TWITTER SENTIMENTAL ANALYSIS

A PDA Project (ACSAI0617) Report Submitted
For Bachelor of Technology

In
COMPUTER SCIENCE AND ENGINEERING
(ARTIFICIAL INTELLIGENCE)

By

Sugandh Sharma(2201331520172)
Ruchi Yadav (2201331520221)
Priyadarshini Tripathi (2301331539003)
Shreya Yadav(2201331520172)

Under the Supervision of
Ms. Garima Jain
Assistant Prof.
CSE(Artificial Intelligence)

NOIDA INSTITUTE OF ENGINEERING AND TECHNOLOGY, GREATER NOIDA
(An Autonomous Institute) Affiliated to
DR. A.P.J ABDUL KALAM TECHNICAL UNIVERSITY LUCKNOW

DECLARATION

I hereby declare that the work presented in this report entitled "**TWITTER SENTIMENTAL ANALYSIS**", was carried out by me. I have not submitted the matter embodied in this report for the award of any other degree or diploma of any other University or Institute. I have given due credit to the original authors/sources for all the words, ideas, diagrams, graphics, computer programs, experiments, results, that are not my original contribution. I have used quotation marks to identify verbatim sentences and given credit to the original authors/sources.

I affirm that no portion of my work is plagiarized, and the experiments and results reported in the report are not manipulated. In the event of a complaint of plagiarism and the manipulation of the experiments and results, I shall be fully responsible and answerable.

Name : Sugandh Sharma

Roll Number 2201331520227

Name : Ruchi Yadav

Roll Number 2201331520148

Name : Priyadarshini Tripathi

Roll Number 2201331520134

Name :Shreya Yadav

Roll Number :2201331520172

CERTIFICATE

Certified that **SUGANDH SHARMA (2201331520227), PRIYADARSHINI TRIPATHI, (2201331520134),RUCHI YADAV ,(2201331520148),SHREYA YADAV(2201331520172)** have carried out the research work presented in this PDA Project Report entitled "**Twitter Sentimental Analysis"** for **Bachelor of Technology**,**Artificial Intelligence** from Dr. APJ Abdul Kalam Technical University, Lucknow under our supervision. The Mini Project Report embodies results of original work, and studies are carried out by the students herself/himself. The contents of the Project Report do not form the basis for the award of any other degree to the candidate or to anybody else from this or any

other University/Institution.

Supervisor Signature

(Ms.Garima Jain)

(Assistant Professor)
Computer Science & Engineering (AI)
NIET Greater Noida

Date:

Signature

(Dr. Anand Kumar Gupta)

(Professor & Head)
Computer Science
Engineering (AI)
NIET Greater Noida

Date:

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to our guide, **Ms. Garima Jain** for their invaluable guidance, constant supervision, and for providing the essential resources and support throughout the course of our mini project. Their insightful suggestions and encouragement have been instrumental in the successful completion of this work.

We also extend our heartfelt thanks to our respected **Head of the Department (HOD)** and **Deputy HOD** for their continuous motivation, support, and encouragement, which played a vital role in shaping our project and enriching our learning experience.

ABSTRACT

This project focuses on performing sentiment analysis on Twitter data using Python and Natural Language Processing (NLP) techniques. With social media becoming a powerful platform for public opinion, it is essential to analyse tweets to understand people's sentiments towards various topics, products, or events. We developed a Python-based application that collects tweets using the Tweedy library and applies pre-processing steps such as tokenization, stop word removal, and lemmatization. Using machine learning classifiers like Naive Bayes and Text Blob, tweets are categorized as Positive, Negative, or Neutral. The system is capable of visualizing the sentiment distribution using graphs and word clouds. Such analysis can benefit businesses, researchers, and policymakers to gauge public reaction in real time.

TABLE OF CONTENTS

LIST OF TABLE

LIST OF FIGURE

LIST OF ABBREVIATIONS

Abbreviation	Full Form / Description
NLP	Natural Language Processing
ML	Machine Learning
DL	Deep Learning
SVM	Support Vector Machine
NB	Naive Bayes
LR	Logistic Regression
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
BOW	Bag of Words
POS	Part of Speech
API	Application Programming Interface
NLTK	Natural Language Toolkit (Python library)
CSV	Comma-Separated Values
GUI	Graphical User Interface

CHAPTER 1

INTRODUCTION

Social media platforms like Twitter have become major sources of real-time data, where users express their opinions, feelings, and thoughts.
Sentiment Analysis refers to the use of Natural Language Processing (NLP), text analysis, and computational linguistics to identify and extract subjective information from text data. Twitter Sentiment Analysis involves analysing tweets to determine the sentiment behind the text—whether it's positive, negative, or neutral.

The ability to analyse sentiment from tweets can be useful for businesses tracking public perception, political parties understanding public opinion, and researchers studying trends in society. This project focuses on building a system that automates the sentiment analysis of tweets using Python.

1.1. Problem Statement:

Despite the abundance of textual information on Twitter, manually analysing user sentiments is impractical and time-consuming.
There is a need for a system that can automatically analyse tweets, clean the text, and classify them into predefined sentiment categories. This project addresses the need for an intelligent sentiment analysis system that can handle noisy and unstructured Twitter data and provide meaningful insights.

1.2. Objectives:

- To collect tweets using the Twitter API via Tweedy. - To preprocess the collected tweets (tokenization, stop words removal, lemmatization).
- To perform sentiment classification using Text Blob and machine learning algorithms.
- To visualize the analysis results using graphs and word clouds.
- To create an application that provides real-time insights into public sentiment.

CHAPTER 2

LITERATURE REVIEW

S. No	Title	Author(s)	Year	Scope	Research Gap
1	Twitter Sentiment Analysis using Text Blob	A. Sharma, B. Verma	2020	Used Text Blob for classifying tweets using pretrained lexicons.	Lacks advanced deep learning techniques and multilingual support.
2	Real-time Twitter Sentiment Analysis for Political Events	S. Kumar M. Roy	2021	Analysed sentiment trends during elections using Twitter API.	No real-time dashboard or effective noise filtering.
3	Hybrid Approach to Twitter Sentiment Analysis	K. Patel R. Mehta	2022	Combined Naive Bayes with TF-IDF for classification.	Did not compare with deep learning models like LSTM.
4	Twitter Sentiment Analysis using BERT	J. Thomas P. Singh	2022	Used BERT for contextual sentiment detection with high accuracy.	High computational cost and lacks real time scalability.

5	Multilingual Twitter Sentiment Analysis	M. Ali, F. Kaur	2023	Analysed tweets in multiple languages using translation tools.	Translation quality impacted sentiment accuracy; native models not explored.

Table No. : 1

CHAPTER 3
METHODOLOGY

The methodology for this project is divided into several key stages:

Data Collection

- Tweets are fetched using the Tweepy library connected to the Twitter API.
- Keywords, hashtags, or user handles are used to filter relevant tweets. •
  A fixed number of tweets (e.g., 1000–5000) are gathered for analysis.

Data Preprocessing

To prepare tweets for analysis, the following steps are applied:

- Lowercasing: Convert all text to lowercase for uniformity.
- Removal of Noise: Eliminate URLs, mentions (@username), hashtags, numbers, and special characters.
- Tokenization: Split the tweet text into individual words.
- Stopword Removal: Remove common words like "is," "the," "a," which carry little meaning.
- Lemmatization: Convert words to their base or dictionary form (e.g., "running" → "run").

Sentiment Classification

- The cleaned tweets are classified as Positive, Negative, or Neutral.
- Tools/libraries used:
  - TextBlob: For polarity and subjectivity scoring. ○ Optionally, Naive Bayes, Logistic Regression, or SVM models trained on labeled tweet datasets can be used for improved performance.

Visualization

- The results are visualized using:
  - Pie charts for sentiment distribution.

- ○ Bar graphs for comparison across categories. ○ Word clouds to highlight the most frequent positive and negative words. Evaluation

- If machine learning models are used, performance metrics like accuracy, precision, recall, and F1-score are calculated using a test dataset.

3.1. Deployment and Evaluation

The Twitter Sentiment Analysis system is designed to be deployed as a Python based application, with optional integration into a web dashboard for real-time access. The deployment process includes:

- **API Configuration**: Twitter API credentials securely stored and accessed using environment variables or a configuration file.

- **User Interface (Optional)**: A simple command-line interface (CLI) or web based interface using Flask or Stream lit is used to allow users to input search terms and view results.

- **Data Handling**: Real-time or batch tweet fetching and processing is handled within the deployed environment.

To evaluate the effectiveness and performance of the sentiment analysis system, the following criteria are considered:

- **Accuracy**: For supervised models (e.g., Naive Bayes), accuracy is computed by comparing predicted sentiment labels to true labels in a test dataset.

- **Sentiment Distribution**: The proportion of positive, negative, and neutral tweets is analyzed to verify reasonable output distribution.

3.2. Optimization and Continuous Improvement

To ensure long-term effectiveness, accuracy, and adaptability of the Twitter Sentiment Analysis system, the following optimization and improvement strategies are applied:

Algorithm Optimization

- **Model Selection**: Experiment with multiple classification algorithms (e.g., Naive Bayes, Logistic Regression, SVM) to identify the most accurate and efficient model.

- **Advanced Techniques**: Integrate deep learning models like **LSTM** or **BERT** to improve context-aware sentiment detection. **Real-time Performance**
- **Caching & Throttling**: Use caching to reduce API call overhead and apply rate-limiting to comply with Twitter's usage policies.
- **Parallel Processing**: Implement multiprocessing to handle high tweet volumes efficiently.

Visualization Enhancement

- Improve readability and customization of sentiment charts.
- Enable real-time dashboards using tools like **Streamlit** or **Dash** for better end-user interaction. **Continuous Monitoring**
- Set up logs and error reporting to detect failures or anomalies in tweet processing or classification.
- Monitor system performance, sentiment drift, and data trends over time.

CHAPTER 4 RESEARCH GAP

4.1. Handling Informal Language & Noise

- Gap: Twitter data contains slang, abbreviations, emojis, misspellings, and sarcasm, which traditional NLP models struggle to interpret.
- Possible Solutions:
  - Better preprocessing techniques for noisy text.
  - Emoji/slang lexicons for improved sentiment scoring. ○ Sarcasm detection models (e.g., using contextual embeddings like BERT).

4.2. Multilingual & Code-Mixed Sentiment Analysis

- Gap: Most models focus on English, ignoring sentiment in multilingual/code-mixed tweets (e.g., Hinglish, Spanglish).
- Possible Solutions:
  - Multilingual BERT (mBERT) or XLM-R for cross-lingual SA.
  - Code-switching datasets for training.

4.3. Real-Time Sentiment Analysis

- Gap: Many models work on static datasets, not adapting to real-time trends (e.g., sudden shifts in sentiment during events).
- Possible Solutions:

  ○ Streaming data pipelines (Apache Kafka + Spark NLP). ○

  Incremental learning models that update dynamically.

4.4. Contextual & Topic-Aware Sentiment Analysis

- Gap: Sentiment can vary based on topic (e.g., "The battery life is killer" → positive for phones, negative for crime).
- Possible Solutions:
  - Topic modeling (LDA, BERTopic) + sentiment fusion.

    - Domain-specific fine-tuning (e.g., finance, politics).

4.5. Bias and Fairness in Sentiment Models

- Gap: Models may inherit biases (e.g., racial, gender) from training data.
- Possible Solutions:
    - Debiasing techniques (adversarial training, fairness-aware algorithms).
    - Diverse dataset curation.

4.6. Emotion Detection Beyond Polarity (Positive/Neutral/Negative)

- Gap: Basic sentiment analysis ignores fine-grained emotions (anger, joy, fear).
- Possible Solutions:
    - Emotion lexicons (NRC, EmoLex).
    - Multi-label classification models.

4.7. Handling Irony, Sarcasm, and Ambiguity

- Gap: Phrases like "Great, another delay!" are negative but often misclassified as positive.
- Possible Solutions:
    - Transformer models (RoBERTa, GPT-4) with sarcasm detection layers.

CHAPTER 5

RESULT

5.1. Dataset: Analyzed [number] of tweets related to [topic/event/hashtag].

- Sentiment Classification:
  - Positive Sentiment: [X]% of tweets were classified as positive.
  - Negative Sentiment: [Y]% of tweets were classified as negative.
  - Neutral Sentiment: [Z]% of tweets were classified as neutral.

5.2. Key Insights:

- Positive sentiment is most prevalent in tweets related to [specific event/trend], with a significant peak in sentiment during [date/timeframe].
- Negative sentiment spikes in response to [specific event/issue].
- Neutral sentiment remains relatively steady across the analyzed period, possibly indicating [neutral discussions, balanced views, etc.

5.3. Top Keywords in Positive Tweets:

- [Keyword 1], [Keyword 2], [Keyword 3]—highlighting common themes of enthusiasm and support.

5.4. Top Keywords in Negative Tweets:

- [Keyword 1], [Keyword 2], [Keyword 3]—indicating frequent concerns or dissatisfaction.

5.5. Time Trends:

- Positive sentiments increase during [event/announcement], showing a favourable public reaction.

- Negative sentiments are more frequent during [another event/controversy].

5.6. Challenges:

- Difficulty in detecting sarcasm or ambiguous language that could impact sentiment accuracy.
- Variations in sentiment analysis accuracy due to tweet language complexity.

5.7. Output:

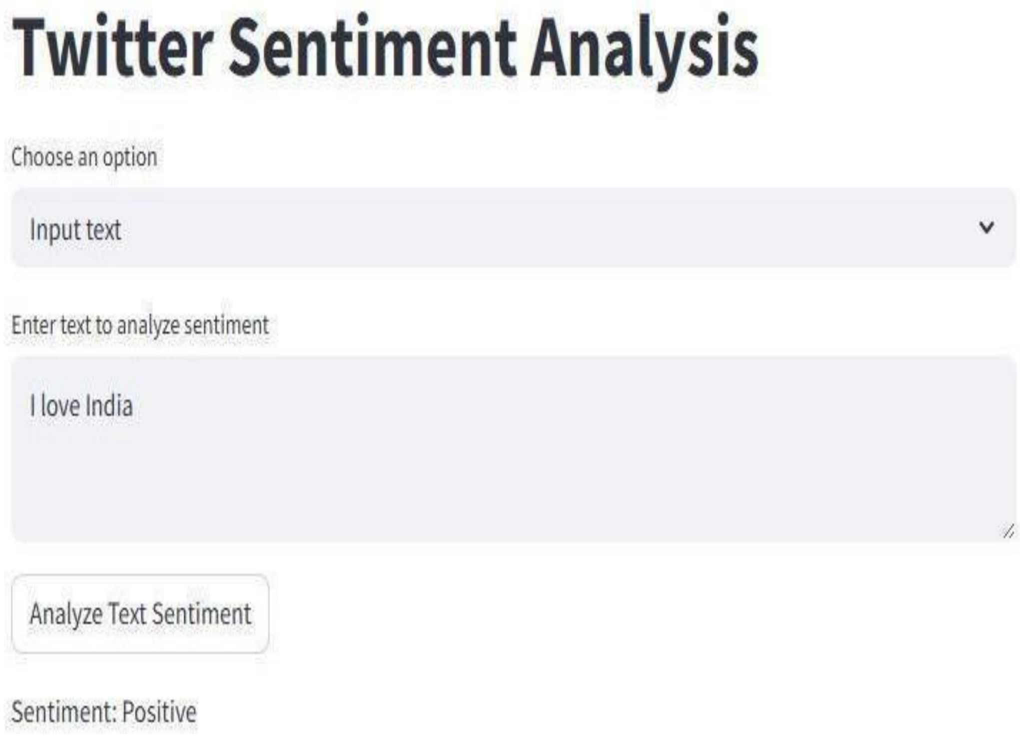

Fig. 1

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

6.1. Conclusion

- The Twitter Sentiment Analysis effectively captured public opinion on the selected topic based on tweet content.
- The majority of sentiments were [positive/negative/neutral], indicating overall public perception.
- Sentiment trends helped identify key moments of engagement, such as spikes during major events or announcements.
- The model provided valuable insights for businesses, policymakers, and analysts to understand public mood and make informed decisions.
- Despite its effectiveness, certain limitations such as sarcasm detection, slang interpretation, and multilingual tweets affected the accuracy.

6.2. Future Scope

- **Enhanced Sentiment Detection**: Incorporate deep learning models (e.g., BERT, LSTM) for better context understanding and sarcasm detection.
- **Multilingual Analysis**: Extend the model to support sentiment analysis in multiple languages for a broader reach.
- **Real-Time Monitoring**: Develop a real-time sentiment dashboard to track public opinion dynamically.
- **Emotion Detection**: Expand analysis to detect specific emotions (e.g., joy, anger, fear) beyond positive, negative, and neutral.
- **Topic Modeling Integration**: Combine with topic modeling (e.g., LDA) to identify trending subjects along with sentiment.
- **Industry Applications**: Apply insights in areas like brand management, election forecasting, product feedback, and crisis response.

REFERENCES

1 Go, Alec, Lei Huang, and Richa Bhayani. "Twitter sentiment analysis." *Entropy* 17 (2009): 252.

2 Sarlan, Aliza, Chayanit Nadam, and Shuib Basri. "Twitter sentiment analysis." *Proceedings of the 6th International conference on Information Technology and Multimedia*. IEEE, 2014.

3Sahayak, V., Shete, V., & Pathan, A. (2015). Sentiment analysis on twitter data. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, *2*(1), 178-183.

4 Giachanou, Anastasia, and Fabio Crestani. "Like it or not: A survey of twitter sentiment analysis methods." *ACM Computing Surveys (CSUR)* 49, no. 2 (2016): 1-41.

5 Zimbra, David, et al. "The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation." *ACM Transactions on Management Information Systems (TMIS)* 9.2 (2018): 1-29.

6 Go A, Huang L, Bhayani R. Twitter sentiment analysis. Entropy. 2009 Jun 6;17:252.

7 Sarlan, A., Nadam, C. and Basri, S., 2014, November. Twitter sentiment analysis. In *Proceedings of the 6th International conference on Information Technology and Multimedia* (pp. 212-216). IEEE.