ICANpred- Twitter sentiment analysis

Devanshi Bavaria
Computer Science and Engineering
Nirma University
Ahmedabad,India
20bce027@nirmauni.ac.in

Dhruv Shah
Computer Science and Engineering
Nirma University
Ahmedabad,India
20bce260@nirmauni.ac.in

***Abstract*—In the present era everyone is addicted to internet and hence opinion on the internet are considered as a valid statement as it is user opinion that he faced which can never be wrong. There are several of social media websites that provides such type of facility such as Facebook, LinkedIn, Google reviews, Twitter and so on. We will be considering the most official and trusted social networking site or application that is Twitter also it is gaining high popularity in terms of any trend It might be political, events and much more furthermore we get abundant data which is in raw format. Reviews will be analyzed and will be represented in a statical format which can be either in the form of bar chart or pie chart. a method has been proposed in which the tweets are classified as Disaster tweets. First, the data is pre-processed and cleaned. After this Machine Learning algorithm called Random Forest is applied for the final classification. The algorithm is applied on the output of both the feature vectors (that is vector from Bag of Words and TF/IDF). The resulting classification depicts that vectorization obtained from TF/IDF gives better classification accuracy.**

***Index Terms*—Twitter, Sentimental analysis, google reviews, dataset with masks and without masks, pie-chart- bar plot, prediction, visualization, positive-negative-neutral content, Bag-of-words, NLP, BERT classification**

I. INTRODUCTION

Basically, every person is on internet nowadays and they all are expressing their opinion with the help of the social media or applications that are easily available on the play-store, app-store and websites. Some of the famous social media websites and applications are LinkedIn, Facebook, Google, Amazon, Flipkart, Snap deal, Alibaba, Myntra and Twitter. Furthermore, these companies provide an opportunity to take and advertise their product online and connect with their buyers or users. Hence based on it, people write their own review on one of the applications which provides an open platform to raise and state their views on a particular thing. Consumer are considered equivalent to god and what they say represents the company/product identity. As the people like to read the reviews that has been stated by the other user and then take a move towards it. So, the reviews can revolve anything and can boom the news to a greater extent. Hence, a company needs to know whether the product they made is performing well or not in the present market. So it will be helpful to them to carry on improvement if needed.

Hence, by going each and every review it becomes tedious task so we will be using Twitter for gathering the reviews which will help the company to conclude whether a product is on the positive side or not. This will be in the form of pictorial representation which will be in bar chart or pie chart. We will be using a pre-defined dataset and it will be filtered means we will be removing the extraneous things which are unnecessary like removal of numerical digits, links and user name to keep it anonymous which will train the model and then we will be testing the reviews that were posted by the users on the twitter. And lastly the conclusion will be drawn on the basis of it. The most popular applications of sentiment analysis in real life:

- Monitoring of Social Media
- Client Assistance
- Customer Opinions
- Management of Reputation and Brand
- Customer and Staff perspectives both are included
- Analysis of Products
- Market Analysis and competitor analysis

II. FRAMEWORKS

We will be using python as a development/base programming tool. Moreover, we are going to use Pandas, Scikit learn (sklearn), Matplotlib and seaborn. Initially, we are going to read the comma separated value file with the help of Pandas library and then we will be using Scikit Learn to divide the data into training and testing and to import the model. Eventually, we will be using Matplotlib and Seaborn to visualize the data. The task that we have done with the frameworks are

- **Classification:** This is a classification application which will be having different categories where categories will be static and will be specified preliminary [1] with the help of our dataset. For an instance “This is bad day” is negative and “what a great day” will be positive.
- **Training:** Some Columns or features from the data will be given to learn as our model will be supervised learning where we are training the model with adequate amount of data. This will be the words that are separated and will be 0 or 1 in Boolean values.
- **Prediction:** The machine will see the features passed to the model and will decide the output as per the trained model

Identify applicable funding agency here. If none, delete this.

978-1-7281-4387-3/19/$31.00 © 2019 IEEE

III. RELATED WORK

In [1] the classification of tweets is carried out as good, bad or neutral relying upon the sentiments of the tweets for different electronic products and a prod- uct review is generated for the same. The classification is performed using the machine learning module called Support Vector Machines. In [2] the authors have done sentimental analysis on the reviews generated by the customers. They have classified these reviews as satisfied, not satisfied or in-between. Based on the count of each type of review they have provided the final conclusion. This method helps other customers who are willing to buy some product. They have used tweets to analyze their method as tweets are producing a lot of opinions nowadays. Micro-blogging sites are a source of lots of information. They provide information even before the traditional media.

Rumors are also present among this information and they can misguide the people. Hence it is necessary to detect these rumors and remove them. Also the classification of these rumors is necessary. So, In[3] the authors have first classified whether it is a rumor or not. The sentiment is analyzed, emotions are analyzed and named entity recognition is performed. In [4] accuracy is analyzed using sentiment classification of the financial tweets along with making use of neural networks and logistic regression. As a mat- ter of first importance, the entire row is deleted if it doesn't contain any emoji and Stop-words are removed. Bigram TF and Unigram TF-IDF are used for text vectorization and for dimensionality reduction Chi squared test is used for capturization into rank features. Problem with this process is conflict between positive and neutral classes. In [5] the Objective is to add semantics in feature vectors to upgrade sentiment classification by utilizing ensemble techniques. Feature Vector Formation is performed using Decision tree, Random Forest and AdaBoost Decision Tree followed by data preprocessing and Synset finding which is of semantic similarities between the tweets using synsets of the WordNet. In [6] Tweets are classified into positive, negative and neutral sentiment using some basic steps of data preprocessing like removing emoji, identification of lower and upper case, compression of words etc. If the dataset is imbalanced then SMOTE technique is used for removing the skewness of the dataset. Later SVM, Random forest and Naive Bayes methods are utilized.

In [7] A sentiment classifier is built that can determine positive, negative and neutral sentiments by extracting the real time tweets and removing the retweets as they will produce bias in the classification process. Naive Bayes algorithm is balanced to fit in the Map Reduce model. Later visualization techniques are used to feature tweets and their related sentiments.

In [8] The author is investigating the kinds of triggers that spark inclines on Twitter without need of outer information news to find breaking news progressively using the printed substance of tweets with the help of tokenization and bag of words approach. SVM arrangement is utilizing the Term Frequency estimation of each term.

Through [9] the author plans to help the users in identifying and sifting through assortments of conceivably tricky news with the investigation of the news which were already observed beforehand. Predictive model is being designed to weigh the pros and cons based on the news available in corpora and then distributed into three different categories: serious fabrications, large scale hoaxes, humorous fakes. Libraries and information science (LIS) is also essential in filtering, vetting and verifying online information.

IV. CONTRIBUTION

To solve the analysis problem of using twitter data properly as an asset, we perform Twitter Sentiment Analysis. We use the data for many of the analysis problems like, stock market, popularity television shows, political issues etc. The Twitter data is firstly fetched using the tweepy library of python. Further we preprocess the data by removing emojis, tokenizing the data and removing the stop words (words like 'a', 'and' 'the' which are of no importance for the classification of the tweets). Then we perform feature extraction on the preprocessed data to produce feature vector for passing it to the training model. Then divide the data into training and testing sets. And finally applying the Classification Algorithm Support Vector Machine or any other classifiers for Classifying the positive and negative tweets.

V. DATASETS

We have used training.1600000.processed.noemoticon.csv to train our model which contains "target", "ids", "date", "flag", "user", "tweets".

- **Target:** - Tweet positive or negative
- **Ids:** - numerical ids of user
- **Date:** - on which it was posted
- **Flag:** - Query or not
- **User:** - username on twitter
- **Tweets:** - what they tweeted

The shape of the data will be 1600000 X 6 which indicates there will be 1600000 rows and 6 columns which are stated above. The data will be like:

- **Target:** - 0
- **Ids:** - 1467810369
- **Date:** - Mon Apr 06 22:19:45 PDT 2009
- **Flag:** - NO_QUERY
- **User:** - $\_TheSpecialOne\_$
- **Tweets:** - @switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. This is a row which is represented column wise. But we will be selecting the target variable and the tweets. Target variable will be in 0 and 4 where 0 denotes negative and 4 will be positive.

VI. DATA PRE-PROCESSING

- Removal of Numeric values
- Removal of links
- Replacing RT from tweets
- Lowering the tweets words

VII. ARCHITECTURE

The basic flow that we will follow is by inserting keyword into the python terminal by giving the credentials that are essentials and with the proper authorization later it will ask the number of posts that we want to analyse which will be added into data frame. The training data will be prestored data and the testing data will be the data accumulated from twitter. Hence, at last user will be given an option to choose a visualization option which will be either bar or pie.

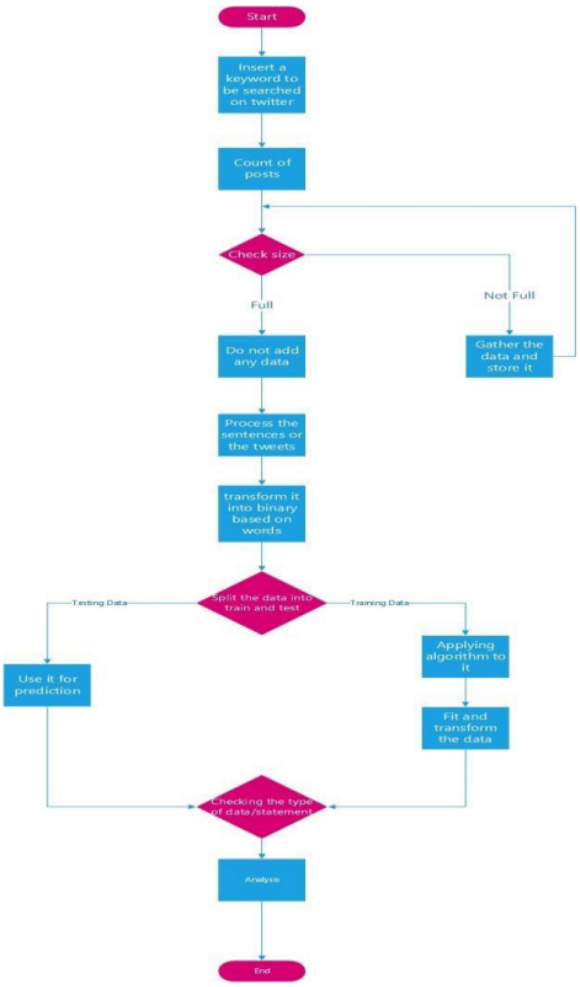

Fig. 1. TF-IDF

VIII. METHODS

To solve the problem of classification of the disaster tweets from the data collected from twitter is solved in this article by the method in which we first preprocess the data using natural language processing, followed by feature extraction, which is a technique of representation of large set or collection of data into selected numerical feature vectors. Feature extraction strategy includes tokenization, counting ad normalization of data. For performing these steps, we have different techniques, two of them on which this article is focused are as following:

- **Bag of Words (BOW):** It serves as a model for the easing of information retrieval from the natural language processing representation system and its simplification. The approach is to take text data and represent as it as a bag of words without regard of grammar and ordering of the words by just maintaining its count and following the general feature extraction flow. The method used in this article uses this technique as well as another technique called which is more reliable as it is dependent on the frequency of the words also rather than just depending on the count of words.
- **Term Frequency-Inverse Document Frequency(TF-IDF):** A numerical statistical model is used to gather the words and keep track of them. while in addition to it, it also intends to reflect the importance of each word in the sentence. It is a metric that indicates how important a word is in a document relative to the entire corpus of documents.

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i})$$

Fig. 2. TF-IDF

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

From the above two feature extraction techniques following the natural language processing, the paper finally states which of the two proposed methods gives a better and improved performance for the classification purpose.

IX. EXPLANATION OF THE FLOW

- **Gathering social networking data:** In today's world, everyone is addicted to social media, because any comment made on the internet is considered legitimate because it is a user's opinion, which can never be incorrect. Several social media platforms, such as Facebook, LinkedIn, Google reviews, Twitter, and others, offer this form of service. look at Twitter, the most official and trusted social networking site or application.
- **Parsing the data:** Parser is a tool that parses text. All inaccessible tweets are removed from the downloaded data by the parser.
- **Preprocessing:** Emoji should be replaced by their oscillation. Targets and URLs should be removed. Acronyms should be expanded. 'Brb' stands for 'be right back,' for example. Stop words should be removed. Tokenization is a term used to describe the process of Case-folding stemming. Delete the punctuation marks. Replace 'hellooooo' with 'hello' in a series of repeating characters.
- **Feature Extraction:** The feature extractor receives the pre-processed data file and generates the feature vector.

The unigram model was the basic (baseline) function. The basic vector for each tweet was created by compiling a list of all specific unigrams across the training collection.

- **Training the model:** It is sufficient to acquire the appropriate values for all weights and bias from labelled cases while training a model. By reviewing several samples and searching for a model that minimises loss, an empirical risk minimization algorithm creates a model for supervised learning.
- **Applying classifier:** Classification is a type of supervised learning when the goals are also given access to the input data. The classifier receives the extracted features. The created model is employed to foretell the tone of the fresh tweets.

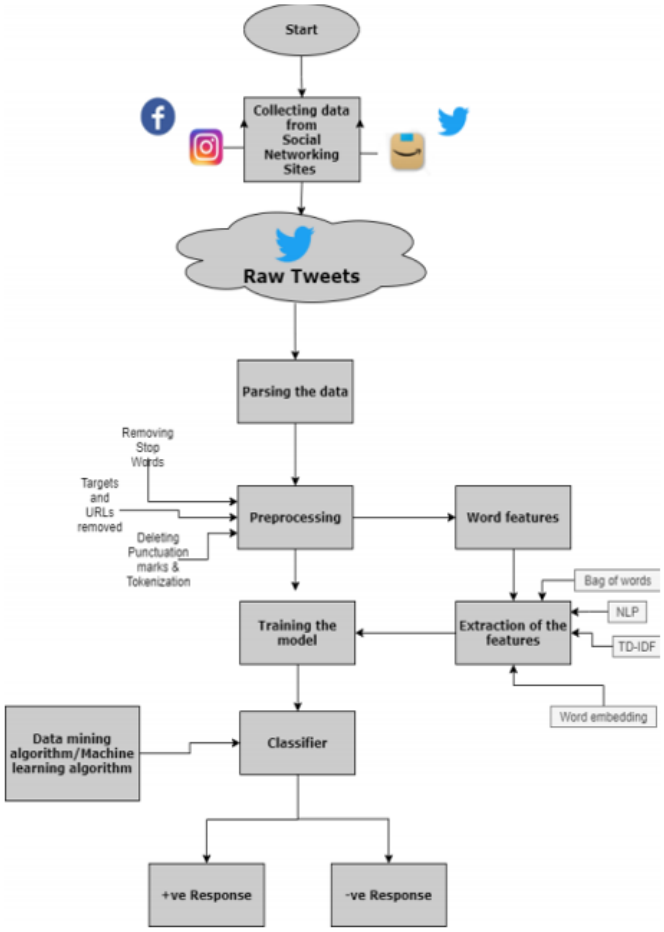

Fig. 3. Flow

X. BERT CLASSIFICATION

One of the most popular current language modelling architectures is BERT. Depending on the requirements of the user, it can be applied to a range of downstream tasks, such as sentiment analysis, NER, link extraction, or question-answering.Almost all the parameters of internal layers of the architecture are fixed because the architecture's core was trained on exceptionally large text corpora. On the other hand, the outermost layers adapt to the goal and are where fine-tuning happens. While learning more about BERT, the base and the big are two crucial structures to consider. The following list identifies the four key areas where the architectures differ: The following variables differ from one another: 2021, 21, 133, 10 of the 21 transformer encoder hidden layers, The number of attention heads, also known as self-attention (12 vs. 16), the hidden size of the feed-forward networks (768 vs. 1024), and finally the maximum sequence length parameter (512 vs. 1024), also known as the maximum allowed input vector size, are all different from one another.

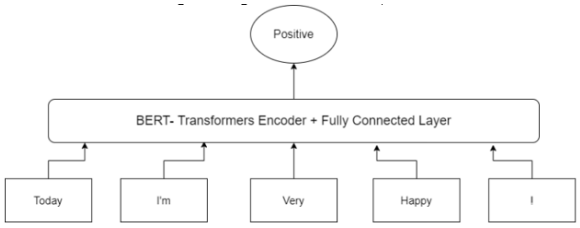

Fig. 4. BERT-Architecture overview

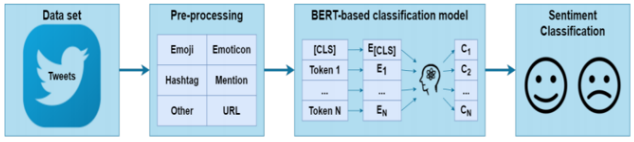

Fig. 5. BERT- classification process

Two more distinct tokens are used in the BERT architecture: SEP is used for segment splitting and CLS is used for classification. Both of these tokens describe the complete sequence and operate as the first input for any classifier. This results into output vector with the same size as the hidden node size H. The final secret state of the first token used as input, which is the transformers' output, can therefore be represented as a vector $C \in R_H$.In other words, a vector $C \in H$ can be used to represent the final concealed state of the initial input token. The vector C is provided as an input to the final fully linked classification layer. Given the parameter matrix $W \in R_{KxH}$ of the classification layer, where K is the number of categories, the softmax function will calculate the likelihood of each category P as follows: P = softmax(CW_T). The outermost classification layer, which made up the entire classification model, was adjusted before the BERT language model was pre-trained. It is a two-step pipeline that uses a pre-trained version of BERT in the second stage after using a variety of pre-processing techniques in the first stage to convert emoticons and emojis from Twitter into plain text. It was created as a two-stage pipeline, where the first stage used a number of pre-processing procedures to convert emoticons and emojis from Twitter into plain text, while the second stage used a pre-trained version of BERT.

XI. RESULTS

We have trained the data on various models to see which algorithms suits the best so the conclusion that we have is shown in the bar graph.

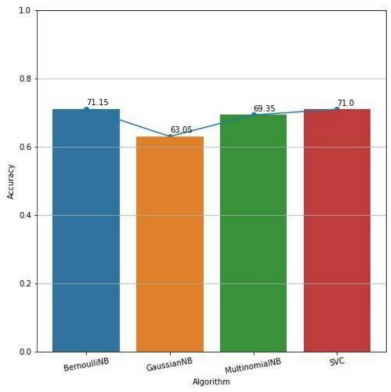

Fig. 6. Comparison between various classifiers

So, from the above-mentioned data we found that Bernoulli Nb was at the peak model among all having 71.15 as accuracy score.

By Using Bernoulli Nb we have tried by different size of data and hence the accuracy that we obtained is shown in the below table.

Data Size	Accuracy
1000	0.615
2000	0.66
5000	0.711
10000	0.712
20000	0.706
40000	0.725375
60000	0.7364167
80000	0.7323125
100000	0.729

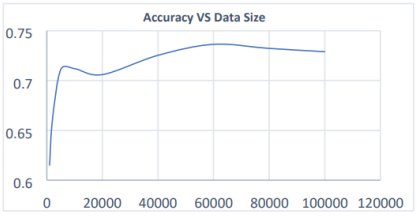

Fig. 7. Plot

The charts show results of predicting live tweets on twitter which are represented in pie chart as well as in bar chart by giving a specific keyword.

XII. CONCLUSION

The goal of this project was to develop an approach that could be used in the real world to analyse sentiment on Twitter using the bag of words, countvectorizer, bernouliiNB, and

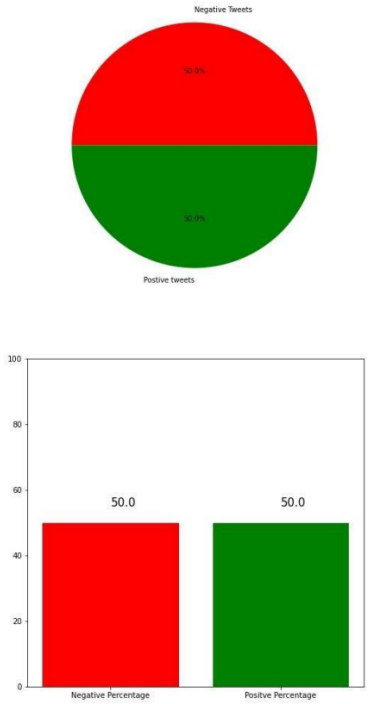

Fig. 8. Pie-Chart and Bar-Graph

BERT language models. The system was created as a two-step pipeline where the first step involved a set of pre-processing steps to convert Twitter language, such as emojis and emoticons, into plain text, and the second step used a pre-trained version of BERT, with a focus on the pre-processing stage, to confirm its effectiveness. Lastly, the suggested strategy's applicability and generalizability will be verified and evaluated against other datasets, languages, and social media platforms, such as Facebook messages.

REFERENCES

[1] A. P. Patel, A. V. Patel, S. G. Butani, P. B. Sawant, Literature survey on sentiment analysis of twitter data using machine learning approaches, In-ternational Journal for Innovative Research in Science Technology 3 .

[2] G. Gautam, D. Yadav, Sentiment analysis of twitter data using machine learning approaches and semantic analysis, in: 2014 Seventh International Conference on Contemporary Computing (IC3), IEEE, 2014, pp. 437–442.

[3] M. Kanakaraj, R. M. R. Guddeti, Nlp based sentiment analysis on twitter data using ensemble classifiers, in: 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), IEEE, 2015, pp. 1–5.

[4] A. Zubiaga, D. Spina, R. Mart´ınez, V. Fresno, Real-time classification of twitter trends, Journal of the Association for Information Science and Tech- nology 66 (3) (2015) 462–473.

[5] K. Stowe, M. Paul, M. Palmer, L. Palen, K. M. Anderson, Identifying and categorizing disaster-related tweets, in: Proceedings of The fourth interna- tional workshop on natural language processing for social media, 2016, pp. 1–6.

[6] S. A. Phand and J. A. Phand, "Twitter sentiment classification using stanford NLP," 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), Aurangabad, India, 2017, pp. 1-5, doi: 10.1109/ICISIM.2017.8122138.

[7] M. R. Hasan, M. Maliha and M. Arifuzzaman, "Sentiment Analysis with NLP on Twitter Data," 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2019, pp. 1-4, doi: 10.1109/IC4ME247184.2019.9036670.

[8] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019, pp. 1-5, doi: 10.1109/ICIC47613.2019.8985884.

[9] L. Mandloi and R. Patel, "Twitter Sentiments Analysis Using Machine Learninig Methods," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-5, doi: 10.1109/INCET49848.2020.9154183.

[10] C. W. Park and D. R. Seo, "Sentiment analysis of Twitter corpus related to artificial intelligence assistants," 2018 5th International Conference on Industrial Engineering and Applications (ICIEA), 2018, pp. 495-498, doi: 10.1109/IEA.2018.8387151.

[11] V. Ikoro, M. Sharmina, K. Malik and R. Batista-Navarro, "Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers," 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2018, pp. 95-98, doi: 10.1109/SNAMS.2018.8554619.

[12] R. Wagh and P. Punde, "Survey on Sentiment Analysis using Twitter Dataset," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 208-211, doi: 10.1109/ICECA.2018.8474783.

[13] R. I. Permatasari, M. A. Fauzi, P. P. Adikara and E. D. L. Sari, "Twitter Sentiment Analysis of Movie Reviews using Ensemble Features Based Naïve Bayes," 2018 International Conference on Sustainable Information Engineering and Technology (SIET), 2018, pp. 92-95, doi: 10.1109/SIET.2018.8693195.

[14] A. Shelar and C. -Y. Huang, "Sentiment Analysis of Twitter Data," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 2018, pp. 1301-1302, doi: 10.1109/CSCI46756.2018.00252.

[15] H. AlSalman, "An Improved Approach for Sentiment Analysis of Arabic Tweets in Twitter Social Media," 2020 3rd International Conference on Computer Applications Information Security (ICCAIS), 2020, pp. 1-4, doi: 10.1109/ICCAIS48893.2020.9096850.

[16] A. P. Jain and V. D. Katkar, "Sentiments analysis of Twitter data using data mining," 2015 International Conference on Information Processing (ICIP), 2015, pp. 807-810, doi: 10.1109/INFOP.2015.7489492.

[17] M. Abdullah and M. Hadzikadic, "Sentiment Analysis of Twitter Data: Emotions Revealed Regarding Donald Trump during the 2015-16 Primary Debates," 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), 2017, pp. 760-764, doi: 10.1109/ICTAI.2017.00120.