COMMENT ANALYSIS USING SENTIMENT ANALYSIS

A PROJECT REPORT

Submitted by

HIRTHIK MATHESH GV	**210701084**
KOUSHIK H	**210701125**
HAYAGRIV KOUSHIK S	**210701081**

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

ANNA UNIVERSITY:: CHENNAI 600 025

MAY 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Report titled "**COMMENT ANALYSIS USING SENTIMENT ANALYSIS**" is the bonafide work of **",“HAYAGRIV KOUSHIK S(210701081)",“ HIRTHIK MATHESH GV (210701084)", KOUSHIK H(210701125)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr.K.Anand, M.E., Ph.D
PROJECT COORDINATOR
Professor,
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai – 602105

Submitted to Project Viva-Voce Examination held on ___________________

Internal Examiner **External Examiner**

ABSTRACT

Sentiment Analysis involves analyzing the sentiments conveyed in tweets to gain a deeper understanding of the emotions expressed. This process is commonly utilized in data analysis for computational purposes. To begin, it is necessary to source the essential datasets from platforms like Kaggle to facilitate the project. At the initial process basic python packages like pandas, mat plot, sci-kit learn. Once this cleaning process is complete, the refined datasets can be inputted into the model for evaluation. Initially, the comments within the datasets may exhibit biased tendencies. However, through the integration of the ER library, these datasets undergo a transformation using oversampling or down sampling/under sampling, becoming more impartial and allowing for a clearer distinction between positive and negative comments. Following this transformation, the datasets undergo additional testing with a focus on applying the concept of majority and minority portions to refine the analysis further. Proceeding with this refined data, the datasets are once again fed into the model. Leveraging tools like Sklearn feature extraction for count vectorization and the sklearn_ linear model with SGDC classifier, the datasets are classified based on the polarity of the comments, leading to a segmentation of positive and negative sentiments. To present the findings, a bar plot is utilized to visualize and interpret the results effectively. As the processed training and test datasets are fed through the prediction function, the outcomes are displayed in a binary format of zeros and ones using F1 score method. Here, zeros signify positive comments while ones represent negative feedback.

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavor to put forth this report. Our sincere thanks to our Chairman **Mr. S.MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **Dr.K.Anand ME,Ph.d** Professor, Department of Computer Science and Engineering. Rajalakshmi Engineering College for his valuable guidance throughout the course of the project.

HAYAGRIV KOUSHIK S (210701081)

HIRTHIK MATHESH GV (210701084)

KOUSHIK H (210701125)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SVM	Support Vector Machine
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
NLP	Natural Language Processing
SGD Classifier	Stochastic Gradient Descent Classifier

CHAPTER 1

INTRODUCTION

Sentiment analysis on Twitter presents unique challenges and opportunities. The brevity and informal nature of tweets, coupled with the rapid evolution of language on social media, require robust and adaptable models. Moreover, the real-time aspect of Twitter data means that sentiment analysis can provide immediate insights into public opinion, making it a powerful tool for businesses and policymakers.

Applications of Twitter sentiment analysis are vast. Businesses use it for brand monitoring and customer feedback analysis, while political analysts gauge public reaction to events and speeches. During crises, sentiment analysis can help in understanding public sentiment and guiding communication strategies.

In conclusion, sentiment analysis on Twitter is a multifaceted process that leverages the platform's rich and dynamic data to extract valuable insights. As techniques in NLP and machine learning continue to advance, the accuracy and applicability of sentiment analysis on Twitter are likely to expand, offering even more profound opportunities for understanding and leveraging public sentiment.

Sentiment analysis, also known as opinion mining, is a field within natural language processing (NLP) that focuses on identifying and extracting subjective information from text. The objective is to determine the sentiment expressed in a piece of text, typically categorizing it as positive, negative, or neutral. One of the most dynamic and rich sources of text data for sentiment analysis is Twitter, a microblogging platform where users post short messages, or tweets, about a myriad of topics in real time.

Twitter sentiment analysis has garnered significant interest from researchers and businesses alike due to the platform's wide reach and the real-time nature of the data it provides. With over 330 million monthly active users generating vast amounts of data daily, Twitter offers a valuable corpus for analyzing public opinion on diverse subjects, including politics, product feedback, social issues, and more.

1.1 GENERAL

Sentiment analysis is a useful tool for finding information from huge amounts of content, such as social media posts and product evaluations. It divides comments as good, negative, or neutral, providing an easily measured rating of public opinion. This enables groups to discover trends, track customer happiness, detect problems, and make data-driven decisions to improve offers and customer experiences.

1.2 OBJECTIVE

Comment analysis with sentiment analysis evaluates emotions and views in textual comments, providing information about public opinion, customer happiness, and company sentiment, allowing for more informed decision-making and a better customer experience.

1.3 EXISTING SYSTEM

Comment analysis systems use machine learning algorithms to automatically classify textual comments or reviews based on their sentiment. These systems collect relevant features, classify sentiment with trained models, and provide additional information through approaches such as sentiment level analysis or aspect-based sentiment analysis.

1.4 PROPOSED SYSTEM

The system uses sentiment analysis to classify comments based on their emotional tone, allowing them to be automatically classified as good, negative, or neutral. This enables companies, advertisers, and researchers to make data-driven decisions and efficiently respond to customer input, while also giving essential decision-making information.

CHAPTER 2

LITERATURE SURVEY

The study tells how to evaluate a hash tag's sentiment in tweets. 17,189 kaggle data points with particular hashtags. Combining lexicon-based and machine learning techniques, sentiments were mentioned; labels for machine learning trials were removed. 9,076 positive and negative data points were used in the machine learning modal. Six classifiers were used, with an 80-20 ratio between training and testing/validation, in Python 3.6 with sklearn. According to the results, the Support Vector Machine and Multilayer Perceptron classifiers were the best, with accuracies of 0.89 and 0.88 and F1 values of 0.8729 and 0.8647. For increased sentiment analysis accuracy, the study makes ideas for possible improvements through dataset improving and parameter tuning.

The paper shows an ordered sentiment analysis dataset with over 200,000 reviews in 22 domains that was obtained from social media comments. The dataset is an important resource for sentiment analysis tasks in the context of Chinese language learning, especially in the field of natural language processing. Three steps are involved in creating the dataset: collecting data from subjects with a high level of discuss and attention, preprocessing to remove unnecessary information, and manual annotation by academics. This unique Chinese text dataset features organized reviews and subject collections, making it easier to create and assess sentiment analysis models that make use of the structured data. This dataset can be used by researchers as an example set for evaluating machine learning, deep learning, or dictionary-based sentiment analysis techniques.

The study uses sentiment analysis of Twitter data to find out how the general public thinks about a variety of topics. Tweet sentiment is classified as positive, negative, or neutral using an ensemble classifier that combines Random Forest, Support Vector Machine, and Decision Tree models. After pre-processing and feature extraction using dimensionality reduction techniques like LDA and wrapper-based feature selection, data is gathered through the Twitter API. With the use of adaptive boosting, the recommended ensemble classifier exceeds other classifiers such as HL-NBC and ConvBiLSTM, reaching a high accuracy of 93.42%. This study highlights how crucial sentiment analysis is for understanding consumer opinions and helping companies make strategic decisions that will improve their products.

George Floyd died suddenly in police custody in May 2020, creating a strong effort against police abuse and an entire revolution. This study analyses the idea of digital citizen agency by analyzing the opinions posted on social media sites in the days following of Floyd's unfortunate loss, especially on the Black Lives Matter (BLM) page. The study shows the importance of enabling people in the age of digital media by showing how digital citizens use online spaces to voice their issues and fight for social change by applying big data and the selective exposure theory. Platforms like BLM allow users to actively participate in a variety of information and communication technologies (ICTs), which not only provide a platform for raising awareness about

This study focuses at the network leaders' expressions on Twitter can be used to determine the general mood of a community by using sentiment and emotion analysis. Focusing on the period range of February 27, 2020, to December 31, 2021, the study focuses at Spanish-language tweets and retweets on COVID-19. By the analysis of tweets only published by famous social media users, the research tries to identify the basic reason behind various points of view and choices. The reduction of authors to network leaders and then to related tweets is significant because it demonstrates the significant information compression that was accomplished. With a compression rate of over 99% for authors and 88% for related tweets, this provides crucial new information for the fields of business administration and healthcare social mood analysis.

In part to the development of online comments created by the quick development of Internet-based services, sentiment analysis (SA) has become an important instrument for academics, governments, and companies together. SA offers a quick, easy, and automated way to determine the opinions and feelings of reviewers. But the reviews of the literature that are now available only cover a small number of studies or focus on a specific area of sentiment analysis work. This comprehensive review of the literature provides light on the purpose of the sentiment analysis task, compares various approaches, looks into the application domains of sentiment analysis, identifies the challenges and limitations faced by researchers, makes suggestions for alternatives, and considers future directions for the field. The study shows the importance of sentiment analysis and the important role of artificial intelligence technologies in automatic text sentiment analysis.

In order to identify abuse on Twitter, this research provides a new approach involving contextual data from tweets that both before and follow abusive posts. This approach takes consideration of the surrounding context to better understand the intent behind abusive language, in contrast to existing methods that only focus on the content of individual abusive tweets. The study identifies the best feature and machine learning algorithm combination for abuse detection through a series of thorough trials. Compared to current methods, the suggested approach achieves an absolute improvement of about 7% in accuracy by testing eight alternative classifiers on content- and context-based variables. This innovative feature set addresses an important problem in ensuring safety and well-being of users on social media platforms by providing a more reliable and efficient method of identifying abuse on Twitter.

The area of government has slowly moved from the physical world to the digital one in the big data period. Players like the public and government can now communicate not only in the physical world but also practically through networks.

Government new media, as the government's online experience, combines the best features of new media communication with the functions of a government agency. Both qualities are essential to digital government. Government Affairs Weibo (GAW) was the first new media platform that the Chinese government used. Now that it has reached a somewhat advanced stage, the public view on it is comparably open. This platform has connected the domains of public opinion and official opinion.

The study offers a paradigm for language-focused sentiment analysis of tweets pertaining to news stories. It uses Natural Language Processing (NLP) approaches to connect tweets to news items and machine learning techniques including Naive Bayes, Complementary Naive Bayes, and Logistic Regression for tweet classification. To be more precise, the Zemberek NLP library is used for morphological analysis and stemming, and the bag-of-words approach is used for mapping. For sentiment analysis, 6000 tweets were gathered and manually classified in order to assess the framework. When it came to tweet classification, Naive Bayes performed admirably. A scalable classifier that runs on Apache Hadoop was developed as a result of the paper's emphasis on the necessity of real-time sentiment analysis.

This study looks into the implications of the NetzDG, a German law aimed at reducing hate speech on social media, with a focus on likely chilling and overblocking effects. Using an original dataset of Facebook posts and comments from ten public pages, the study looks at how user engagement and content deletion rates changed before, during, and following the law's application. The data shows not much evidence of over blocking or chilling effects, ignoring concerns about excessive content deletion and self-censorship. Rather, it shows a small increase in deleted comments per post after the law's full its operation. By exposing the difficulties of evaluating such policies in practice and offering empirical insights into the effects of platform regulation in the real world, this study adds to a wealth of literature.

CHAPTER 3

SYSTEM DESIGN

3.1 DEVELOPMENT ENVIRONMENT

3.1.1 HARDWARE SPECIFICATIONS

This project uses minimal hardware but in order to run the project efficiently without any lack of user experience, the following specifications are recommended

Table 3.1.1 Hardware Specifications

Component	**Specification**
Processor	Intel Core i5
RAM	4GB or above (DDR4 RAM)
GPU	Intel Integrated Graphics
Hard Disk	6GB
Processor Frequency	1.5 GHz or above

3.1.2 SOFTWARE SPECIFICATIONS

The software specifications in order to execute the project has been listed down in the below table. The requirements in terms of the software that needs to be preinstalled and the languages

needed to develop the project has been listed out below.

Table 3.1.2 Software Specifications

Category	Technologies/Tools
Front End	HTML, CSS, Bootstrap, JavaScript
Back End	Python, Django
Frameworks	Torch, TensorFlow
IDE/Software	Visual Studio, Jupiter Notebook

3.2 SYSTEM DESIGN

3.2.1 ARCHITECTURE DIAGRAM

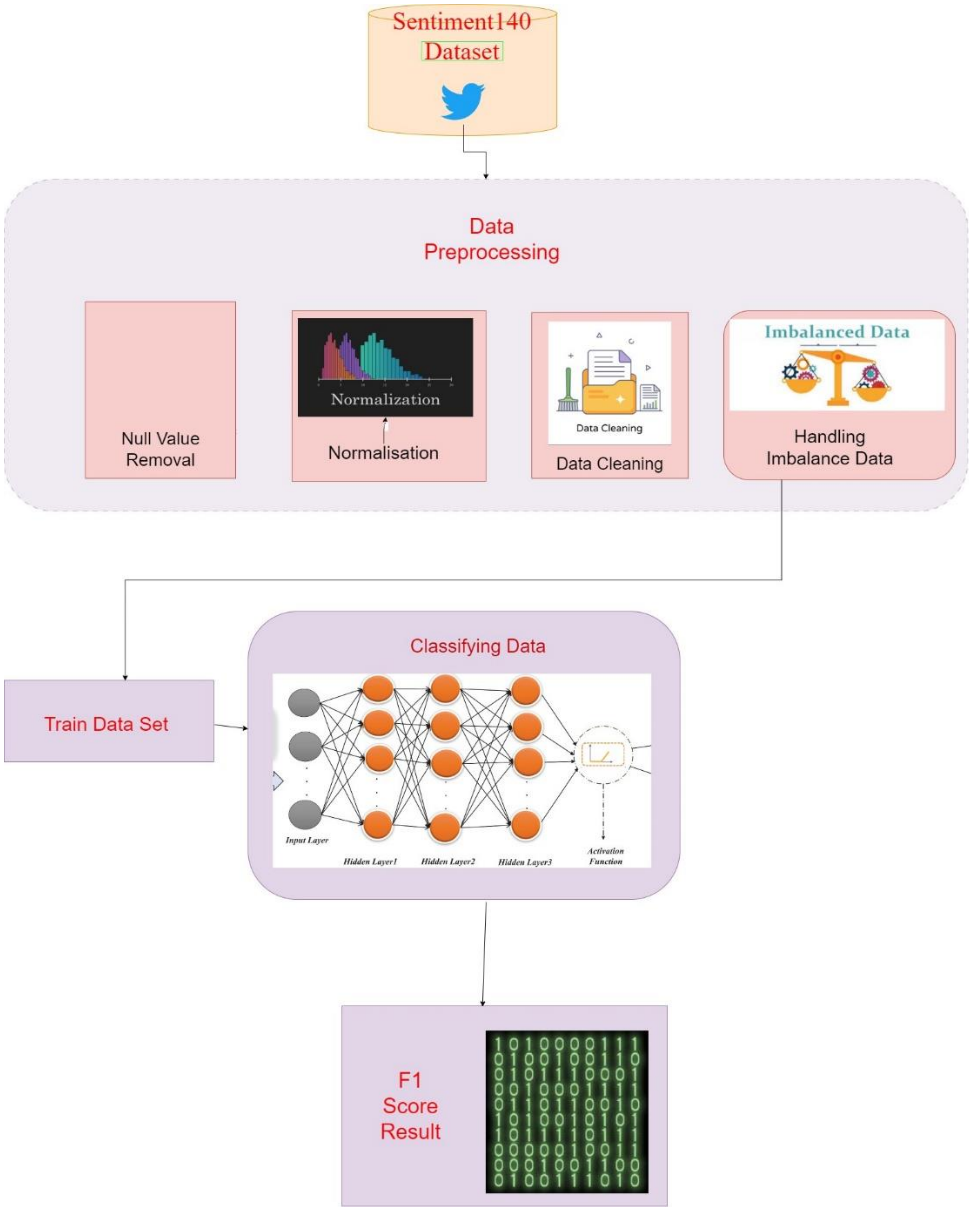


Fig 3.2.1 Comment Analysis Model

PRE-PROCESSING:

Expanding the model's capacity for estimation by adding new features or changing the existing ones. This could involve creating polynomial features, communication terms, or feature extraction from date or text variables. Combining data from several sources into a single dataset and providing connectivity and consistency across datasets. Decreasing the dataset's feature weigh to boost the accuracy of processing and lower the chance of over fitting. This can be achieved by using methods such as methods of selecting features or principal component analysis (PCA).

By doing the execution of these preprocessing procedures, analysts can ensure that the data is correctly formatted for analysis, leading in more exact and trusted results from machine learning models.

TRAINING SET:

The train dataset's variety and quality of data have significant impacts on the sentiment analysis model's function and ability for generalization. The model is able to recognize reliable patterns of sentiment phrase with the help of a varied dataset spreading a broad range of topics, areas, and writing types. Better labels further ensure that the model develops knowledge from exact ground truth sentiment annotations. In sentiment analysis, the train dataset gives the data needed for training, evaluation, and efficiency, and it is the base upon which reliable model constructing is built. The ultimate sentiment analysis system's effectiveness and ability for generalization is greatly affected by its quality, variety, and the training techniques used.

CHAPTER 4

PROJECT DESCRIPTION

4.1 MODULE DESCRIPTION

4.1.1 DATA PRE-PROCESSING:

Expanding the model's capacity for estimation by adding new features or changing the existing ones. This could involve creating polynomial features, communication terms, or feature extraction from date or text variables. Combining data from several sources into a single dataset and providing connectivity and consistency across datasets. Decreasing the dataset's feature weigh to boost the accuracy of processing and lower the chance of over fitting. This can be achieved by using methods such as methods of selecting features or principal component analysis (PCA). Doing the execution of these preprocessing procedures, analysts can ensure that the data is correctly formatted for analysis, leading in more exact and trusted results from machine learning models.

4.1.2 TRAINING SET:

The train dataset's variety and quality of data have significant impacts on the sentiment analysis model's function and ability for generalization. The model is able to recognize reliable patterns of sentiment phrase with the help of a varied dataset spreading a broad range of topics, areas, and writing types. Better labels further ensure that the model develops knowledge from exact ground truth sentiment annotations. In sentiment analysis, the train dataset gives the data needed for training, evaluation, and efficiency, and it is the base upon which reliable model constructing is built. The ultimate sentiment analysis system's effectiveness and ability for generalization is greatly affected by its quality, variety, and the training techniques used.

4.1.3 NORMALIZATION:

Normalization is a vital part of database design that improves data integrity, improves storage, and decreases the risk of anomalous data. The creation of dependable, in effect databases that fulfill user needs is an option for database designers who follow to the normalizing rules. To be able to ensure that the text data is reliable, consistent, and ready for analysis, data cleaning is a vital step in the sentiment analysis the creation process.

4.1.4 DATA CLEANING:

This includes handling special characters and symbols, turning all characters to lowercase, and removing punctuation from the text so as to format it regularly. Splitting a text into smaller parts called tokens, like words or phrases, to make analysis simpler. This phase helps in text consistency and gets it ready for the next step. Removing terms that are frequently used but don't really mean anything, like "and," "the," and "is." Removing stop words improves the focus on sentiment-expressing words while decreasing data noise. Both approaches seek to reduce words to their most basic. Reducing words to their base or root form is the goal of both techniques.

4.1.5 TRAINING MODEL CLASSIFIER:

The method of identifying textual inputs into predefined sentiment categories—typically positive, negative, or neutral—is known as sentiment analysis the categorization This is an extensive description of the classification procedure. The written data has been preprocessed to create standards and clean the input before classification. Stop words—common words with little linguistic meaning, like "and" or "the"—and special characters, punctuation, and other noise may need to be eliminated in order accomplish this. The text can also be designated to divide it into individual words or n-grams, which are word sequences.

4.1.6 HANDLING IMBALANCE DATA:

In sentiment analysis, handling unbalanced data is vital to producing accurate and reliable outcomes. When one sentiment group occurs in the dataset, it generates unbalanced information, which results in biased models that fall short of on minority classes. The following are some methods to deal with this problem. Reducing involves not enough or over sampling the majority class, depending on which is chosen. Copying minority class cases or developing artificial samples with methods like SMOTE (Synthetic Minority Over-sampling Technique) are instances of the oversampling techniques. To equalize the data set, under sampling decreases the number of cases of the majority class.

4.1.7 F1 SCORE:

The F1 score gives an honest assessment of the model's performance through putting both recall and accuracy into a single number. It is computed as the precision and know harmonic means, with the same weight assigned to each metric. As such, models with low recall or accuracy are affected by the F1 score, resulting in a balance among the two. A high F1 score is often chosen in sentiment analysis applications as it shows that the model reduces false positives and false negatives while correctly recognizing positive, negative, and neutral sentiment in the text. An F1 score is a helpful indicator to measure the overall effectiveness of sentiment analysis systems because it shows that the model is capable of predicting sentiment across different types of written content. To sum up, the F1 score is an important sentiment analysis metric that impacts a balance between accuracy and recall, providing a thorough assessment of the model's ability of recognizing sentiment in textual data. A high F1 score indicates solid accuracy across various sentiment categories, as the model achieves both high accuracy and high memory.

CHAPTER 5

IMPLEMENTATION AND RESULTS

5.1 IMPLEMENTATION

The study begins with data collecting and preprocessing and then focuses on sentiment analysis in comment analysis. Customer reviews, internet events, and social media platforms were the sources of a variety of comments. Lemmatization and stemming were used in the preprocessing of the comments to organize the text and eliminate noise. SVM, Naive Bayes, and Random Forest classifiers were among the machine learning and deep learning approaches used for sentiment analysis. Using methods like word embeddings and TF-IDF, each remark was represented as a number vector.

We investigated convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their relatives, such as LSTM and GRU, as well as other deep learning architectures like Torch and TensorFlow. With each comment labelled with the appropriate sentiment, the models were trained using labelled data. Order to improve model performance and provide stability against overfitting, which cross-validation techniques were employed, along with hyperparameter change.

5.2 RESULTS

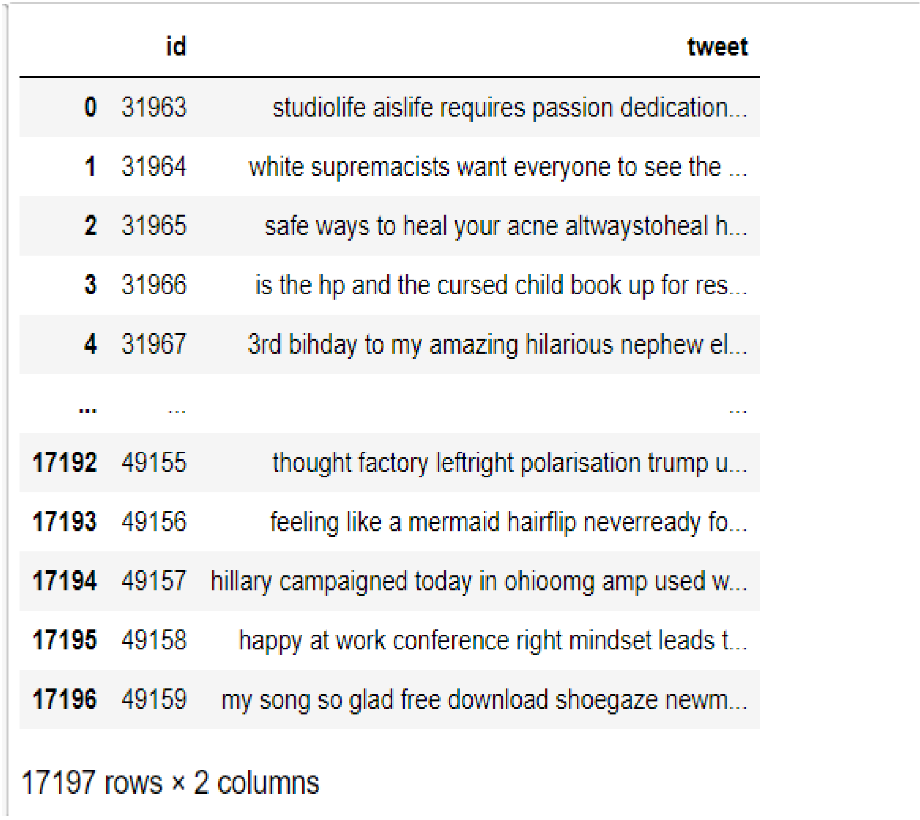
We analyzed the trained models and found that the accuracy, precision, recall, and F1-score measures of sentiment categorization showed encouraging results. The machine learning models performed reasonably, particularly when they were trained on large and complete datasets. At the same time, deep learning models showed greater skill in obtaining complex language details and contextual data, resulting in improved sentiment analysis precision.

Moreover, real-time monitoring and analysis of comments across many platforms was made possible by the sentiment analysis system's implementation. These insights could be used by companies and organizations to determine areas for improvement, measure consumer satisfaction, and get a sense of public opinion. Based on the sentiment trends found, the technology enabled informed choices and responsive actions.

Overall, it was successful in gathering useful details from textual data by implementing sentiment analysis in comment analysis.

5.2 OUTPUT SCREENSHOTS

The analysis explains the sentiment of positive or negative comment of a hashtag. Initially by cleaning the data by removing unwanted symbols and transforming the tweets which are understood by ML Model. On the next then un balanced data are balanced by using sampling process (under sampling/over sampling). The analysis on Comment Analysis in Tweets gives around the values of 0 or 1 according to the hate level of tweets. If the value comes less than 0 then the result is a positive comment else it is a negative comment which shows the F1 values is more than 1.



	id	tweet
0	31963	studiolife aislife requires passion dedication...
1	31964	white supremacists want everyone to see the ...
2	31965	safe ways to heal your acne altwaystoheal h...
3	31966	is the hp and the cursed child book up for res...
4	31967	3rd bihday to my amazing hilarious nephew el...
...	...	...
17192	49155	thought factory leftright polarisation trump u...
17193	49156	feeling like a mermaid hairflip neverready fo...
17194	49157	hillary campaigned today in ohioomg amp used w...
17195	49158	happy at work conference right mindset leads t...
17196	49159	my song so glad free download shoegaze newm...

17197 rows × 2 columns

Fig 5.2.1 Cleaned Data Set

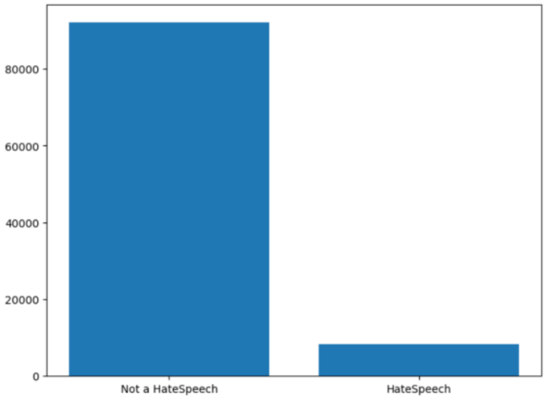


Fig 5.2.2 Before Sampling

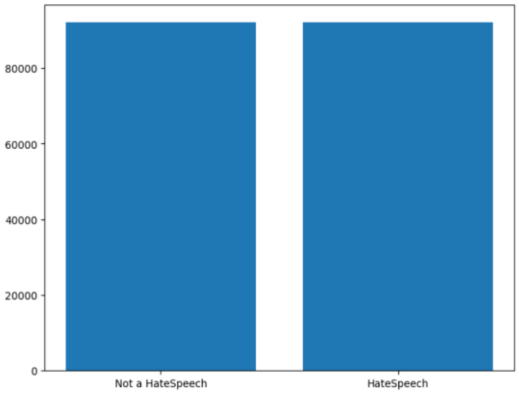
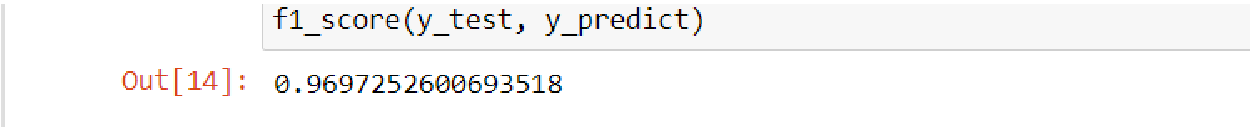


Fig 5.2.3 After Sampling



```
f1_score(y_test, y_predict)
```

```
Out[14]: 0.9697252600693518
```

Fig 5.2.4 F1 Score for the following tweets

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENTS

6.1 CONCLUSION

Since Twitter is one of the biggest names in social media, plenty of studies has been done on sentiment analysis using the tweets. In this research, we discuss the initial findings of the system we propose that takes subjects from tweets using natural language processing (NLP) and describes tweet a component's by looking at sentiment words linked to subjects. The results of the research show that the suggested solution performs more effectively than the Alchemy API, but enhancements are still needed because SVM exceeds it. Further research are intended to improve the precision of sentiment analysis. It's also important to keep in mind that, because language and misspelled words are frequently found in tweets, obtaining sentiment words without first regulating them can be difficult.

6.2 FUTURE ENHANCEMENTS

In the future, sentiment analysis machine learning projects aimed at comment analysis could improve accuracy, adaptability, and utility by: refining the model on larger, more varied datasets; introducing aspect-based analysis for more detailed insights; integrating multimodal analysis to incorporate emojis, images, or audio; improving contextual understanding for sarcasm and cultural nuances; enabling real-time analysis for instant feedback; offering user customization options; supporting entity-level sentiment analysis; expanding multilingual support; putting in place continuous learning mechanisms; and integrating with other natural language processing tools for increased functionality.

APPENDIX

SOURCE CODE:

```
import pandas as pd
train = pd.read_csv('train.csv')
print("Training Set:"% train.columns, train.shape, len(train))
test=pd.read_csv('test.csv')
print("Test Set:"% test.columns, test.shape, len(test))
import re
def  clean_text(df, text_field):
    df[text_field] = df[text_field].str.lower()
    df[text_field] = df[text_field].apply(lambda elem: re.sub(r"(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)|^rt|http.+?", "", elem))
    return df
test_clean = clean_text(test, "tweet")
train_clean = clean_text(train, "tweet")
test_clean
train_clean
import seaborn as sns
import matplotlib.pyplot as plt
fig = plt.figure()
ax = fig.add_axes([0,0,1,1])
langs = ['Not a HateSpeech','HateSpeech']
data = [len(train_clean[train_clean.label==0]),len(train_clean[train_clean.label==1])]
ax.bar(langs,data)
plt.show()
from sklearn.utils import resample
train_majority = train_clean[train_clean.label==0]
train_minority = train_clean[train_clean.label==1]
train_minority_upsampled = resample(train_minority,
                                    replace=True,
                                    n_samples=len(train_majority),
                                    random_state=123)
train_upsampled = pd.concat([train_minority_upsampled, train_majority])
train_upsampled['label'].value_counts()
fig = plt.figure()
```

```
ax = fig.add_axes([0,0,1,1])
langs = ['Not a HateSpeech','HateSpeech']
data =
[len(train_upsampled[train_upsampled.label==0]),len(train_upsampled[train_upsampled.la
bel==1])]
ax.bar(langs,data)
plt.show()
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.linear_model import SGDClassifier
pipeline_sgd = Pipeline([
    ('vect', CountVectorizer()),
    ('tfidf',  TfidfTransformer()),
    ('nb', SGDClassifier()),])
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(train_upsampled['tweet'],
                                                    train_upsampled['label'],random_state = 0)
model = pipeline_sgd.fit(X_train, y_train)
y_predict = model.predict(X_test)
from sklearn.metrics import f1_score
f1_score(y_test, y_predict)
```

REFERENCES

[1]Çam, H., Cam, A., Demirel, U., & Ahmed, S. (2024, January 1). Sentiment analysis of financial Twitter posts on Twitter with the machine learning classifiers. Heliyon. https://doi.org/10.1016/j.heliyon.2023.e23784

[2]Zhang, W., & Zhang, S. (2024, February 1). A Structured Sentiment Analysis Dataset Based on Public Comments from Various Domains. Data in Brief. https://doi.org/10.1016/j.dib.2024.110232

[3]Kp, V., Ab, R., Hl, G., Ravi, V., & Krichen, M. (2024, May 1). A Tweet Sentiment Classification Approach using an Ensemble Classifier. International Journal of Cognitive Computing in Engineering. https://doi.org/10.1016/j.ijcce.2024.04.001

[4]Scotland, J., Thomas, A., & Jing, M. (2024, April 1). Public Emotion and Response immediately following the Death of George Floyd: A Sentiment Analysis of Social Media Comments. Telematics and Informatics Reports. https://doi.org/10.1016/j.teler.2024.100143

[5]López, J. N., Aguarón, J., Moreno-Jiménez, J. M., & Turón, A. (2024, January 1). Social mood during the Covid-19 vaccination process in Spain. A sentiment analysis of tweets and social network leaders. Heliyon. https://doi.org/10.1016/j.heliyon.2023.e23958

[6]Mao, Y., Liu, Q., & Zhang, Y. (2024, April 1). Sentiment analysis methods, applications, and challenges: A systematic literature review. Journal of King Saud University. Computer and Information Sciences/Mağalaẗ Ğam’aẗ Al-malīk Saud : Ùlm Al-ḥasib Wa Al-ma’lumat. https://doi.org/10.1016/j.jksuci.2024.1020

[7]Hussain, K., Saeed, Z., Abbasi, R. A., Sindhu, M. A., Khattak, A. S., Arafat, S., Daud, A., & Mushtaq, M. (2024, April 1). Towards Understanding the Role of Content-based and Contextualized Features in Detecting Abuse on Twitter. Heliyon. https://doi.org/10.1016/j.heliyon.2024.e29593

[8]Li, M., & Shi, Y. (2023, August 1). Sentiment analysis and prediction model based on Chinese government af airs microblogs.Heliyon. https://doi.org/10.1016/j.heliyon.2023.e19091

[9]A Scalable Approach for Sentiment Analysis of Turkish Tweets and Linking Tweets to News. (2016, February 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/7439388

[10]Maaß, S., Wortelker, J., & Rott, A. (2024, February 1). Evaluating the regulation of social media: An empirical study of the German NetzDG and Facebook. Telecommunications Policy. https://doi.org/10.1016/j.telpol.2024.102719