polling layers to minimize the output size of one stack layer to the next, preserving only the important information [10].

IV. Working with Dataset

A. Dataset

A large collection of datasets are available now-a-days for almost any type of research related topic. Kaggle is a very popular digital source of datasets where we have collected our dataset from.

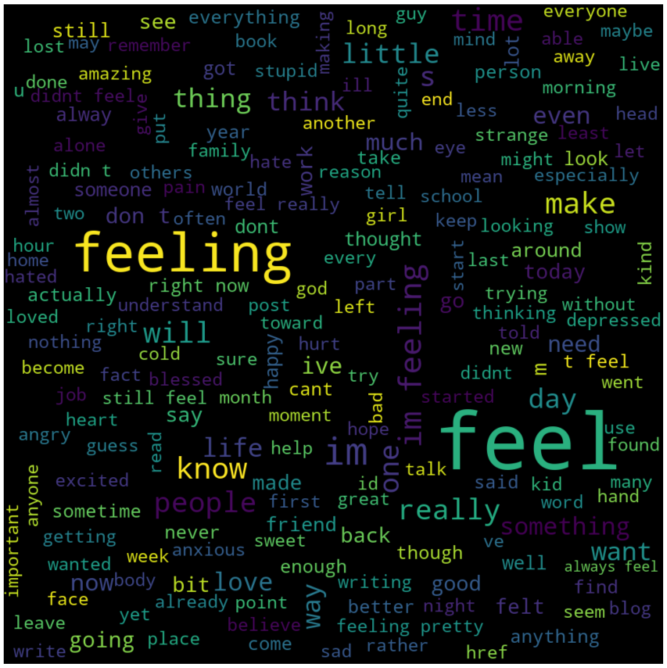

Fig. 1. Most Common Word in Dataset

B. Dataset Description

In our collected dataset there are a total of 21,460 rows, each of them containing a sentence and its corresponding sentiment. The dataset is utilizable for six basic types of emotions: happy, sadness, anger, surprise, fear and love.

C. Data Preprocessing

Any kind of data needs to be preprocessed before putting into an algorithm for analysis. Raw data often contains inappropriate types of value that hamper the training and testing. At first, we cleaned all the unnecessary characters and strings to make the dataset as convenient as possible. For training data, we used 80

V. Proposed Methodology

Our project begins with the collection of sentimental information from social media twitter in the form of a dataset. Then we performed data visualization and data preprocessing to prepare data for algorithms. Next, we applied Random Forest Classifier, Logistic Regression, Multinomial Naive Bayes, KN Classifier, CNN classifier respectively. Following that, we have analyzed output from each algorithm and also measured accuracy.

VI. Experimental Result analysis

The experiment gives us an overview of the rate of six basic emotions people tend to share on twitter.

From the graph, it appears that the sentiment happy and sadness are on top of the list with the amount of over 6000

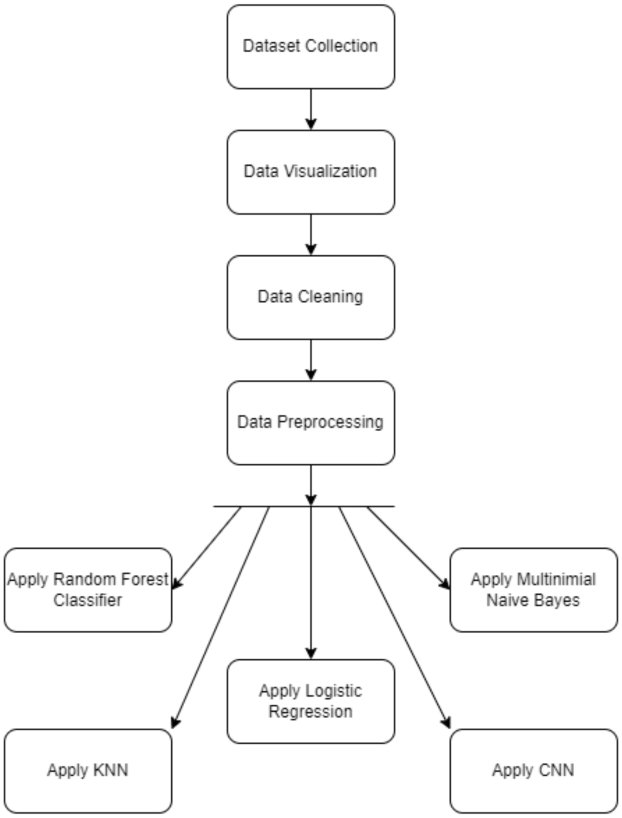

Fig. 2. Proposed Mathedology

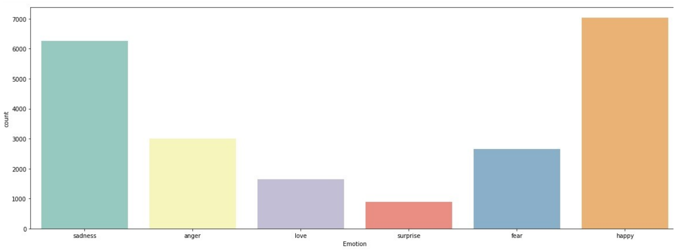

Fig. 3. Caption

times appearance. It means people are very likely to share their happy and sad moments of their life on twitter. Similarly, anger and fear are very close to each other if we look at the figure, showing up for about 3000 times. Next, love and surprise are at the bottom with the number below 2000.

A. Accuracy and Summary of all models:

Model	Accuracy	Precision	Recall	F1 score
Random Forest Classifier	0.86	0.84	0.82	0.83
Multinomial Naive Bayes	0.80	0.88	0.64	0.80
Logistic Regression	0.85	0.86	0.77	0.85
KNN	0.97	0.98	0.98	0.96

After the completion of training and testing, the result shows that KNN comes out to be the most accurate algorithm for our research with an accuracy 97

Then there is the Random Forest Classifier with an accuracy of 86

Lastly, the Multinomial Naive Bayes model performed a little bit worse compared to the other ones as it showed an accuracy of 80

So, keeping in view the results of all the models, we would like to recommend the KNN algorithm for better performance when it comes to analyzing human sentiment related research.

VII. Conclusion

To achieve relatively refined results when searching emotion in text, a comprehensive suggestion system is proposed. The Random Forest Classifier, Logistics Regression, Multinomial Naive Bayes, KNeighbor Classifier and CNN are therefore employed for this entire system. To find a perfect emotion from text we made an accuracy based program. Because the proposed approach is so straightforward, it will also alleviate the burden of devoting an inordinate amount of effort to ensuring that one's preferences are taken into account when making important decisions.

References

[1] Yassine Al Amrani, Mohamed Lazaar, and Kamal Eddine El Kadiri. Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*, 127:511–520, 2018.

[2] Palak Baid, Apoorva Gupta, and Neelam Chaplot. Sentiment analysis of movie reviews using machine learning techniques. *International Journal of Computer Applications*, 179(7):45–49, 2017.

[3] Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. *Expert Systems with Applications*, 72:221–230, 2017.

[4] Novelty Octaviani Faomasi Daeli and Adiwijaya Adiwijaya. Sentiment analysis on movie reviews using information gain and k-nearest neighbor. *Journal of Data Science and Its Applications*, 3(1):1–7, 2020.

[5] Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2):1–41, 2016.

[6] Soudamini Hota and Sudhir Pathak. Knn classifier based approach for multi-class sentiment analysis of twitter data. *Int. J. Eng. Technol*, 7(3):1372–1375, 2018.

[7] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[8] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[9] Anjuman Prabhat and Vikas Khullar. Sentiment classification on big data using naïve bayes and logistic regression. In *2017 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–5. IEEE, 2017.

[10] Anwar Ur Rehman, Ahmad Kamran Malik, Basit Raza, and Waqar Ali. A hybrid cnn-lstm model for improving accuracy of movie reviews sentiment analysis. *Multimedia Tools and Applications*, 78(18):26597–26613, 2019.

[11] Maryem Rhanoui, Mounia Mikram, Siham Yousfi, and Soukaina Barzali. A cnn-bilstm model for document-level sentiment analysis. *Machine Learning and Knowledge Extraction*, 1(3):832–847, 2019.