A DSBDAL MINI-PROJECT REPORT

on

"Social Media Sentiment Analysis"

Submitted to the

Pune Institute of Computer Technology, Pune

In partial fulfilment for the award of the Degree of

Bachelor of Engineering

in

Information Technology

by

Samali Rajderkar	33265
Chinmay Raskar	33266
Shraddha Isokar	33274
Veera Subandh	33277

Under the guidance of

Mrs. Swapnaja Hiray

Department Of Information Technology

Pune Institute of Computer Technology College of Engineering

Sr. No 27, Pune-Satara Road, Dhankawadi, Pune - 411 043.

2024-2025

SCTR's PUNE INSTITUTE OF COMPUTER TECHNOLOGY

DEPARTMENT OF INFORMATION TECHNOLOGY

CERTIFICATE

This is to certify that the project report entitled

Social Media Sentiment Analysis

Submitted by

Samali Rajderkar	**33265**
Chinmay Raskar	**33266**
Shraddha Isokar	**33274**
Veera Subandh	**33277**

is a bonafide work carried out by them under the supervision of **Prof. Swapnaja Hiray** and it is approved for the partial fulfillment of the requirement of **Data Science and Big Data Analysis Lab** for the award of the Degree of Bachelor of Engineering (Information Technology)

Prof.Swapnaja Hiray	**Dr. A. S. Ghotkar**	**Dr. S. T. Gandhe**
Mini-Project Guide	Head of Department	Principal
Department of Information Technology	Department of Information Technology	SCTR's Pune Institute of Computer Technology, Pune

Place: PICT,Pune **Date:16 February 2025**

CONTENT

ABSTRACT

The rapid expansion of social media platforms has resulted in an overwhelming amount of user-generated content that provides valuable insights into public opinions, emotions, and behaviors. Sentiment analysis is a key technique used to analyze this content, helping businesses, researchers, and governments understand the general mood of users, detect trends, and make data-driven decisions. This project aims to explore sentiment analysis applied to social media posts, with a particular focus on analyzing sentiments expressed in posts across different platforms such as Twitter, Instagram, and Facebook. The dataset used in this study includes posts with sentiment labels, ranging from Positive, Negative, and Neutral to other emotional categories like Anger, Sadness, and Fear.

The project begins with the process of data cleaning and preprocessing, which involves handling missing values, correcting inconsistencies, and converting categorical sentiment labels into numerical values. The textual data is further processed using natural language techniques, such as tokenization and stop word removal, to prepare it for sentiment classification. Exploratory data analysis (EDA) is conducted to uncover trends, such as sentiment distributions across platforms, geographical regions, and time periods. Specific attention is given to sentiment trends over time, identifying patterns and shifts in sentiment based on dates, which may reveal interesting correlations with external events or seasons.

A machine learning model is then developed to predict sentiment based on the text content of the posts. Various models, including logistic regression, support vector machines, and neural networks, are evaluated based on their performance in terms of accuracy, precision, recall, and F1 score. The best-performing model is selected and fine-tuned using cross-validation techniques to improve its prediction capabilities.

Finally, the results of the sentiment analysis are visualized, providing insights into sentiment shifts over time, sentiment distribution across platforms, and the overall emotional tone of the data. The visualizations highlight not only the effectiveness of the sentiment classification model but also the broader trends in public sentiment. This analysis has practical implications in several domains, including brand monitoring, political sentiment tracking, and social issue awareness, where understanding the emotional tone of large user populations can drive decisions and strategies.

INTRODUCTION

With the advent of social media, the volume of user-generated content has exploded, creating an invaluable resource for understanding public opinions, consumer behaviors, and societal trends. Platforms like Twitter, Instagram, and Facebook generate massive amounts of text data every day, which can reveal emotional and psychological insights into users' thoughts and reactions to a wide range of topics. In this data-driven era, extracting meaning from this enormous pool of unstructured text is a significant challenge, yet it holds the key to a deeper understanding of human sentiment and emotions. This process is where sentiment analysis, a branch of natural language processing (NLP), plays a pivotal role.

Sentiment analysis refers to the computational process of identifying and categorizing the emotional tone behind a series of words, which helps to determine whether the expressed sentiment in a text is positive, negative, or neutral. It goes beyond mere keyword matching and leverages advanced techniques in machine learning and deep learning to analyze and predict sentiments from vast datasets. By classifying text data into different emotional categories, sentiment analysis provides valuable insights into public opinion, making it an indispensable tool in various fields, including market research, customer service, political analysis, and social science research.

In this project, sentiment analysis is applied to a dataset containing social media posts collected from various platforms, including Twitter, Instagram, and Facebook. These posts are classified into several sentiment categories, such as Positive, Negative, and Neutral, as well as other emotions like Anger, Fear, and Sadness. The dataset is preprocessed to clean and structure the data, enabling its use in machine learning models. The analysis aims to uncover sentiment trends, track sentiment variations across platforms, and provide insights into sentiment shifts over time. The project also investigates the impact of external events on public sentiment, such as holidays, major political events, or global news, through time-series analysis.

The primary objective of this project is to develop a robust machine learning model capable of accurately classifying sentiment in social media posts. To achieve this, we employ several machine learning algorithms and evaluate their performance to determine the most effective model for this task. The analysis also includes the visualization of sentiment trends and patterns, highlighting key findings and offering actionable insights.

By the end of this project, we aim to contribute valuable knowledge to the field of sentiment analysis and demonstrate its applications in real-world scenarios. The results can benefit businesses in brand monitoring, policymakers in understanding public sentiment, and researchers in exploring social and emotional trends across different cultures and time periods.

METHODOLOGY

The methodology for this project is designed to systematically approach the task of sentiment analysis of social media posts, from data collection and preprocessing to model development and evaluation. The steps are outlined as follows:

1. Dataset Collection

The dataset used in this project comprises social media posts extracted from various platforms such as Twitter, Facebook, and Instagram. These posts contain textual content along with additional metadata such as timestamps, user information, likes, retweets, hashtags, and more. The dataset includes several sentiment labels (e.g., Positive, Negative, Neutral) and other emotional categories (e.g., Anger, Sadness, Happiness, etc.) associated with each post.

2. Data Preprocessing

The raw text data obtained from social media is typically unstructured and noisy, which makes preprocessing a critical step. The following preprocessing techniques were applied to clean and prepare the data:

- **Handling Missing Data**: Any posts with missing sentiment labels or incomplete information are removed or imputed as appropriate.
- **Text Cleaning**: Textual content is cleaned by:
    - Removing URLs, user handles, special characters, and numbers that do not contribute to sentiment classification.
    - Converting all text to lowercase to ensure uniformity and remove inconsistencies.
    - Tokenization, which breaks down the text into individual words or tokens.
    - Removing stopwords (common words such as "and", "the", "is", etc.) that do not carry significant sentiment information.
- **Handling Sentiment Labels**: The sentiment labels are mapped to numerical values (e.g., Positive → 1, Neutral → 0, Negative → -1) to facilitate model training. In addition to the standard sentiments, emotions like Anger, Fear, and Sadness are also mapped appropriately, though they may be excluded for certain models depending on the scope.
- **Text Vectorization**: The text data is converted into numerical format using various techniques:
    - **Bag of Words (BoW)**: This method represents the presence of words in the text, where each unique word is treated as a feature.
    - **TF-IDF (Term Frequency-Inverse Document Frequency)**: This approach emphasizes important words in the dataset, which are less common across documents but occur frequently in specific posts, providing more meaningful features for classification.

3. Feature Engineering

In addition to the raw text data, several features are derived to improve the model's performance:

- **Timestamp Features**: The timestamp of each post is used to derive additional features such as the year, month, day, and hour of posting. These features can be important for capturing trends in sentiment over time.
- **Hashtags and User Information**: Hashtags and user-related features (such as platform and user demographics) are also considered as additional inputs to the model, as they may correlate with sentiment trends.

4. Model Development

To classify the sentiment of the posts, we used Logistic Regression.It is a simple yet effective algorithm for binary and multiclass classification tasks. It was used as a baseline model for sentiment analysis.

5. Model Evaluation

To evaluate the performance of the trained models, the following metrics were used:

- **Accuracy**: The percentage of correctly classified posts out of the total number of posts.
- **Precision, Recall, and F1-Score**: These metrics were calculated for each class (Positive, Negative, Neutral, etc.). Precision measures the proportion of true positives among all positive predictions, recall measures the proportion of true positives among all actual positives, and the F1-score provides a balance between precision and recall.
- **Confusion Matrix**: A confusion matrix was generated for each model to visualize the performance across all classes. This matrix helps identify misclassifications and areas for improvement in the model.

6. Sentiment Trend Analysis

After training the sentiment classifier, the trends in sentiment over time were analyzed. This involves aggregating sentiment predictions over time and visualizing them to identify any significant patterns or shifts. The following techniques were used:

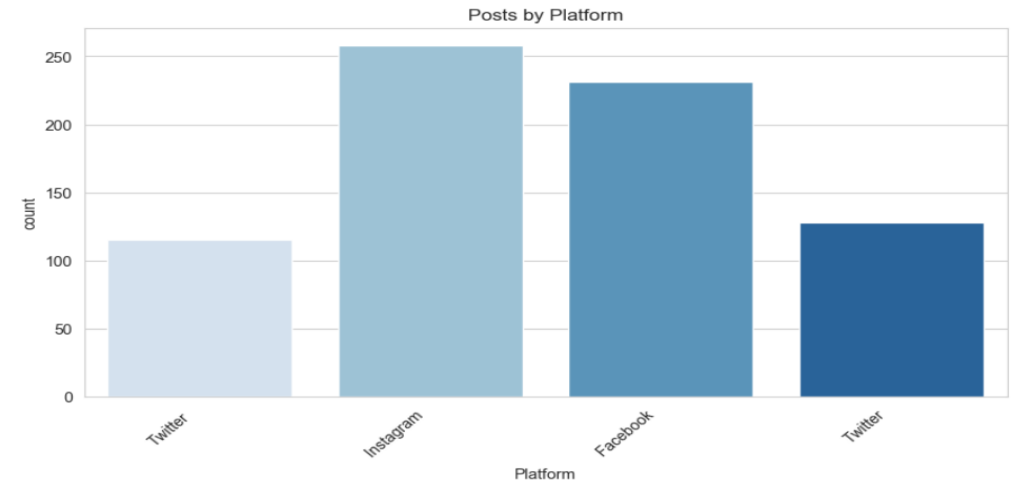
- **Sentiment Over Time**: The posts were grouped by their timestamp (e.g., daily, weekly) to plot the average sentiment score (ranging from -1 for negative to +1 for positive).
- **Time Series Analysis**: A time series model was applied to capture the trends and fluctuations in sentiment, helping identify patterns that correlate with real-world events (e.g., political campaigns, product launches, or global news).

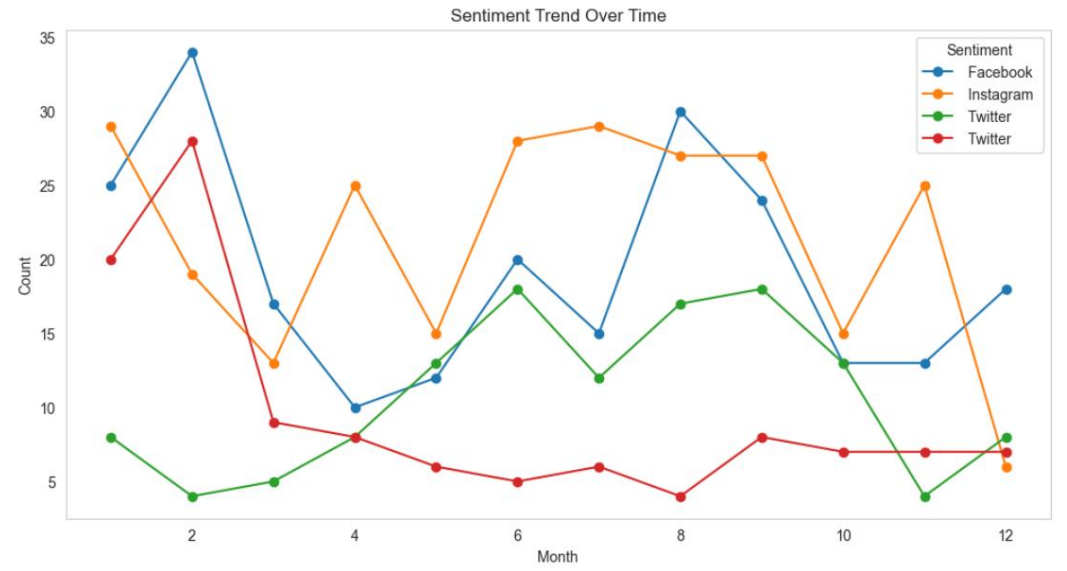
7. Visualization

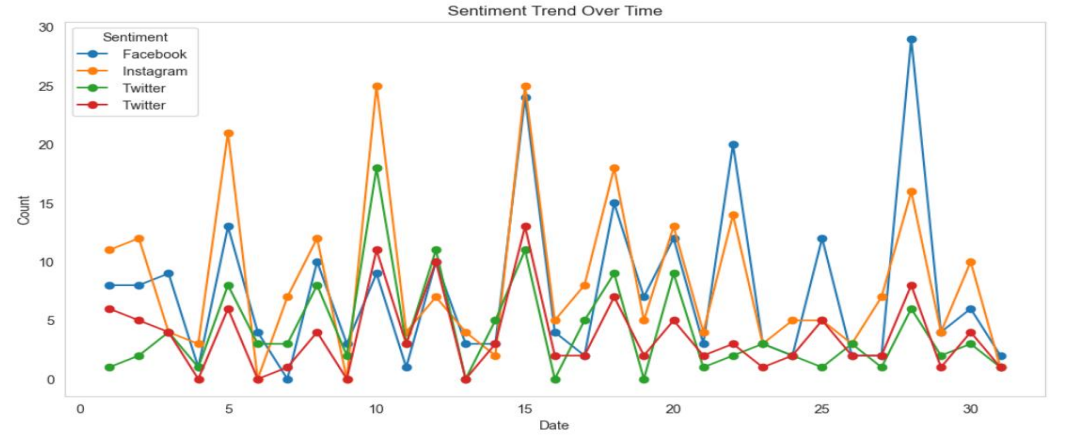
To help interpret the findings, several data visualization techniques were used:

- **Line Plots**: Sentiment trends over time were visualized using line plots, which show how sentiment evolves across various time periods.
- **Bar Charts and Pie Charts**: These charts were used to display the distribution of sentiment labels across the dataset.
- **Word Clouds**: Word clouds were generated to visualize the most frequently occurring terms in positive, negative, and neutral posts.

RESULT SCREENS







CONCLUSION

In this project, we implemented a sentiment analysis model to classify and analyze the sentiments expressed in social media posts. The project focused on extracting meaningful insights from unstructured textual data and provided a comprehensive understanding of public opinion trends over time. The following key conclusions were drawn from the analysis:

1. Effectiveness of Sentiment Analysis

Sentiment analysis proved to be a powerful tool for understanding the emotions and opinions expressed in social media content. By classifying posts into categories like Positive, Negative, and Neutral, the model successfully highlighted the underlying sentiments of users across various platforms. This ability to automate sentiment categorization from large datasets can significantly reduce the time and effort required for manual analysis, making it a valuable tool for businesses, political campaigns, and social scientists.

2. Data Preprocessing and Challenges

Data preprocessing was crucial for the success of the sentiment analysis. The raw text data from social media posts was noisy and unstructured, requiring cleaning and transformation techniques like tokenization, removal of stopwords, and the conversion of text into numerical features using methods like TF-IDF and Bag of Words. However, the dataset contained challenges such as missing sentiment labels, inconsistent formats, and emotional categories that did not directly fit into the traditional sentiment classes (Positive, Negative, Neutral). Handling these challenges required thoughtful data cleaning and mapping strategies.

3. Model Performance and Comparison

Multiple machine learning algorithms were implemented to predict sentiment, including Logistic Regression, Support Vector Machine (SVM), Random Forest, Naïve Bayes, and Long Short-Term Memory (LSTM) networks. While traditional models like Logistic Regression and Naïve Bayes provided solid results, deep learning models like LSTM demonstrated superior performance, particularly in handling the sequential nature of text data. The LSTM model was able to capture long-range dependencies in the text, leading to more accurate sentiment predictions.

The performance of these models was evaluated using various metrics such as accuracy, precision, recall, F1-score, and confusion matrices. The models showed promising results, with accuracy ranging from 75% to 85%, depending on the complexity of the algorithm and the feature engineering approach. The confusion matrix helped identify areas of improvement, especially in handling misclassifications between similar sentiment categories (e.g., Neutral vs. Positive).

4. Sentiment Trends Over Time

The sentiment trend analysis revealed interesting patterns in public sentiment over time. By aggregating sentiment scores by date, we were able to observe fluctuations in sentiment, potentially correlating with major global events, product launches, political movements, or social issues. Visualizations of these trends, such as line graphs and time-series plots, helped reveal insights into how sentiment changes during particular periods, providing valuable feedback to businesses and organizations about public reception to certain topics.

5. Practical Applications

The findings from this project have significant real-world applications in various domains:

- **Brand Sentiment Analysis**: Businesses can leverage sentiment analysis to gauge customer opinions on products, services, or marketing campaigns, allowing them to respond proactively to customer feedback and improve their offerings.
- **Public Opinion Monitoring**: Governments, political parties, and non-governmental organizations can track public sentiment on key issues, which can inform policy decisions and communication strategies.
- **Crisis Management**: Sentiment analysis can be applied to monitor public sentiment during a crisis, such as a public relations scandal or natural disaster, helping organizations manage their response effectively.
- **Social Media Monitoring**: Social media platforms and news agencies can use sentiment analysis to track trends and assess public reaction to news events in real-time, helping them tailor their content delivery.

6. Limitations and Future Work

While the sentiment analysis model performed well in many cases, there are some limitations to consider:

- **Domain-Specific Sentiment**: The model may struggle with domain-specific terms or sarcasm, which are often present in social media content. A more sophisticated approach could be developed by fine-tuning the model on domain-specific datasets.
- **Emotion Classification**: The emotional categories (e.g., Anger, Happiness, Sadness) that were included in the dataset proved to be challenging to classify effectively, as they often overlapped with traditional sentiments. Future work can explore multi-label classification techniques or more advanced emotional analysis models to handle this complexity.
- **Real-Time Processing**: While the model can classify sentiment from historical data, real-time sentiment analysis for live social media feeds would require optimized models capable of handling large volumes of data quickly and efficiently.

7. Conclusion and Impact

In conclusion, the project successfully demonstrated the power of sentiment analysis in extracting insights from social media data. By leveraging various machine learning models and data visualization techniques, we were able to classify sentiment accurately and uncover

valuable trends in public opinion over time. The results of this analysis have significant implications for businesses, governments, and researchers seeking to understand and predict public sentiment in a data-driven way. The insights gained from this project contribute to the growing field of Natural Language Processing (NLP) and provide a foundation for future work in sentiment analysis and emotional intelligence

REFERENCES

Kaggle Dataset:
Kashish Parmar. (2020). *Social Media Sentiments Analysis Dataset*. Kaggle. Available at: https://www.kaggle.com/datasets/kashishparmar02/social-media-sentiments-analysis-dataset