INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

SENTIMENT ANALYSIS

Shivam
Batch 246

INNOMATICS
RESEARCH LABS

About me

Proactive Data Science trainee passionate about leveraging data to derive insights and solve complex problems. Proficient in Python with foundational knowledge in machine learning, deep learning, object detection, and data analysis techniques. Eager to apply skills to real-world projects and support data-driven decision-making. Skilled in using GitHub for version control and collaboration, demonstrating adherence to best practices. Competent in Microsoft Office Suite, facilitating effective communication and data presentation.

INNOMATICS RESEARCH LABS

Agenda

Business Problem

- **Objective:** To analyze and understand public sentiment towards various topics on Twitter by classifying tweets into positive, neutral, and negative sentiments.
- **Use Case:** Sentiment analysis can help businesses monitor customer opinions, track brand reputation, and make informed decisions based on public feedback.

Business Problem and Use Case Domain Understanding Domain

Social Media Sentiment Analysis

INNOMATICS RESEARCH LABS

Objective of the Project

The primary objective is to build a robust machine learning model to classify tweets into positive, neutral, and negative sentiments. This involves preprocessing the data, building and tuning an LSTM model, and deriving insights from the analysis to help stakeholders understand public sentiment trends.

Summary of the Data

Dataset

•Columns: target, ids, date, flag, user, text

•Target: Sentiment labels (0 and 4) representing negative and positive sentiments.
•Text: Actual tweet content.
•Date: Timestamp of the tweet.
•User: User who posted the tweet.
•Flag: Non-informative column (constant value).

Exploratory Data Analysis

	target	ids	flag	user	text	time	date_only	year
0	negative	1467810369	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...	22:19:45	2009-04-06	2009
1	negative	1467810672	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...	22:19:49	2009-04-06	2009
2	negative	1467810917	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...	22:19:53	2009-04-06	2009
3	negative	1467811184	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire	22:19:57	2009-04-06	2009
4	negative	1467811193	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....	22:19:57	2009-04-06	2009

Data Cleaning Steps

1.Column Renaming:

•Renamed columns for better readability.,Example: date to date_only, text to tweet_text.

2.Target Mapping:

•Mapped target values: 0 to negative, 4 to positive.

3.Date Parsing:

•Parsed date column to extract date_only, time, and year.

4.Dropped Irrelevant Columns:

•Dropped the flag column as it contained a single constant value.

INNOMATICS RESEARCH LABS

Exploratory Data Analysis

Summary statistics:

	target	ids	date	flag	\
count	1.600000e+06	1.600000e+06	1600000	1600000	
unique	NaN	NaN	774363	1	
top	NaN	NaN	Mon Jun 15 12:53:14 PDT 2009	NO_QUERY	
freq	NaN	NaN	20	1600000	
mean	2.000000e+00	1.998818e+09	NaN	NaN	
std	2.000001e+00	1.935761e+08	NaN	NaN	
min	0.000000e+00	1.467810e+09	NaN	NaN	
25%	0.000000e+00	1.956916e+09	NaN	NaN	
50%	2.000000e+00	2.002102e+09	NaN	NaN	
75%	4.000000e+00	2.177059e+09	NaN	NaN	
max	4.000000e+00	2.329206e+09	NaN	NaN	

	user	text
count	1600000	1600000
unique	659775	1581466
top	lost_dog	isPlayer Has Died! Sorry
freq	549	210
mean	NaN	NaN
std	NaN	NaN
min	NaN	NaN
25%	NaN	NaN
50%	NaN	NaN
75%	NaN	NaN
max	NaN	NaN

Data Manipulation Steps

1.Tokenization:

- Converted tweet text into sequences of tokens using Keras Tokenizer.

2.Padding:

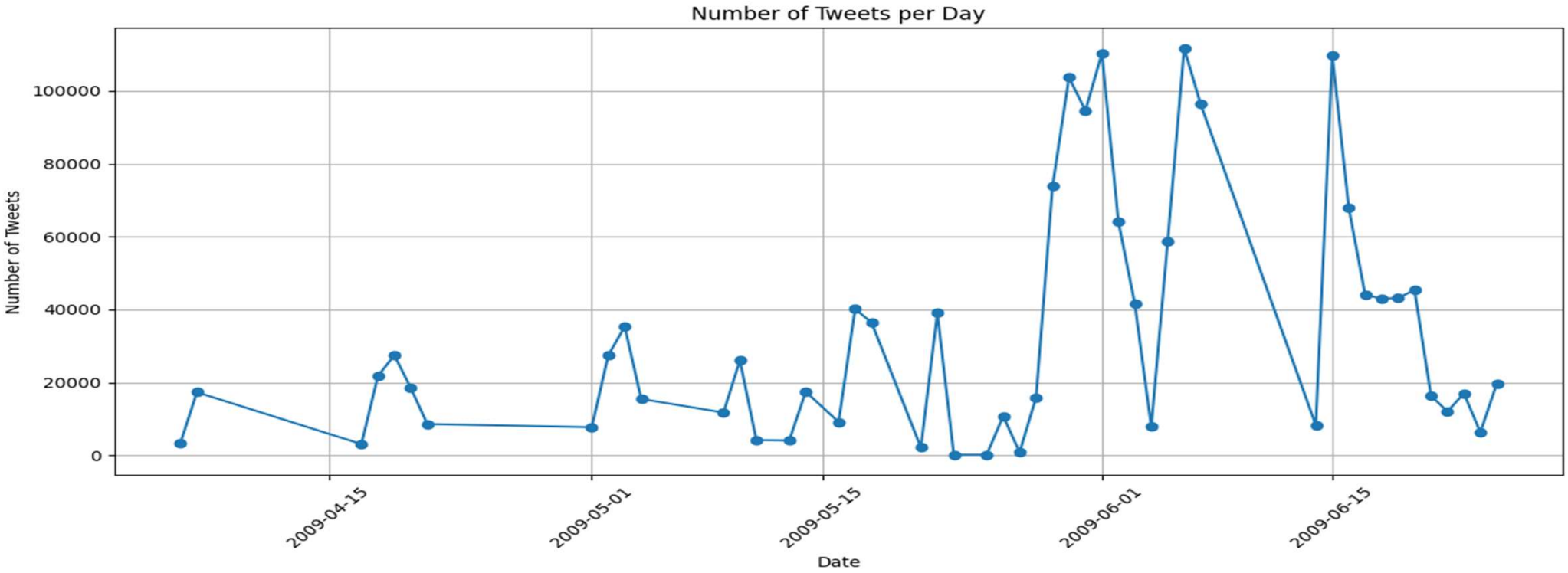
- Padded sequences to a fixed length (300) for consistent input to the model.

3.Label Encoding:

- Encoded target labels into numerical values for model training.

INNOMATICS RESEARCH LABS

Exploratory Data Analysis



Univariate Analysis Steps

1.Class Distribution:

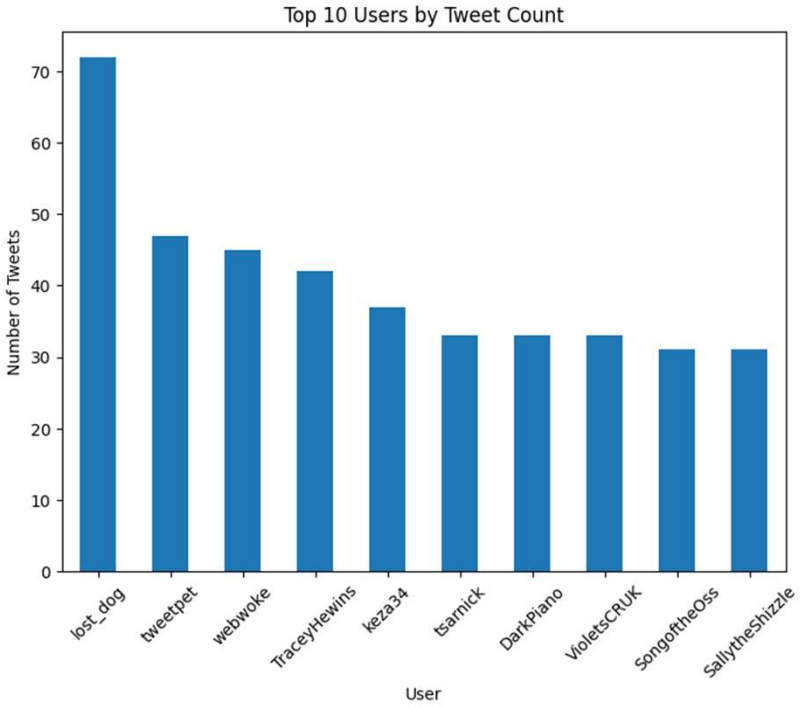
•Plotted the distribution of sentiment classes to check for balance.

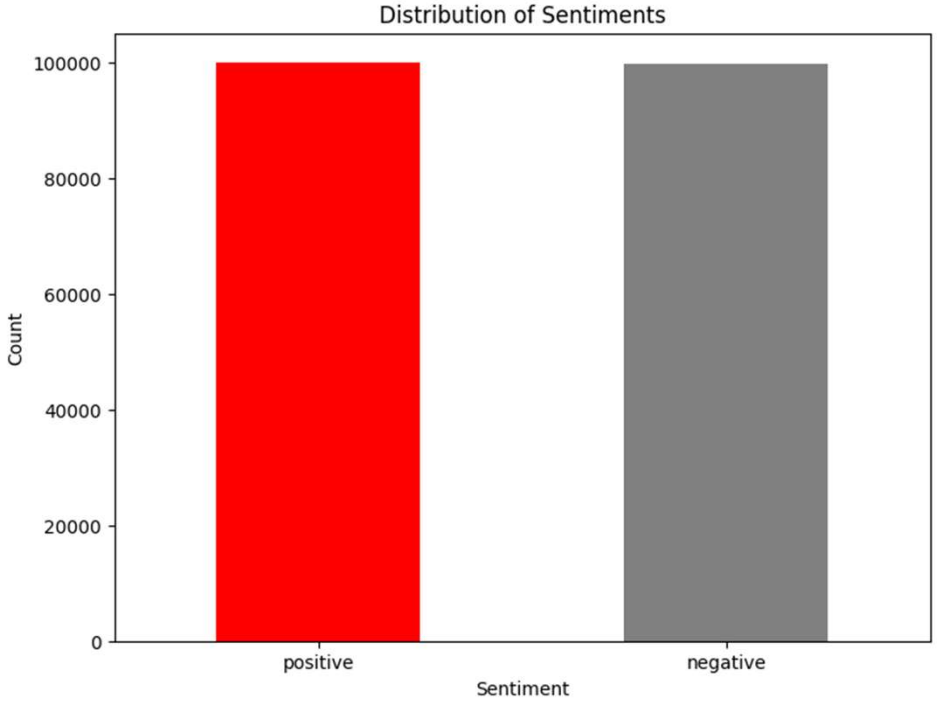
2.Tweet Frequency by Date:

•Analyzed the number of tweets per day to identify peak activity periods.

INNOMATICS RESEARCH LABS

Exploratory Data Analysis





Bivariate Analysis Steps

1.Sentiment by Date:

- Examined sentiment trends over time to understand how sentiment varies by date.

2.Sentiment by User:

- Analyzed the distribution of sentiments by user to identify any significant patterns.

INNOMATICS RESEARCH LABS

Key Business Question

How does public sentiment towards various topics on Twitter fluctuate over time, and what are the primary factors influencing these trends?

Conclusion (Key Findings Overall)

1. **Data Quality**:
   Cleaned and preprocessed data, including handling date-time formats and dropping irrelevant columns.
2. **Sentiment Distribution**:
   Identified potential class imbalance, with more tweets labeled as negative than positive.
3. **Peak Activity**:
   Found certain days with higher tweet activity, indicating potential events or news driving public engagement.
4. **Model Performance**:
   Built and tuned an LSTM model, achieving good accuracy in classifying tweet sentiments.
5. **Future Improvements**:
   Addressing class imbalance, exploring advanced embeddings, and using cross-validation could further improve model performance.

INNOMATICS RESEARCH LABS

Q&A Slide

1. **What is the purpose of this sentiment analysis project?**
   To classify tweets into sentiments to understand public opinion trends.
2. **How was the data collected?**
   For this project, a pre-collected dataset was used. In practice, tools like Twitter API can be used for data collection.
3. **What preprocessing steps were taken?**
   Data cleaning, tokenization, padding, and label encoding.
4. **What model was used for sentiment analysis?**
   An LSTM model with hyperparameter tuning using Keras Tuner.
5. **What were the key findings?**
   Discovered class imbalances and peak tweet days, built a performant LSTM model.

INNOMATICS RESEARCH LABS

Challenges:

•Handling imbalanced data which can affect model performance.

•Parsing complex date-time formats and extracting meaningful features.

•Ensuring consistent and accurate preprocessing steps to avoid errors in subsequent analysis

Experience:

•Gained hands-on experience in data preprocessing, exploratory data analysis, and building machine learning models.

•Learned to use tools like Keras Tuner for hyperparameter optimization.

INNOMATICS RESEARCH LABS

Final Conclusion

After extensive data preprocessing, exploratory analysis, text processing, and model building with hyperparameter tuning, the LSTM model achieved satisfactory performance in classifying tweets into positive and negative sentiments

INNOMATICS RESEARCH LABS

Recommendations for Future Work:

- **Addressing Class Imbalance**: Implement techniques like class weighting or resampling to handle class imbalance.
- **Advanced Embeddings**: Experiment with more advanced embeddings (e.g., BERT, GloVe) to capture richer text semantics.
- **Model Complexity**: Explore more complex architectures or hybrid models to further improve performance.
- **Cross-Validation**: Use cross-validation to ensure robustness and generalizability of the model.

INNOMATICS RESEARCH LABS

THANK YOU

INNOMATICS
RESEARCH LABS