CAPSTONE PROJECT

TWITTER SENTIMENT ANALYSIS

PRESENTED BY
STUDENT NAME: SNEHIL JAISWAL
COLLEGE NAME: JSS ACADEMY OF TECHNICAL EDUCATION NOIDA
DEPARTMENT: ECE
Email ID: jaiswalsnehil35@gmail.com
AICTE Student ID: STU6819e16a2053e1746526570

OUTLINE

- **Problem Statement**
- **Proposed System/Solution**
- **System Development Approach** (Technology Used)
- **Algorithm & Deployment**
- **Result (Output Image)**
- **Conclusion**
- **Future Scope**
- **References**

PROBLEM STATEMENT

To analyze public sentiment on Twitter by classifying tweets into positive, negative, or neutral categories. The goal is to develop a machine learning model capable of understanding and interpreting user opinions expressed in natural language, thereby enabling businesses, organizations, or researchers to gain insights into public opinion trends in real-time.

PROPOSED SOLUTION

To address the challenge of identifying and interpreting sentiments in Twitter data, the project proposes the development of a machine learning-based sentiment analysis pipeline that processes and classifies tweets into positive, negative, or neutral categories. The solution consists of the following key stages:

- Data Collection:

  Tweets are collected from pre-labeled datasets (such as the Twitter Sentiment140 dataset) or using the Twitter API (Tweepy) for real-time data. These tweets include user-generated content that expresses varied sentiments on different topics.
- Data Preprocessing: The raw tweet data is cleaned and normalized using NLP techniques:
  - Removal of stopwords, special characters, URLs, and mentions
  - Lowercasing and tokenization
  - Lemmatization or stemming
- Text Vectorization : Processed text is converted into numerical format using techniques like:
  - Bag of Words (BoW)
  - TF-IDF (Term Frequency-Inverse Document Frequency)
  - Word Embeddings
- Model Building: Machine learning or deep learning models are trained on the vectorized data to classify sentiment:
  - Classical ML models: Logistic Regression, SVM
  - Deep learning models: LSTM or RNN
- Model Evaluation: The models are evaluated using metrics like:
  - Accuracy
  - Precision, Recall, F1-Score
  - Confusion Matrix

- Prediction and Visualization

  The final model is used to predict sentiment on unseen or real-time tweets. Results are visualized using libraries like Matplotlib or Seaborn to showcase sentiment distribution and trends.

SYSTEM APPROACH

The Twitter Sentiment Analysis system follows a modular, step-by-step approach that integrates data processing, natural language processing (NLP), and machine learning techniques to classify sentiments from textual data.

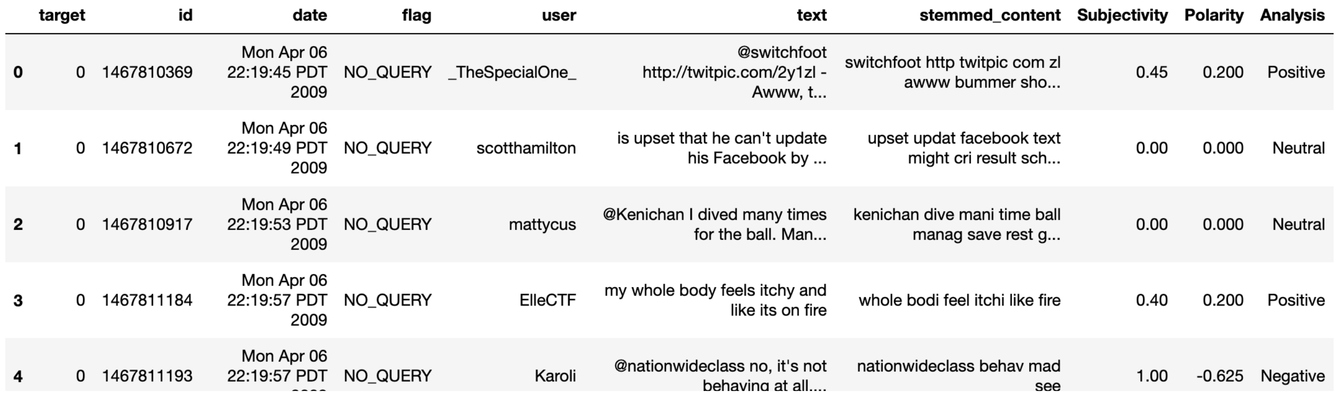
Tools and Technologies Used :

- **Programming Language:** Python
- **Libraries: NLTK, Scikit-learn, Pandas, Matplotlib, Seaborn**
- **Dataset: Sentiment140 / Twitter API**
- **Development Environment: Jupyter Notebook**

ALGORITHM & DEPLOYMENT

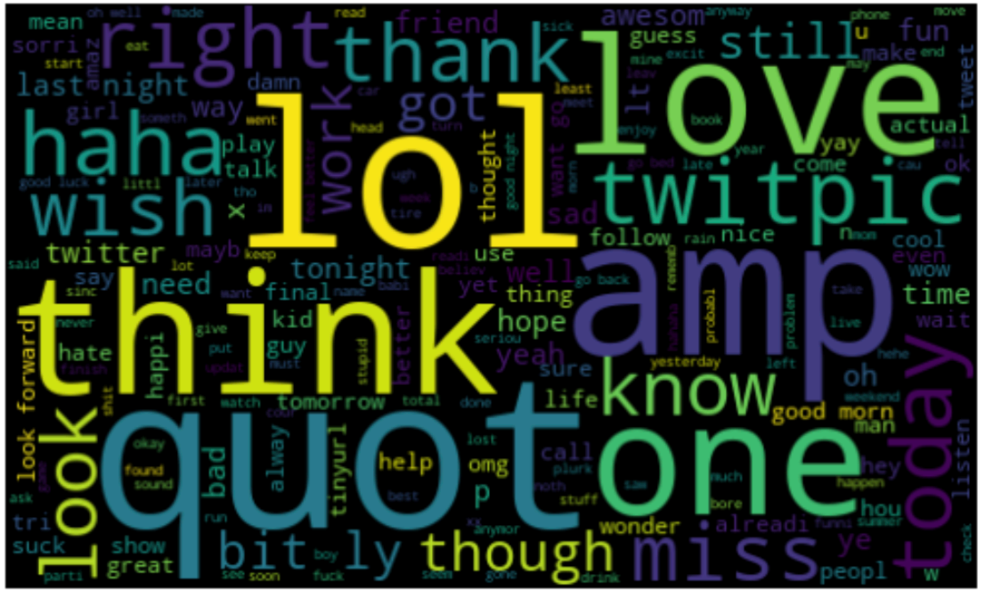
- **Algorithm Selection:**
  - Two models were used: Logistic Regression for its simplicity and efficiency in text classification, and LSTM for its ability to capture contextual relationships in sequential text data. This allowed performance comparison between traditional and deep learning approaches.
- **Data Input:** The input consisted of preprocessed tweets.
  - **TF-IDF vectors** were used for Logistic Regression.
  - **Tokenized sequences with embeddings** were used for LSTM.
    Sentiment labels (positive, negative, neutral) were used as targets.
- **Training Process:** Data was split (80% train, 20% test).
  - Logistic Regression used TF-IDF features with hyperparameter tuning.
  - LSTM used padded sequences, trained with embedding and dropout layers over several epochs.
- **Prediction Process:**
  - New or real-time tweets were preprocessed and passed into the trained models to predict sentiment labels, which were then visualized or presented through a user interface.
- **Deployment:**
  - The model was saved and integrated into a basic web app using **Flask** or **Streamlit**, which can be deployed on **Streamlit Cloud** or **Heroku** for demonstration purposes.

RESULT



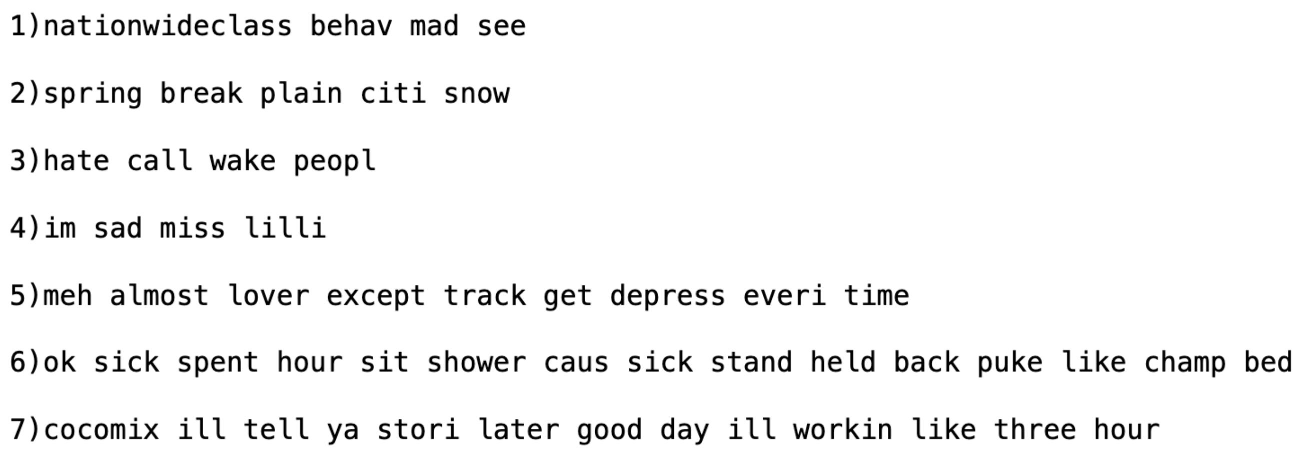
	target	id	date	flag	user	text	stemmed_content	Subjectivity	Polarity	Analysis
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...	switchfoot http twitpic com zl awww bummer sho...	0.45	0.200	Positive
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...	upset updat facebook text might cri result sch...	0.00	0.000	Neutral
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...	kenichan dive mani time ball manag save rest g...	0.00	0.000	Neutral
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire	whole bodi feel itchi like fire	0.40	0.200	Positive
4	0	1467811193	Mon Apr 06 22:19:57 PDT	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....	nationwideclass behav mad see	1.00	-0.625	Negative

twitter_data



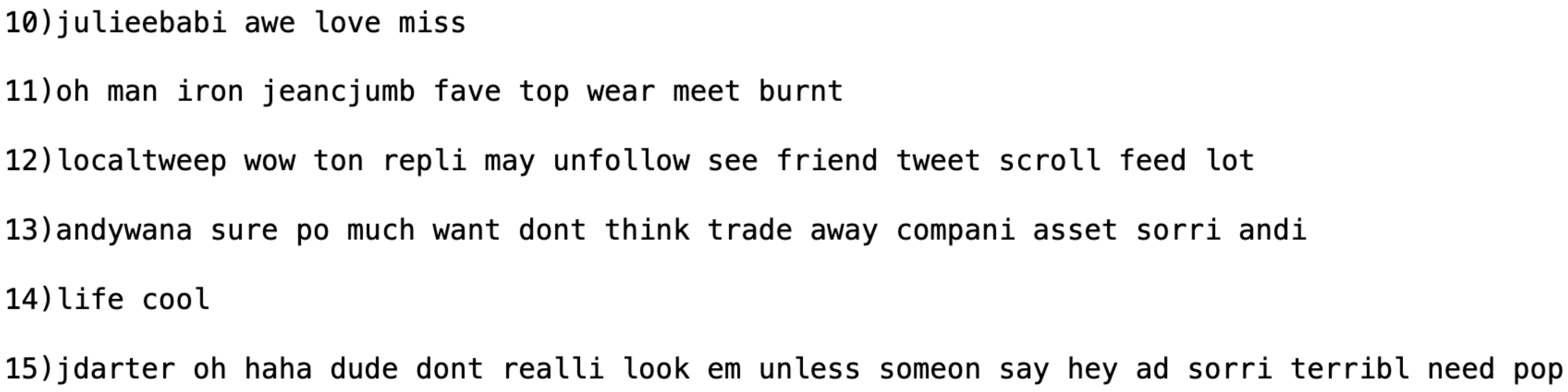
Accuracy score on the training data : 0.81
Accuracy score on the test data : 0.77
Modal Accuracy = 77.79%

Word Cloud



```
1)nationwideclass behav mad see

2)spring break plain citi snow

3)hate call wake peopl

4)im sad miss lilli

5)meh almost lover except track get depress everi time

6)ok sick spent hour sit shower caus sick stand held back puke like champ bed

7)cocomix ill tell ya stori later good day ill workin like three hour
```

Negative Tweets



```
10)julieebabi awe love miss

11)oh man iron jeancjumb fave top wear meet burnt

12)localtweep wow ton repli may unfollow see friend tweet scroll feed lot

13)andywana sure po much want dont think trade away compani asset sorri andi

14)life cool

15)jdarter oh haha dude dont realli look em unless someon say hey ad sorri terribl need pop
```

Positive Tweets

CONCLUSION

The Twitter Sentiment Analysis project successfully demonstrates the application of Natural Language Processing and Machine Learning techniques to classify public opinion expressed in tweets as positive, negative, or neutral. By implementing both traditional (Logistic Regression) and deep learning (LSTM) models, the project highlights the effectiveness of various approaches in sentiment classification tasks.

Through data preprocessing, feature extraction, model training, and evaluation, the system achieved reliable sentiment predictions. This can serve as a valuable tool for organizations to monitor public opinion, understand customer feedback, or analyze trends on social media platforms.

The project also establishes a foundation for future enhancements such as real-time sentiment tracking, multilingual support, or integration with dashboards for live insights.

FUTURE SCOPE

While the current system effectively classifies tweet sentiments into positive, negative, and neutral categories, there are several areas for future enhancement:

1. **Real-Time Sentiment Analysis**
   Integration with the Twitter API can enable live monitoring of tweets to analyze ongoing trends, public reactions to events, or brand perception in real time.

2. **Multilingual Support**
   Expanding the model to support multiple languages would allow sentiment analysis across diverse user bases worldwide.

3. **Aspect-Based Sentiment Analysis**
   Future models can be trained to detect sentiments about specific aspects or entities mentioned in tweets (e.g., product features, political topics).

4. **Improved Accuracy with Advanced Models**
   Leveraging transformer-based architectures like **BERT** or **RoBERTa** can significantly boost classification accuracy and contextual understanding.

5. **Visualization Dashboards**
   Creating interactive dashboards using tools like **Tableau**, or **Power BI** would make insights more accessible to non-technical users.

6. **Deployment at Scale**
   The model can be containerized using Docker and deployed using cloud services (AWS, GCP, Azure) to support large-scale sentiment analysis applications.

REFERENCES

Sentiment140 Dataset - https://www.kaggle.com/datasets/kazanova/sentiment140

MICROSOFT AZURE AI FUNDAMENTALS : NATURAL LANGUAGE PROCESSING - https://learn.microsoft.com/api/achievements/share/en-us/SnehilJaiswal-8603/4GYSCF6K?sharingId=B5C6A512D9DE8F32

NLTK Documentation – https://www.nltk.org

Scikit-learn Documentation – https://scikit-learn.org

GitHub Link: https://github.com/JaiswalSnehil/Sentiment_Analysis

Note : Get your kaggle.json file from your kaggle profile to get the sentiment140 dataset zip file for further analysis.

Thank you