**Text Mining with Sentiment Analysis using Python**

Project report submitted in partial fulfillment of the Requirements for the

Award of the Degree of B. Tech. in

Computer Science and Engineering

**BY**

**Yatharth Shrivastava**

**(University Roll No. 2017145)**

**B. Tech. (CSE)**

****

**Department of Computer Science and Engineering**

**Graphic Era** (**Deemed to** **be University)**

**Dehradun-248002**

**CONTENTS**

* **Acknowledgement**
* **Introduction**

1. **Problem statement**
2. **Motivation**

* **Tools used**

1. **Pycharm**
2. **Libraries**
3. **Tweepy**
4. **re**
5. **time**
6. **flair**

* **Methodology**
* **Results**
* **Conclusion**
* **Bibliography**

**ACKNOWLEDGEMENT**

Hereby I am submitting the project report on **“Text Mining with Sentiment Analysis using Python”** as per the scheme of Graphic Era Deemed to be University, Dehradun.

I would like to express my sincere gratitude to **Ms. Vishu Tyagi, Associate** Professor Graphic Era Deemed to be University, for providing a congenial environment to work in and carry out my project.

I am very much thankful to all the faculty members of the Department of Computer Science and Technology, friends and my parents for their constant encouragement, support and help throughout the period of project conduction.

Yatharth Shrivastava

**Roll no. 2017145**

**INTRODUCTION**

**Problem Statement: Text Mining with Sentiment Analysis using Python.**

An application to detect the sentiments of the tweets in real time for the specific keyword through text mining.

**Motivation:**

Text mining is also known as text analysis. It is the process of transforming unstructured text

into structured data for easy analysis. Text mining needs natural language processing (NLP), enabling devices to learn the human language and process it automatically. It is defined as the process of extracting essential data from standard language text. Some data that we generate via text messages, documents, emails, files are written in common language text. Text mining is generally used to draw beneficial insights or patterns from such data. Text mining is an automatic procedure that uses natural language processing to derive valuable vision from unstructured text. It can be transforming data into information that devices can learn, text mining automates the process of classifying texts by sentiment, subject and intent.

Text mining is required if organizations and individuals are to make sense of these vast information and data resources and leverage value. The resources need first to be processed, accessed, analyzed, annotated and related to existing information and understanding. The processed data can then be 'mined' to identify patterns and extract valuable information and new knowledge. How these information and data resources are analyzed depends on their format. Structured data can be relatively easily mined as the structure can be used to aid processing. Using a computer to automatically analyze information contained in documents is however much more difficult. Most digital documents consist of unstructured text containing flat data, rather than structured and meaningful information, which cannot directly be automatically processed by a computer in a useful way. Text mining therefore involves more complicated processes than structured data mining, and it is the processes involved that give rise to the conflict with copyright law. Given the volume of text generated by business, academic and social activities in for example competitor reports, research publications or customer opinions on social networking sites text mining is, however, highly important. Text mining is used in various fields like fraud detection, sentiment analysis, banking and finance, cybersecurity, digital analysis, risk management etc. Companies use text mining and natural language processing to improve customer service. Text analytics software are used to extract key and relevant information from customer surveys, call notes and other communications. Sentiment analysis (opinion mining) is a text mining technique that uses machine learning and natural language processing (NLP) to automatically analyze text for the sentiment of the writer (positive, negative, neutral, and beyond). The overall purpose of text mining is to derive high-quality information and actionable insights from text, allowing businesses to make informed decisions. Powerful machine learning algorithms can easily recognize statements as Positive, Negative, or Neutral.

**Tools Used:**

**Pycharm :** PyCharm is a hybrid-platform developed by JetBrains as an IDE for Python. It is commonly used for Python application development. It has an intelligent code editor which also helps in identifying errors more efficiently, better code navigation, it helps developers create web applications in Python. It supports popular web technologies such as HTML, CSS, and JavaScript. Developers have the choice of live editing with this IDE. At the same time, they can preview the created/updated web page. It also supports popular frameworks such as Django and has wide variety of libraries at its disposal.

**Libraries involved:**

**1) Tweepy:** Tweepy is open-sourced, hosted on GitHub and enables Python to communicate with Twitter platform and use its API. Tweepy supports accessing Twitter via Basic Authentication and the newer method, OAuth. Twitter has stopped accepting Basic Authentication so OAuth is now the only way to use the Twitter API. Main Model classes in the Twitter API are Tweets, Users, Entities and Places. Access to each returns a JSON formatted response and traversing through information is very easy in Python.

**2) re:** A regular expression is a powerful tool for matching text, based on a predefined pattern. It can detect the presence or absence of a text by matching with a particular pattern, and also can split a pattern into one or more sub patterns. The Python standard library provides a re module for regular expressions. Its primary function is to offer a search, where it takes a regular expression and a string.

**3) time:** The time library in Python is used to obtain time in the real world and perform various tasks related to it. One can even manipulate execution time using this module.

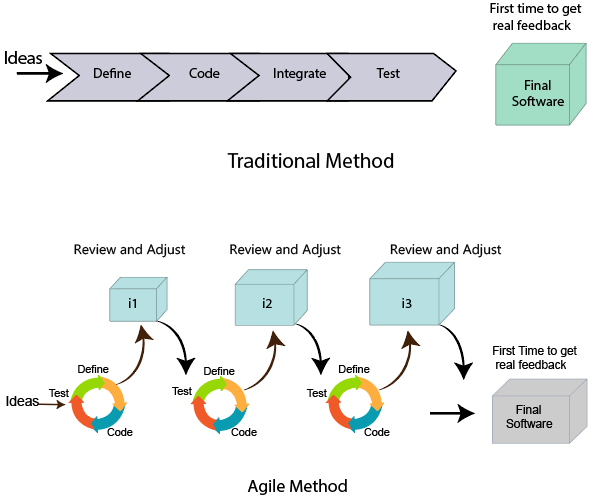
4) **flair:** Flair is a powerful open-source library for natural language processing. It is mainly used to get insight from text extraction, word embedding, named entity recognition, parts of speech tagging, and text classification. All these features are pretrained in flair for NLP models. It also supports biomedical data that is more than 32 biomedical datasets already using flair library for natural language processing tasks. Easily integrated with Pytorch NLP framework for embedding in document and sentence.

Tweepy is used to get the tweets in real time for the corresponding keyword. Flair is a pretrained sentiment analysis model used to detect the sentiments of the tweets.

**Methodology:**

Project follows an agile methodology,

An agile methodology is an iterative approach to software development. Each iteration of agile methodology takes a short time interval. The agile development process is aligned to deliver the changing business requirement. It distributes the software with faster and fewer changes. In single-phase development, all the requirement gathering and risks management factors are predicted initially. The agile software development process frequently takes the feedback of workable product.



Obtaining twitter API tokens through twitter developer docs.

The text mining process contains the following steps to extract the data from the files which are as follows −

**Document Gathering** − In the first step, the text documents are collected, which are present in several formats. The document can be in form of pdf, word, html doc, css, etc.

**Document preprocessing** − In this process, the given input document is processed for eliminating redundancies, inconsistencies, independent words, stemming and files are prepared for the next step, and the stages implemented are as follows −

**Tokenization** − The given document is treated as a string and recognized single word in the document that is the given document string is split into one unit or token.

**Removal of Stop word** − In this process the removal of constant words such as a, an, but, and, of, the, etc.

**Stemming** − A stem is a natural set of words with similar meanings. This approach defines the base of a specific word. There are two types of methods are Inflectional and derivational stemming. One of the famous algorithms for stemming is porter’s algorithm such as if a document pertains to words like resignation, resigned, resigns then it will be treated as resigning after using the stemming method.

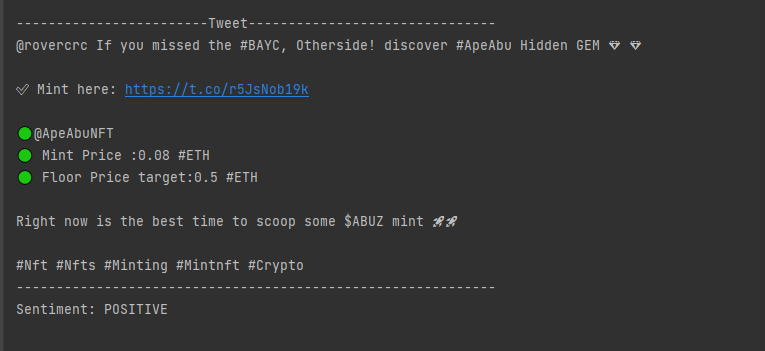
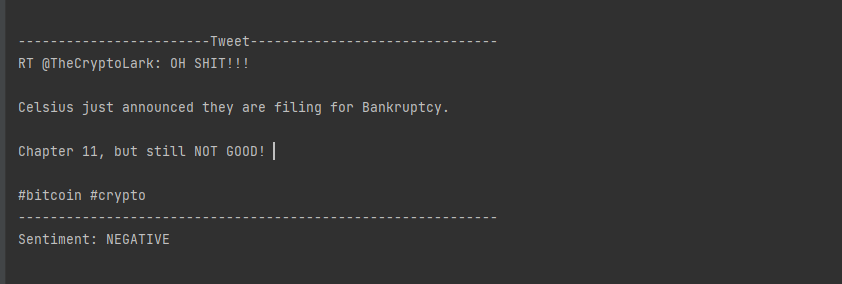
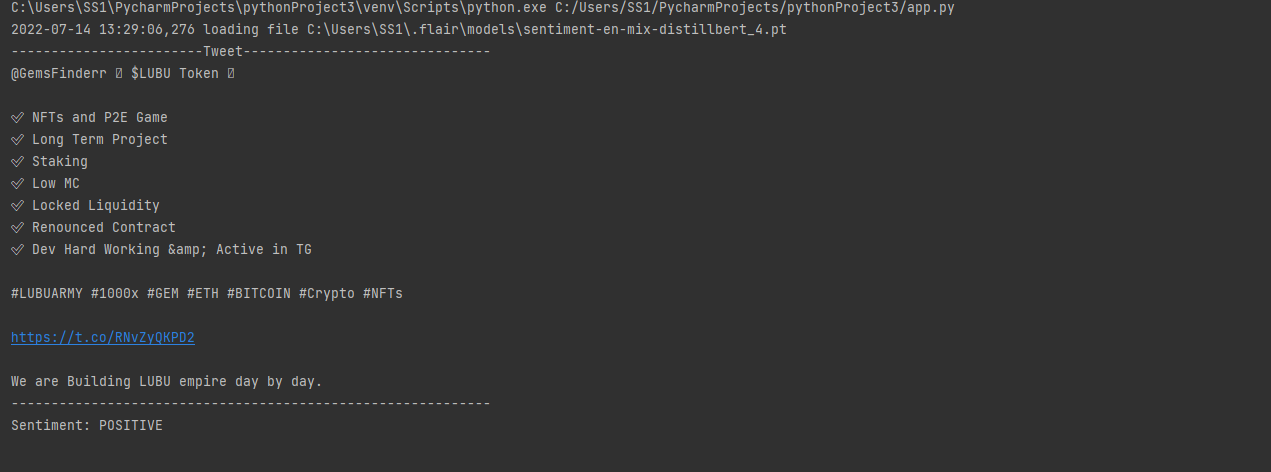
**Text Transformation** − A text document is a set of words (feature) and their appearances. There are two methods for representations of such documents are Vector Space Model and Bag of words.

**Feature Selection (attribute selection)** − This approach results in providing low database space, minimal search methods by taking out irrelevant natures from the input document.

**Data mining/Pattern Selection** − In this process, the conventional data mining process combines with the text mining process. A structured database facilitates classic data mining techniques that resulted from an earlier stage.

**Evaluate** − This stage calculates the outcome. This resulting outcome can be focused away or can be used for the following set of sequences.

**Results:**

****

**Conclusion:**

Text mining with sentiment analysis offers powerful data analysis insights and dynamic results, no matter the type of text you need to analyze. And once you train a sentiment analyzer to your specific needs, you can analyze your unstructured text at speeds and levels of accuracy you never thought possible. This program is a simple explanation to how this kind of application works. It’s only for academic purposes, as the program described here is by no means production level.

**Bibliography:**

1. https://docs.python.org
2. https://developer.twitter.com
3. https://pypi.org/project/flair
4. https://www.tweepy.org
5. https://www.tutorialspoint.com
6. Python for data analysis by Wes McKinney