**DATA SCIENCE BCSC 4227**

**BITC01/0054/2019  
KIMUTAI GRIFFINS**

**Project Scope:** Analyzing tweets collected using the query term “World Cup Finals”

**Project Report**

Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights.

**Data Collection**

The first step of text mining is gathering the data. With this, the python tweepy library came in handy to facilitate the pulling of data from Twitter. The data pulled from the API is unstructured data- (This data does not have a predefined data format. It can include text from sources, like social media or product reviews, or rich media formats like, video and audio files.)

The data collected has to be cleaned and transformed into usable format.It involves the use of techniques such as language identification, tokenization and syntax parsing.

**Information retrieval**

Returning relevant information based on a pre-defined set of queries.

Tasks performed include :

Tokenization - breaking out long-form text into sentences

**Stemming:** separating the prefixes and suffixes from words to derive the root word form and meaning.

### Natural language processing (NLP)

Ability to understand text and spoken words in much the same way human beings can.

Tasks involved include :

**Sentiment analysis**: This task detects positive or negative sentiment from data sources, allowing you to track changes in customer attitudes over time.

**Summarization** : provides a synopsis of long pieces of text to create a concise, coherent summary of a document’s main points.

### Information extraction

Extracting structured information from free text and storing these entities, attributes, and relationship information in a database.

**Feature selection,** or attribute selection,is the process of selecting the important features (dimensions) to contribute the most to output of a predictive analytics model.

**Feature extraction** is the process of selecting a subset of features to improve the accuracy of a classification task. This is particularly important for dimensionality reduction.

**Named-entity recognition (NER)** also known as entity identification or entity extraction, aims to find and categorize specific entities in text, such as names or locations. For example, NER identifies “California” as a location and “Mary” as a woman’s name.

**Importance of the performed analysis**

**Sentiment analysis-**

provides information about perceptions of brands, products, and services. These insights can propel businesses to connect with customers and improve processes and user experiences.

Providing a mechanism for companies to prioritize key pain points for their customers, allowing businesses to respond to urgent issues in real-time and increase customer satisfaction.

**Risk management**

**Give insights into industry trends and financial markets by tracking shifts in sentiment and pulling information from analyst reports and whitepapers This is especially beneficial to financial organizations since it gives them more confidence when contemplating company investments in diverse areas.**

Text mining tools analyze documents to identify entities and extract relationships between them, unlocking hidden information to help researchers:

****Challanges encountered during the proccess****

****Short Informal texts** - one of the difficulties in sentiment analysis is short informal text.They are restricted in length, usually spanning one or less sentences.They frequently use slang words, misspellings, and truncated word forms.**

Many of the collected tweets are stripped of context or too short to serve as real carriers of meaning. Most of the tweets have the meaning hidden in images or links to websites.

Twitter API has a request limit that limits the amount of data to collect and analyse.