Group members: Rushikesh Kshirsagar Data: 2nd May 2021

Vishal Brambhatt

Vinay Basavaraja

**BIG DATA COURSE PROJECT**

**Twitter sentiment analysis**

**Introduction:**

The project is mainly for the sentiment analysis in the twitter; that is categorize tweets according to the sentiment expressed in them which can be positive, negative, or neutral. Twitter is an online micro-blogging and social-networking platform which allows users to express by writing short stories updates of maximum length 140 characters. It is a rapidly expanding service with over 200 million registered users - out of which 100 million are active users and half of them log on twitter on a daily basis - generating nearly 250 million tweets per day.

Due to the huge amount of usage we try to achieve an opinion of public sentiment by analysing the sentiments expressed in the tweets. Analysing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting elections and socio-demographic phenomena like stock exchange.

**Objective:**

To learn streaming of data from Twitter, which is one popular social media platform, and applying machine learning techniques using big data analytical tools like Apache Spark.

**Twitter Tweet Extraction:**

We have used Apache Spark and Python 3.5. We have used Tweepy to extract the tweets from the training dataset provided to us, then supplied the tweets to the sentiment analysis algorithm. The training and testing datasets are as provided.

**Functionality and Design:**

The process of designing a functional classifier is divided into four basic categories for sentiment analysis. They are as follows:

1. **Data Acquisition**

Data in the form of raw tweets is extracted by using the python library which provides a package for simple twitter streaming API that is tweestream [3].

A tweet extracted by this method has a lot of raw information in it which may or may not find useful for our particular application. It usually in the form of python dictionary data type with a lot of key-value pairs.

We further strain out the tweets to be labelled so that we have the large amount of variation in tweets without the losing generality. The filtering criteria which can be applied are mentioned below:

• Remove Retweets

• Remove very short tweets

• Remove similar tweets

1. **Human Labelling**

We labelled the tweets in three different classes according to sentiments observed in the tweets in train data: positive, negative, neutral.

1. **Feature Extraction**

We have used the tweets with relevant words as features and filtered waste words.

add\_stopwords = ["http","https","amp","rt","t","c","the","@"]

stopwordsRemover = StopWordsRemover(inputCol="words", outputCol="filtered").setStopWords(add\_stopwords)

1. **Classification**

Pattern classification is the process through which data is classified into different classes according to some usual patterns which are found in one class which differ to some degree with the patterns found in the other classes. The main aim of this project is to build the classifier which accurately categorise tweets in the following three sentiment classes: positive, negative, neutral.

**Twitter Authentication:**

API Key: DkQSTXh66NS6vdN97hijLHNwj

API Key Secret: DzIqNUur2Mw7RY7v7P3ZG6lNYKoLlcQdwOekun1fRZSVQWZxzE

Access Token: 946090488880234496-1nswb3KRxBCe0Vmno1giURgNNuTcyQd

Access token Secret: ENbJnpjykvIGrIaNNOR8feCTgdjKEEGkUc29rlZQ2JyYh

**Output:**

The output of the machine learning algorithm stored in a database as stated below:

![Graphical user interface, text, application

Description automatically generated]()

The root folder with all the code is attached with this report, along with the outputs of various steps executed by the code.

**References**

[1] Maggie Shields, Technology Reporter, BBC News. Twitter co-founder Jack Dorsey re-joins company. <<https://www.bbc.com/news/business-12889048>>

[2] Ben Parr. Twitter Has 100 Million Monthly Active Users; 50% Log-In Every day.

<<https://mashable.com/2011/10/17/twitter-costolo-stats/>>

[3] Tweet Stream: Simple Twitter Streaming API Access <<https://pypi.org/project/tweetstream/>>