CCT College Dublin

|  |  |
| --- | --- |
| Module Title: | Advanced Data Analysis |
| Assessment Title: | Decoding Twitter Emotions: Uncovering Sentiment Patterns and Extracting Insights |
| Lecturer Name: | David McQuaid |
| Student Full Name: | Haider Riaz Ghuman |
| Student Number: | 2022564 |
| Assessment Due Date: | 07/01/2024 |
| Date of Submission: | 07/01/2024 |

Declaration

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Table of Contents

[Abstract 3](#_Toc154792562)

[Introduction 3](#_Toc154792563)

[Hypothesis 3](#_Toc154792564)

[Problem Statement 4](#_Toc154792565)

[Objective: 4](#_Toc154792566)

[Literature Review 5](#_Toc154792567)

[Introduction: 5](#_Toc154792568)

[Social Media: 5](#_Toc154792569)

[Social Media data analysis: 5](#_Toc154792570)

[Methodology 6](#_Toc154792571)

[Data Collection: 6](#_Toc154792572)

[Pre-processing: 6](#_Toc154792573)

[Sentiment Classification: 6](#_Toc154792574)

[Insights and Visualizations: 6](#_Toc154792575)

[Results: 6](#_Toc154792576)

[Conclusion: 7](#_Toc154792577)

[REFERENCES 9](#_Toc154792578)

Decoding Twitter Emotions:

Uncovering Sentiment Patterns and Extracting Insights.

# Abstract

We have many social media applications, such as YouTube, Facebook, and Twitter, and there are different types YouTube is a video sharing type, and Twitter is a social media platform where people express their emotions in the form of text, so Twitter has provided researchers with valuable opportunities to gain insights into public sentiment and attitudes by accessing their tweets using the API (for that, we need a Twitter developer account).

So, in this project report, we focus on preprocessing using a big data tool and storing data in a NoSQL database, MongoDB Atlas on Cloud. After that, we analyze Twitter sentiment to identify meaningful patterns and categories the tweets (positive, neutral, or negative) of timeseries tweets using NLP and machine learning techniques. Sentiment analysis can be performed to classify tweets as positive, negative, or neutral, allowing for a comprehensive understanding of public opinion (their tweets).

# Introduction

Twitter is one of the popular social media platforms for the individuals to express their thoughts, opinions, and emotions. Twitter stands out as a valuable source of real-time data due to its vast user base and dynamic nature. Analyzing Twitter sentiment enables us to extract valuable insights into public sentiment, allowing for a deeper understanding of people's emotions and attitudes towards various topics. and also understand the kinds of people over the twitter based on the sentiment.

# Hypothesis

Twitter is one of the best social media platforms to get the data(text) for understand the sentiment. Twitter data can uncover significant patterns and trends in the expression of emotions, opinions, and attitudes, providing valuable insights. Specifically, by decoding Twitter emotions, we can identify sentiment patterns related to different topics, understand the demographics, kind of users who express specific emotions, and analyze the language and sentiment used to characterize discussions on Twitter. Through this analysis, we aim to extract meaningful insights that can inform decision-making processes, such as market research, Brand popularity, and public sentiment analysis.

# 

# Problem Statement

Understanding public sentiment is crucial for various applications, including market research, reputation management, and policy-making. However, manually analyzing a large volume of tweets is time-consuming and impractical. Therefore, automated sentiment analysis techniques, such as machine learning and natural language processing, are needed to efficiently process and analyze Twitter data, providing valuable insights into public sentiment.

Social media users generated a massive amount of data from the past few years, with a significant portion coming from social media platforms (Twitter, YouTube, Facebook, etc.) These platforms serve as a hub for people to share pictures, status updates, and thoughts, making the data obtained from them highly valuable for market research and understanding people's thoughts.

In our study, we have taken 1 million time-series tweets for sentiment analysis and uncover hidden patterns. By leveraging machine learning and natural language processing, we aim to extract meaningful insights from this vast amount of data.

## Objective:

To explore the concept of sentiment analysis on Twitter and its relevance in understanding public emotions and opinions. This involves reviewing previous studies and research that have focused on sentiment analysis techniques specifically applied to Twitter time series data.

To apply natural language processing (NLP) and machine learning techniques for sentiment analysis on a large dataset of Twitter data also include preprocessing data using big data technique and store data in NoSQL database or SQL database. The objective is to classify tweets into different sentiment categories, such as positive, negative or neutral, and evaluate the effectiveness of the chosen techniques in capturing Twitter emotions.

To uncover sentiment patterns and trends within the analyzed Twitter data. By analyzing the classified tweets, the aim is to identify common themes, popular sentiments, and any significant shifts in sentiment over time. This step involves visualizations, statistical analysis, and other exploratory techniques.

# 

# Literature Review

## Introduction:

Previous studies have extensively explored sentiment analysis of Twitter data(tweets). Researchers have employed various techniques, such as lexicon-based approaches, machine learning algorithms, and deep learning models, to classify sentiment in tweets. These studies have demonstrated the effectiveness of sentiment analysis in understanding public opinion and sentiment trends. In this study include different big data techniques for pre-processing such as pySpark, Spark. SQL, NLP technique for text cleaning, storing data in NoSQL databases (mongo DB), and for the un-labelled text forecasting we use library textblob etc.

One study conducted by Pak and Paroubek (2010) focused on sentiment analysis of Twitter data and compared different machine learning algorithms for classification. The study found that Support Vector Machines (SVM) and Naive Bayes classifiers performed well in identifying sentiment in tweets.

Another study by Go et al. (2009) explored sentiment analysis on Twitter using a large dataset of tweets. The study utilized a lexicon-based approach, where words were assigned sentiment scores, and the overall sentiment of a tweet was calculated based on the scores of the words present. The findings showed that sentiment analysis on Twitter can provide valuable insights into public opinion.

## Social Media:

Twitter, as a microblogging platform, has unique characteristics that make sentiment analysis challenging. The limited length of tweets, the use of abbreviations and slang, and the presence of hashtags and mentions add complexity to sentiment classification. Researchers have addressed these challenges by developing techniques specifically tailored to Twitter data.so the Text must be preprocessed before get analysis that required to use different library and Regular expression.

## Social Media data analysis:

We see many Applications of Twitter sentiment analysis in various domains, including politics, marketing, and brand management. Researchers have used sentiment analysis to track political sentiment during elections, assess customer satisfaction and sentiment towards brands, and monitor public sentiment towards specific events or topics.

# Methodology

The methodology for this analysis involves several steps:

Data Collection:

A large dataset of tweets related to the target topic or timeframe is pre-processed by the Twitter API by a third-party application that has access to this data, saving it to MongoDB for future action.

Pre-processing:

The collected tweets are pre-processed to remove noise, including URLs, special characters, etc. Tokenization and stemming techniques may also be applied by using regular expressions after reading by sparks and pySpark (a Python package that uses Hadoop distributed systems internally), etc.

Sentiment Classification:

The pre-processed tweets are classified into positive, negative, or neutral sentiments using textblob, which uses machine learning algorithms to classify the text based on polarity.

Insights and Visualizations:

The classified tweets are analysed to identify patterns, trends, and common themes. Visualisation techniques, such as word clouds, sentiment distribution charts, and trend graphs, are employed to present the findings in an easily understandable format.

Results:

The results of the sentiment analysis provide valuable insights into the public sentiment towards the target topic. The analysis reveals the distribution of positive, negative, and neutral tweet. here we find the people over twitter,

The Positive sentiment Percentage is higher than neutral and negative.

The negative sentiment is on average 20%(+-2%) over weekly, monthly and yearly.

The Neural and positive tweets percentage are close to each other.

# Conclusion:

So based on above results, I concluded that the overall sentiment expressed by users on Twitter tends to be positive and on Twitter higher percentage of positive sentiment compared to other tweets sentiment(neutral and negative sentiments) it suggests that the majority of users exhibit a positive attitude in their tweets.

Analysis of these tweets The consistent average of 20% (±2%) for negative sentiment across weekly, monthly, and yearly timeframes indicates that a certain portion of users consistently express negative emotions or opinions in their tweets or may use Twitter for negative reviews or spreading negative thought. This suggests that negative sentiment is present but relatively stable over time.

The analysis of neutral and positive sentiment The fact that percentages of neutral are close to each other implies that a significant portion of users express their thoughts or opinions without necessarily leaning towards a strong positive or negative sentiment. This suggests a more balanced sentiment distribution in terms of neutral and positive tweets. The percentage of positive and neutral tweets is approx. 80% (+-2).

Sentiment analysis on Twitter provides valuable insights into public sentiment and opinion over the time.

The application of various techniques, including machine learning algorithms and lexicon-based approaches, or pre-trained model has proven effective in classifying sentiment in tweets.

Despite the challenges posed by the unique characteristics of Twitter data, sentiment analysis on this platform offers opportunities for understanding public sentiment in real-time.

# REFERENCES

Google Scholar.com

<https://aclanthology.org/L10-1263/>

<https://dl.acm.org/doi/10.1145/1557019.1557114>

<https://datascienceparichay.com/article/get-data-from-twitter-api-in-python-step-by-step-guide/>

<https://www.toptal.com/apache/apache-spark-streaming-twitter>

<https://archive.org/details/twitterstream?sort=-publicdate>

1.          Details of the data storage and processing activities carried out, including preparation of the data and processing the data in a MapReduce/ Spark environment;**[0-20]**

2.          A discussion of the rationale and justification for the choices you have made in terms of data processing and storage, programming language choice, machine learning models and algorithms that you have implemented**.[0-40]**

3.          Comparative analysis for at least two databases using any benchmarking tool. (For example, ycsb)**[0-10]**

4.          Your analysis of  any change sentiment that occurs over the time period that you have selected.**[0-10]**

5.          Your forecast of the sentiment at 1 week, 1 month and 3 months going forward**[0-10]**

6.          Presentation of results by making appropriate use of figures along with caption, tables, etc and your dashboard for your forecast**.[0-10]**

Note that MapReduce-style processing in this instance is considered to include platforms such as Apache Spark.

Marks and feedback will be provided for each module separately based upon the learning outcomes for each of the modules.

**All documentation, code, examples, and any other files MUST be evidenced in your Version Control repository. (Git or Similar) using your CCT email address ONLY. Your repository MUST show your continual development of the project throughout its life-cycle and include notes detailing progress at each commit.**

**You may be called to a Viva to defend your work.**

**SUBMISSION:**

●        The final report must be submitted to Moodle on/before the deadline. Submissions received after the deadline will be subject to penalties.

●        Format of word file name should follow the format, studentID\_Integrated\_CA including your link to the Version Control repository

**Submissions that are suspected of plagiarism and/or inclusion of AI (CHATGBT, BARD etc…) Generated content will be referred to the college authorities.**