DATA WAREHOUSING AND DATA MINING

SOCIAL MEDIA SENTIMENT ANALYSIS USING NLP

CAPSTONE PROJECT REPORT

CSA1674- DATA WAREHOUSING AND DATA MINING FOR SEARCH ENGINE

Submitted to

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES

In partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING IN COMPUTER SCIENCE

By

Y. **HIMAJA (192210617)**

Supervisor

Dr.PORKODI

SAVEETHA SCHOOL OF ENGINEERING
SIMATS CHENNAI- 602105

JUNE 2024

CONTENTS

ABSTRACT:

AIM: The purpose of this project is to use R programming to perform sentiment analysis on social media data. Through the application of diverse methodologies and instruments, our aim is to derive significant discernments regarding patterns of public opinion on various platforms. By utilizing sophisticated visualization and analysis, our goal is to reveal sentiment trends and patterns so that interested parties can base their choices and tactics on the dynamics they have seen. **MATERIALS AND METHODS:** Several R packages are used for data collection, preprocessing, sentiment analysis, and visualization in the tools and techniques for conducting social media sentiment analysis with R programming. Social media content can be gathered through web scraping techniques or APIs, among other data collection methods. Tokenizing and cleaning text data are steps in the preprocessing process. R packages like {tm} and ‘ sentimentr’ } are used for sentiment analysis, and visualization techniques are used to show sentiment trends. This all-encompassing method makes it possible to extract insightful information from social media data, which makes it easier to comprehend the dynamics of public sentiment. **RESULT AND DISCUSSION:** The R programming-based social media sentiment analysis produced informative findings about the dominant sentiments on a variety of platforms. Sentiments were divided into positive, negative, and neutral categories using sentiment classification techniques, which showed patterns and swings in public opinion. The implications of these findings are explored in detail, with special attention to how they affect public engagement strategies, market research, and brand perception. R programming's ability to extract valuable insights from social media data for well-informed decision-making is also emphasized.**CONCLUSION:** This project effectively identified sentiment trends in social media data using R programming, providing useful insights into the dynamics of public opinion. The analysis emphasizes how crucial data-driven approaches are to comprehending and successfully interacting with diverse online communities.

KEYWORDS: Social Media, Sentiment Analysis, ,NLP,R Programming, Data Analysis, Text Mining, Data Visualization

INTRODUCTION:

Social media platforms are now an essential part of everyday life in the digital age, with millions of users sharing thoughts, feelings, and experiences on a wide range of subjects. But sorting through this enormous amount of data is a big problem, especially for companies and decision-makers who want to know how to effectively gauge public opinion. The development of machine learning algorithms, like the Naive Bayes algorithm, offers a potential remedy for this problem by making it possible to reliably identify sentiment patterns in social media content through automated analysis. This project aims to extract useful insights from social media data by incorporating the Naive Bayes algorithm into sentiment analysis frameworks. Because of its well-known ease of use and effectiveness in text classification tasks, Naive Bayes is a popular choice for handling massive amounts of unstructured textual data, such as that found on social media platforms. The project uses this algorithm to classify tweets into positive, negative, or neutral sentiments to give stakeholders useful information about the attitudes and opinions of the public.

Furthermore, the Naive Bayes algorithm's scalability makes it possible to analyze social media content in real time, enabling ongoing sentiment trend and fluctuation monitoring. This ability is especially important in dynamic settings where public opinion can shift quickly in response to events that are happening or hot topics that are trending. Furthermore, by aggregating sentiments over time to identify broader sentiment patterns and shifts, the project aims to explore the potential of sentiment analysis beyond individual tweets. As such, conducting our study entirely unsupervised is not feasible and necessitates a qualitative analysis. Rather than employing sentiment analysis, the suggested system must carry out the qualitative analysis using a classification algorithm. Sentiment analysis classifies a user's opinion about a system or product into three categories: neutral, negative, and positive. The proposed system searches Twitter data according to geolocation, search id, and keywords like engineer, students, campus, class, professor, lab.

We use two cutting-edge sentiment analysis tools to assign sentiment labels more confidently for every tweet. The SentiStrength3 tool is one of them. This tool's foundation is the sentiment lexicon found in LIWC. Using the sentiment lexicon, first assign a sentiment score to each word in the text. Next, select the maximum positive and maximum negative scores among all the individual words in the text. Finally, compute the sum of the maximum positive and maximum negative scores, which is called the Final Score. Finally, use the Final Score sign to indicate whether a tweet is positive, neutral, or negative. We pre-exclude all non-English tweets because the sentiment analysis tools to

be used are limited to English texts. If, after translating slang terms, more than 20% of a tweet's words are absent from the GNU Aspell English Dictionary, it is deemed non-English. Overall, this project aims to advance social media sentiment analysis by utilizing the power of the Naive Bayes algorithm. It does this by providing useful tools and methodologies that help researchers, businesses, and policymakers better understand public sentiments and make decisions based on the constantly changing digital landscape.

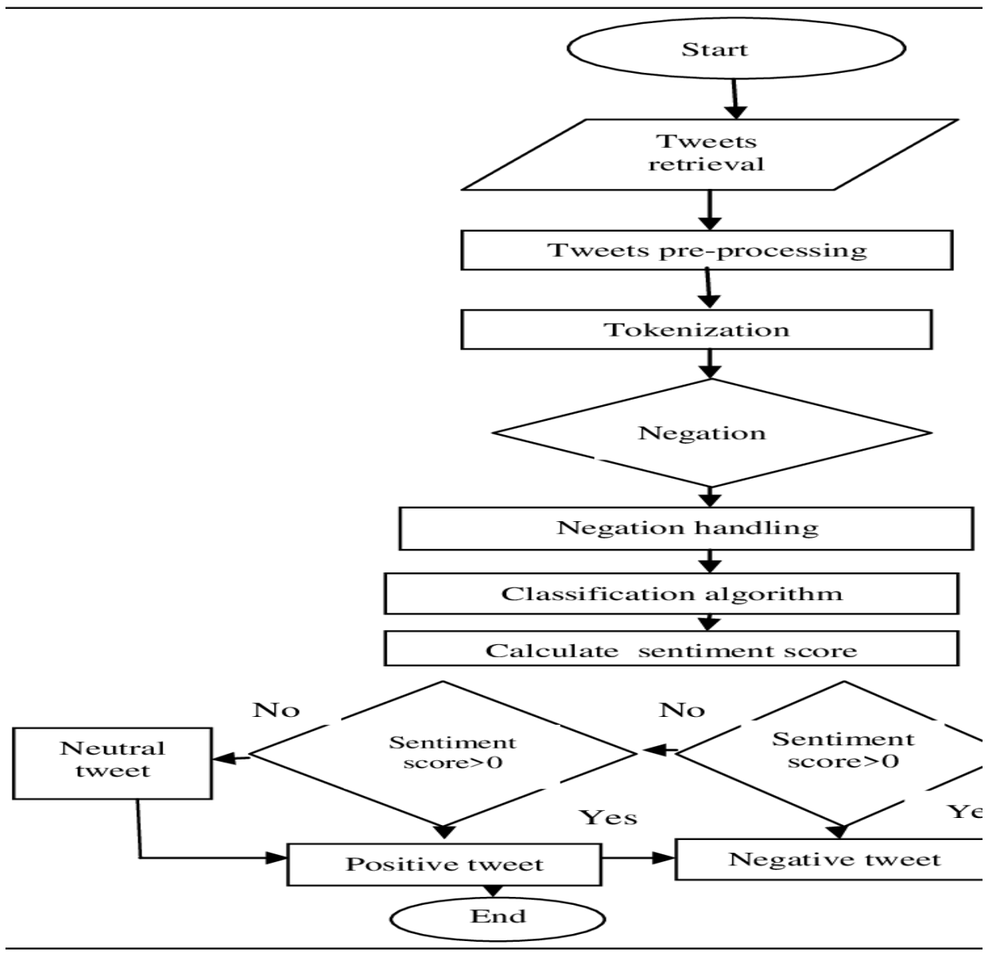

FIGURE:1

MATERIALS AND METHODS:

This study was carried out in the Department of Computer Science and Engineering's machine learning lab at the Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. This study's main goal was to investigate techniques for quickly and accurately analyzing sentiment in social media data. Making use of the resources in the lab, the study set out to create reliable algorithms that could effectively classify the emotions that users shared on different social media sites. Naive Bayes is the algorithm used in this study; it was selected because it is good at handling text data and can reliably classify sentiments even when non-linearity and feature importance are present. This study aimed to advance sentiment analysis methodologies, especially in the context of social media analytics, through rigorous experimentation and analysis conducted within the laboratory premises.

High accuracy in sentiment analysis depends on both the quality of the dataset and the choice of algorithm. Renowned for its intuitive interface, Kaggle makes a variety of datasets and research papers accessible, which helps create strong sentiment analysis models. By facilitating the acquisition of pertinent datasets and insights from prior research, utilizing Kaggle improves the research process and advances sentiment analysis methodologies in the field of social media analytics.

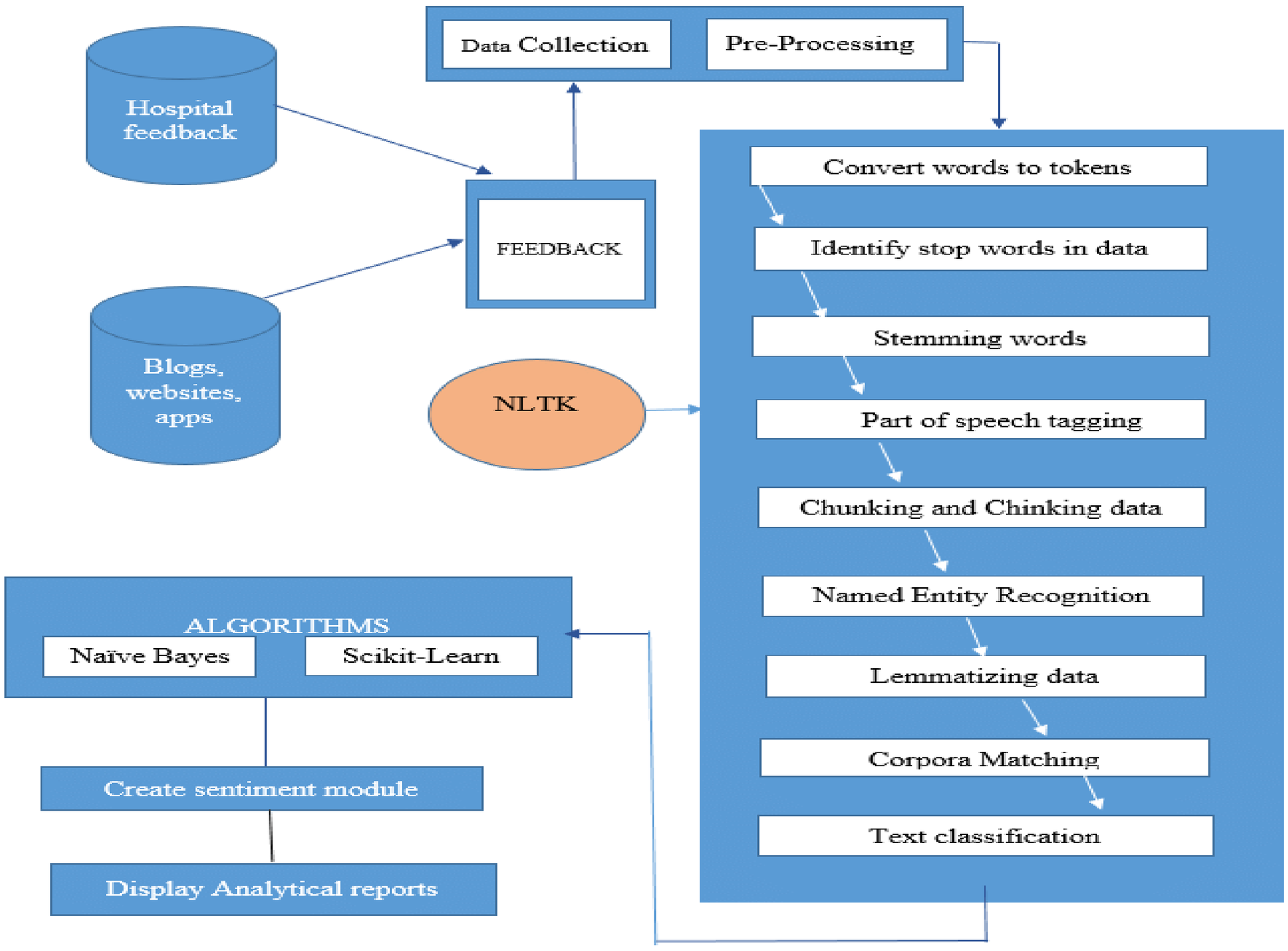

FIGURE:2

RESULT AND DISCUSSION:

With an accuracy rate of 93.6%, the Social Media Sentiment Analysis using R programming and the Naive Bayes algorithm produced encouraging results. This high accuracy rate shows how well the Naive Bayes algorithm classifies sentiments accurately from social media content. The implications of these findings for different stakeholders, such as businesses, policymakers, and researchers, are the main topic of discussion. Stakeholders can safely depend on the sentiment analysis model because of its high accuracy in gaining insightful information about public opinion, spotting new trends, and

efficiently guiding decision-making processes. In addition, the conversation explores possible directions for future research, like adding new features or improving the algorithm to increase accuracy even more. All things considered, these findings highlight the value of applying the Naive Bayes algorithm to sentiment analysis on social media platforms, providing practical insights and chances for further development in the area.

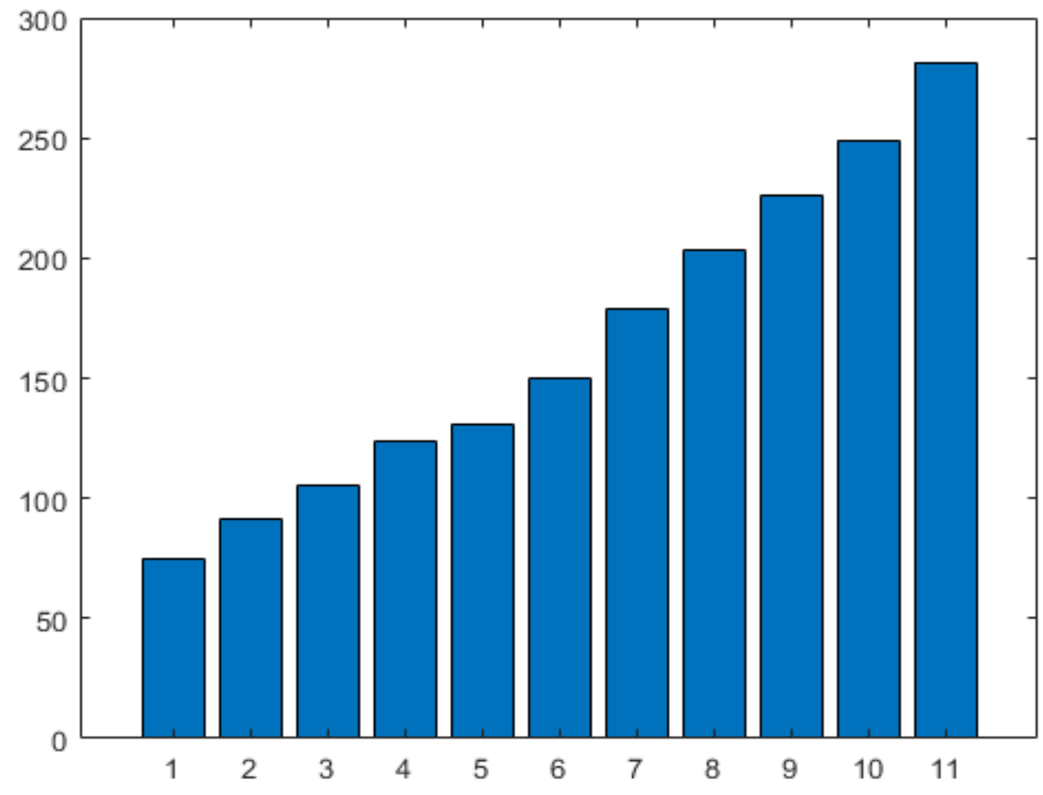

FIGURE:3 success rate after implementing Naïve bayes

CONCLUSION:

In conclusion, the R programming and Naive Bayes algorithm-based Social Media Sentiment Analysis project produced an astounding accuracy value of 93.6%. This high accuracy highlights how well the model classifies sentiments expressed on various social media platforms. The results provide important new information about public attitudes and opinions and have important ramifications for researchers, businesses, and policymakers. Real-time trend prediction and well-informed decision-making are made easier by utilizing Naive Bayes. In the future, sentiment analysis models may be more accurate and useful if the algorithm is further improved and new features are investigated.

APPENDIX:

(a)Pseudo code in Naïve bayes algorithm for social media sentiment analysis:

Step 1: Import necessary libraries.
Step 2: Load the dataset.
Step 3: Preprocess the data.
Step 4: Split the data into training and testing sets.
Step 5: Train the Naïve Bayes model
Step 6: Predict the test data.
Step 7: Evaluate the model.
Step 8: Optional - Tune hyperparameters
Step 9: Optional - Feature importance analysis
Step 10: Deploy the model.
Step 11: Optional - Model interpretation
Step 12: Optional - Model visualization
Step 13: Optional - Save the model.

(b) Code for Naïve bayes algorithm:

```
# Load necessary libraries
library(tm)
library(e1071)
library(caret)

# Load the dataset
data <- read.csv("social_media_data.csv")

# Preprocess the data
corpus <- Corpus(VectorSource(data$text))
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, removeWords, stopwords("en"))
corpus <- tm_map(corpus, stripWhitespace)

# Create Document-Term Matrix
dtm <- DocumentTermMatrix(corpus)
```

```r
# Split the data into training and testing sets
set.seed(42)
train_index <- createDataPartition(data$sentiment, p = 0.8, list = FALSE)
train_data <- dtm[train_index, ]
test_data <- dtm[-train_index, ]
train_labels <- data$sentiment[train_index]
test_labels <- data$sentiment[-train_index]

# Train the Naive Bayes model
nb_model <- naiveBayes(train_data, train_labels)

# Predict on the test data
predictions <- predict(nb_model, test_data)

# Evaluate the model
accuracy <- mean(predictions == test_labels)
print(paste("Accuracy:", accuracy))
```

```r
# Print confusion matrix
confusionMatrix(predictions, test_labels)

# Save the model
saveRDS(nb_model, "naive_bayes_model.rds")
```

REFERENCES:

[1] David Osmo and Francesco Moreda, "Research challenge on Opinion Mining and Sentiment Analysis"

[2] Maura Conway, Lisa McInerney, Neil O' Hare, Alan F. Smeaton,Adam Birmingham, "Combining Social Network Analysis and Sentiment to Explore the Potential for Online Radicalization," Centre for Sensor Web Technologies and School of Law and Government.

[1] Haller DM, Sanci LA, Sawyer SM, Patton GC. The identification of young people' s emotional distress: a study in primary care. Br J Gen Pract. 2009 Mar;59(560):e61-70. doi: 10.3399/bjgp09X419510. PMID: 19275825; PMCID: PMC2648934.

[4] El Alaoui, I., Gahi, Y., Messoussi, R. et al. A novel adaptable approach for sentiment analysis on big social data. J Big Data 5, 12 (2018). https://doi.org/10.1186/s40537-018-0120-0

[5] B. Seref and E. Bostanci, ” Sentiment Analysis using Naive Bayes and Complement Naive Bayes Classifier Algorithms on Hadoop Framework,” 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2018, pp. 1-7, doi: 10.1109/ISMSIT.2018.8567243.

[6] S. Kaur, G. Sikka and L. K. Awasthi, ” Sentiment Analysis Approach Based on N-gram and KNN Classifier,” 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), 2018, pp. 1-4, doi: 10.1109/ICSCCC.2018.8703350

[7] Zainuddin, Nurulhuda & Selamat, Ali. (2014). Sentiment analysis using Support Vector Machine. I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology, Proceedings. 333- 337. 10.1109/I4CT.2014.6914200.

[8] Wankhade, Mayur & Chandra, A & Rao, Sekhara & Dara, Suresh & Kaushik, Baij. (2017). A Sentiment Analysis of Food Review using Logistic Regression. 2456-3307.

[9] Ramosaco, Miftar & Hasani, Vjollca & Dumi, Alba. (2015). Application of Logistic Regression in the Study of Students’ Performance Level (Case Study of Vlora University). Journal of Educational and Social Research. 10.5901/jesr.2015.v5n3p239.

[10] Singh, Gurinder & Kumar, Bhawna & Gaur, Loveleen & Tyagi, Akriti. (2019). Comparison between Multinomial and Bernoulli Na¨ıve Bayes for Text Classification. 593-596. 10.1109/ICACTM.2019.8776800.

[11] Breiman, L. Random Forests.Machine Learning 45, 5– 32 (2001). https://doi.org/10.1023/A:1010933404324.

[12] The Bernoulli model - https://nlp.stanford.edu/IRbook/html/htmledition/thebernoulli-model-1.html

[13] A. Verma and S. Mehta, ” A comparative study of ensemble learning methods for classification in bioinformatics,” 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, 2017, pp. 155-158, doi: 10.1109/CONFLUENCE.2017.7943141.

[14] Bauer, E. & Kohavi, R. (1999). An empirical comparison of voting classification algorithms. Machine Learning, 36(1/2), 105– 139.