SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE

A
MINI PROJECT REPORT
ON
“SENTIMENT ANALYSIS OF TWEETS”

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE
IN THE FULFILLMENT OF THE REQUIREMENT
OF

Data Science and Big Data Analytics Laboratory
Third Year Computer Engineering
Academic Year 2023-24

BY

Name of Students:	Roll No.:
Sanskruti Kabadi	**3101065**

Under the Guidance of
Mr. S. D. Dighe

DEPARTMENT OF COMPUTER ENGINEERING
STES’S SINHGAD INSTITUTE OF TECHNOLOGY AND SCIENCE
NARHE, PUNE – 411041

DEPARTMENT OF COMPUTER ENGINEERING
STES'S SINHGAD INSTITUTE OF TECHNOLOGY AND SCIENCE
NARHE, PUNE – 411041

CERTIFICATE

This is to certify that,

Name of Students:	Roll No.:
Sanskruti kabadi	**3101065**

studying in TE Computer Engineering Course SEM-VI has successfully completed their DSBDA Lab Mini-Project work titled **SENTIMENT ANALYSIS OF TWEETS** at Sinhgad Institute of Technology and Science, Narhe in the fulfillment of the Bachelor's Degree in Computer Engineering in **Savitribai Phule Pune University**, during the academic year 2023-2024.

Mr. S. D. Dighe	**Dr. G. S. Navale**	**Dr. S. D. Markande**
Guide	Head of Department	Principal

SINHGAD INSTITUTE OF TECHNOLOGY AND SCIENCE, NARHE, PUNE

Place : Pune

Date :

ACKNOWLEDGEMENT

I take this opportunity to acknowledge each and every one who contributed towards our work. We express our sincere gratitude towards guide **Mr. S. D. Dighe**, Assistant Professor at Sinhgad Institute of Technology and Science, Narhe,Pune for her valuable inputs, guidance and support throughout the course.

I wish to express our thanks to **Dr. G. S. Navale**, Head of Computer Engineering Department, Sinhgad Institute of Technology and Science, Narhe for giving us all the help and important suggestions all over the Work.

I thank all the teaching staff members, for their indispensable support and priceless suggestions. We also thank our friends and family for their help in collecting data, without their help DSBDA Lab Mini Project report have not been completed. At the end our special thanks to **Dr. S. D. Markande**, Principal Sinhgad Institute of Technology and Science, Narhe for providing ambience in the college, which motivate us to work.

Name of students	Signature
Sanskruti kabadi	

CONTENT

INTRODUCTION

In today's digital age, social media platforms like Twitter have become invaluable sources of data for understanding public opinion and sentiment on various topics. Leveraging this vast repository of textual data through sentiment analysis not only provides insights into public perception but also offers opportunities for exploring trends and attitudes in specific domains. This mini-project focuses on sentiment analysis applied to Twitter data, specifically targeting tweets related to data science, a field pivotal in shaping modern technological advancements. The dataset chosen for this endeavor, sourced from Kaggle, offers a comprehensive collection of tweets specifically centered around data science topics. With over 10,000 tweets, this dataset presents a diverse array of opinions, discussions, and sentiments expressed by users within the data science community. Analyzing these tweets can offer valuable insights into the prevailing sentiments, challenges, and advancements within the data science domain.

In addition to sentiment analysis, this mini-project integrates data visualization and analysis techniques to provide a comprehensive exploration of the dataset. Data visualization serves as a powerful tool for uncovering patterns, trends, and outliers within the textual data. By visually representing the sentiment distribution, word frequency, and other relevant metrics, we cangain a deeper understanding of the underlying patterns and dynamics present in the dataset. Data analysis plays a crucial role in preprocessing the textual data, extracting meaningful features, and preparing it for sentiment classification. Techniques such as text cleaning, tokenization, and feature engineering are essential steps in this process, enabling the creation of effective models for sentiment classification. Furthermore, exploratory data analysis (EDA) techniques allow us to gain insights into the characteristics of the dataset, such as the distribution of sentiment labels, the most common words/phrases, and potential correlations between variables.

PROBLEM STATEMENT

Use the following dataset and classify tweets into positive and negative tweets. https://www.kaggle.com/ruchi798/data-science-tweets

The problem at hand is to perform sentiment analysis on Twitter data. Given a dataset of tweets, the goal is to classify each tweet as either positive or negative based on the sentiment expressed in the text. Sentiment analysis has various applications, such as understanding public opinion, monitoring brand sentiment, and analyzing customer feedback.

The challenge in this task is to develop a machine learning model that can accurately classify tweets into positive or negative sentiment categories. This requires understanding and processing natural language data, extracting relevant features, and training a model that can generalize well to unseen data. The model should be able to capture the nuances and contextof tweets, considering factors such as sarcasm, irony, and abbreviations commonly used in social media.

REQUIREMENTS

Hardware Requirements:

- Processor: Intel Core i5 or equivalent (later generations recommended)
- RAM: 4GB (8GB or more recommended for better performance)

Software Specifications:

- Operating System: Windows 10 (or macOS, Linux)
- Languages: Python (3.6.3 or later)
- Software: Anaconda, Jupyter Notebook
- Dataset: data_visualization.csv
- Libraries: nltk, SentimentIntensityAnalyzer, vader_lexicon

ALGORITHM USED

Sentiment Intensity Analyzer Algorithm:

1. Install NLTK and Download Resources: First, ensure you have the Natural Language Toolkit (NLTK) installed.

2. Next, download the VADER lexicon (required for sentiment analysis).

3. Sentiment Analysis with VADER: VADER is a pre-trained sentiment analysis tool that provides polarity scores for text. It’s particularly useful for social media content and doesn’t require extensive text preprocessing.

The resulting DataFrame will contain sentiment scores for each review, including positive, negative, neutral, and compound scores. The compound score summarizes overall sentiment.

IMPLEMENTATION

Data Science, Data Analytics and Data Visualization:

1. **Data Science:**

- **Focus:** Extracting knowledge and insights from data.
- **Skills:** Statistics, machine learning, programming.
- **Description:** Data science encompasses the entire life cycle of data, from collection and cleaning to analysis and modeling.

2. **Data Analytics:**

- **Focus:** Analyzing data to identify trends and patterns.
- **Skills:** Statistical methods and tools.
- **Description:** Data analytics is a subfield of data science that specifically dives into analyzing data. Analysts use statistical methods and tools to clean, transform, and uncover patterns within datasets.

3. **Data Visualization:**

- **Focus:** Communicating data insights through visual elements.
- **Skills:** Design, communication.
- **Description:** Data visualization is the art of presenting data findings through visuals like charts, graphs, and maps. It helps people understand complex information quickly and effectively.

Pandas, NLKT and Plotly:

1. **pandas (pd)**: This library is imported as pd for convenience. Pandas is a powerful tool for data analysis, especially for working with tabular data. It provides data structures like DataFrames (similar to spreadsheets) and Series (one-dimensional arrays) that allow for efficient manipulation, cleaning, and analysis of data.

2. **nltk**: The Natural Language Toolkit (nltk) is a library specifically designed for working with human language data. It provides a comprehensive set of functionalities for various tasks in Natural Language Processing (NLP), including tokenization (breaking text into smaller units like words), stemming/lemmatization (reducing words to their base form), text classification, sentiment analysis, and more.

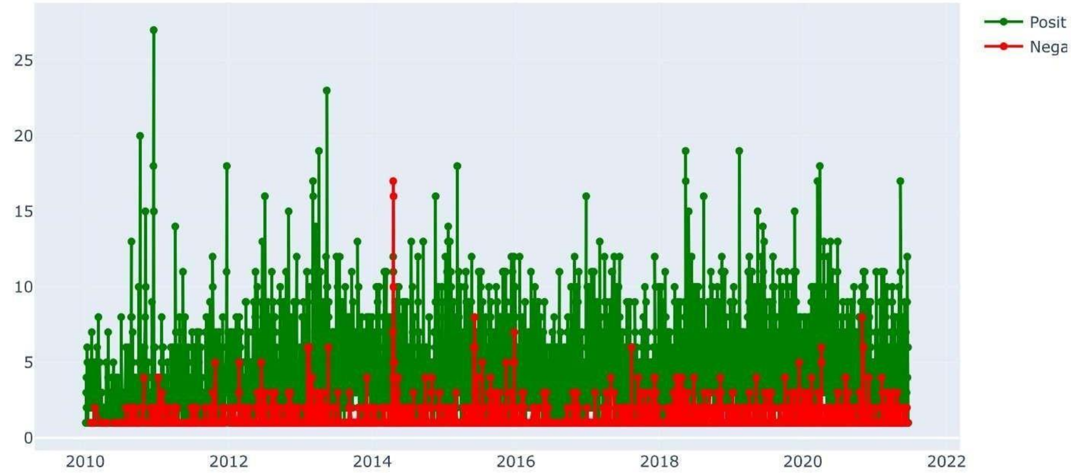
3. **plotly**: This library offers functionalities for creating interactive and visually appealing data visualizations. Plotly allows you to generate various chart types like bar charts, scatter plots, line charts, and even geographical visualizations. It's particularly known for its ability to create web-based interactive charts that can be explored dynamically.

In essence, this code snippet imports essential tools for a data science workflow:

- **pandas** helps you wrangle and prepare your data.
- **nltk** assists in analyzing textual data.
- **plotly** enables you to create insightful visualizations to present your findings.

OUTPUT:

	tweet	date	id	sentiment	sentiment_category
0	Take your storytelling to using...	2021-	1406335989484822531	0.7089	positive
1	Choosing Fonts for Your Data Visualization | b...	2021-06-19	1406292636789526537	0.0000	neutral
2	This data visualization our greate...	2021-	1406082288035811330	0.0000	neutral
3	Looking for examples of stellar charts made so...	2021-06-18	1405948260796100610	0.4019	positive
4	With #WISQARS Data Visualization, you can disp...	2021-06-18	1405942146960613376	-0.4215	negative



CONCLUSION

The provided code exemplifies the application of data analysis and machine learning techniques to analyze sentiments in Twitter data. The code performs a range of tasks including data exploration, preprocessing, feature extraction, model training, and evaluation.

Throughout the code, insights into the dataset are gained by visualizing the distribution of tweets, identifying frequently occurring words, and exploring hashtags associated with different sentiment categories. This analysis aids in understanding the characteristics of thedata and lays the groundwork for subsequent modeling efforts. By harnessing the capabilitiesof libraries such as NumPy, Pandas, Matplotlib, Seaborn, NLTK, Gensim, Scikit-learn, WordCloud, and Tqdm, the code streamlines the data analysis process and facilitates efficient model development.