A REPORT ON

Data Science and Big Data Analytics Mini Project

on

TWITTER SENTIMENT ANALYSIS

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE IN THE PARTIAL FULFILLMENT FOR THE AWARD OF

THE DEGREE OF BACHELOR OF ENGINEERING IN COMPUTER ENGINEERING

BY

Aarti Swami 3202065

Under The Guidance

Mr. S.D. Sachin Dighe

DEPARTMENT OF COMPUTER ENGINEERING

Sinhgad Institute of Technology & Science Narhe, Pune – 411041

Sinhgad Institutes

Department of Computer Engineering,

Sinhgad Institute of Technology and Science, Narhe

CERTIFICATE

This is to certify that the report entitled

“TWITTER SENTIMENT ANALYSIS”

Submitted by

Aarti Swami 3202065

is a bonafide work carried out by them under the supervision of **Mr. S.D. Dighe** and is approved for the partial fulfillment of the requirement of **Data Science and Big Data Analytics Laboratory** course in Third Year Computer Engineering, in the academic year 2022-2023 prescribed by Savitribai Phule Pune University, Pune.

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

Place: Pune
Date :22/05/2023

Acknowledgement

I take this opportunity to acknowledge each and every one who contributed towards my work. I express my 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 my thanks to Dr. G. S. Navale, Head of Computer Engineering Department, Sinhgad Institute of Technology and Science, Narhe for giving me all the help and important suggestions all over the Work.

I thank all the teaching staff members, for their indispensable support and priceless suggestions.

I also thank my friends and family for their help in collecting data without which their help Database Management Laboratory report have not been completed. At the end my 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.

Aarti Swami

Contents

1. INTRODUCTION

In the age of social media, Twitter has emerged as a popular platform for expressing opinions, sharing thoughts, and engaging in public discourse. With millions of tweets being generated every day, there is a vast amount of valuable information that can be extracted from this data. One such valuable insight is understanding the sentiment or emotional tone behind these tweets. Sentiment analysis, also known as opinion mining, is the process of automatically determining the sentiment expressed in a piece of text, whether it is positive, negative, or neutral.

The ability to accurately analyse sentiment on Twitter has numerous applications in various domains. It can provide insights into public opinion, customer feedback, brand perception, and even predict stock market trends. Businesses can leverage sentiment analysis to gauge customer satisfaction, identify emerging trends, and make data-driven decisions. Government agencies can monitor public sentiment on social issues and policy decisions. Market researchers can assess the popularity and sentiment surrounding products or services.

The purpose of this report is to present an analysis of sentiment on Twitter data using machine learning techniques. The objective is to develop a model that can accurately classify tweets as either positive or negative based on their content. By analysing sentiments expressed in tweets, we can gain valuable insights into the prevailing public opinion and sentiment on various topics. The scope of this project includes several stages of analysis. We begin by pre-processing the Twitter data, which involves cleaning and transforming the text data to make it suitable for analysis. We then perform exploratory data analysis to gain a deeper understanding of the dataset, such as the distribution of sentiments and the most frequently used words. Additionally, we use visualization techniques like word clouds to visualize the most common words associated with positive and negative sentiments.

To enable machine learning-based sentiment classification, we employ word embedding techniques to convert the textual data into numerical vectors. Specifically, we utilize the Word2Vec model, which maps words to a high-dimensional vector space, capturing semantic relationships between words. This allows the machine learning model to understand the contextual meaning of words and make better predictions.

Next, we train and evaluate various classifiers on the labelled Twitter data. We use machine learning algorithms such as Random Forest, Logistic Regression, Decision Tree, Support Vector Machines (SVM), and XGBoost. We measure the performance of these models using evaluation metrics such as accuracy, F1 score, and confusion matrix. This allows us to assess the effectiveness of different classifiers in accurately predicting sentiment.

In summary, this report aims to analyse sentiment on Twitter data using machine learning techniques. By understanding the sentiment behind tweets, we can gain insights into public opinion, customer feedback, and brand perception. The findings of this analysis can be applied in various domains such as market research, customer service, and social listening, empowering businesses and organizations to make informed decisions based on the insights derived from sentiment analysis on Twitter data.

2. PROBLEM STATEMENT

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 analysing 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 context of tweets, considering factors such as sarcasm, irony, and abbreviations commonly used in social media.

The solution to this problem involves several steps, including data preprocessing, feature extraction, model training, and evaluation. It requires the use of machine learning algorithms and techniques, as well as appropriate libraries and tools for natural language processing and text analysis.

The ultimate objective is to build a robust and accurate sentiment analysis system that can effectively analyse large volumes of tweets and provide valuable insights into public sentiment and opinions on various topics.

3. Dataset

The dataset is divided into two parts: a training set (train_tweet.csv) and a test set (test_tweets.csv).

The train_tweet.csv file contains labelled tweets, where each tweet is associated with a sentiment label of either 0 or 1. The label 0 represents a tweet with a non-negative sentiment (neutral), while the label 1 represents a tweet with a negative sentiment. This dataset is used for training and evaluating the sentiment analysis model.

The test_tweets.csv file contains unlabelled tweets, and it is used to test the trained model's performance on unseen data. The objective is to predict the sentiment labels for these tweets using the trained model.

These datasets are typically collected from real-world Twitter data, and they provide a representative sample of tweets from various users and topics. The dataset used in the code allows for the development and evaluation of a sentiment analysis model specifically tailored to Twitter data.

4. Python Libraries Used

The libraries used in the code include:

1. Numpy: A library for numerical computing in Python.
2. Pandas: A library for data manipulation and analysis, used for handling structured data and dataframes.
3. Matplotlib: A plotting library for creating visualizations in Python.
4. Seaborn: A data visualization library built on top of Matplotlib, providing additional highlevel visualization capabilities.
5. Warnings: A module for issuing and handling warnings in Python.
6. NLTK (Natural Language Toolkit): A library for natural language processing tasks, including tokenization and stemming.
7. Gensim: A library for topic modeling, document similarity, and word embedding techniques.
8. Sklearn (Scikit-learn): A machine learning library in Python, providing a wide range of tools for classification, regression, clustering, and more.
9. Wordcloud: A library for creating word clouds, visual representations of word frequency in a text.
10. Tqdm: A library for creating progress bars in Python.

These libraries are commonly used in data science and machine learning tasks and provide various functionalities for data preprocessing, visualization, and model training and evaluation.

5. CONCLUSION

In conclusion, the provided code demonstrates the application of data analysis and machine learning techniques to analyze sentiment in Twitter data. The code performs various tasks such as data exploration, data preprocessing, feature extraction, model training, and evaluation.

Throughout the code, we gained insights into the dataset by visualizing the distribution of tweets, identifying frequent words, and exploring hashtags associated with different sentiment categories. This analysis helps in understanding the characteristics of the data and provides a foundation for further modeling.

The code also employs machine learning algorithms, such as Random Forest, Logistic Regression, Decision Tree, Support Vector Machine (SVM), and XGBoost, to train classification models for sentiment analysis. These models are evaluated using accuracy and F1 score metrics, and their performance is assessed through confusion matrices.

By leveraging the power of libraries like Numpy, Pandas, Matplotlib, Seaborn, NLTK, Gensim, Sklearn, Wordcloud, and Tqdm, the code streamlines the data analysis process and enables efficient model development.

In summary, the code showcases the workflow of sentiment analysis on Twitter data and demonstrates the potential of machine learning in extracting insights from social media text. This code can serve as a starting point for further exploration, experimentation, and refinement of sentiment analysis tasks in the context of social media data.