CIND 820

Capstone Project Abstract

Name: Pallavi Thirunavukarasu

Student Number: 501114919

Supervisor: TBD

Date: January 24, 2022

## Theme:

Text Mining and Sentiment Analysis

## Objective:

The purpose of this project is to apply Text Mining and Sentiment Analysis using Natural Language Processing techniques on Twitter data:

* Identify the top topics/context within the tweets
* Group the tweets based on the topics
* Classify the tweets by the sentiment they express within each topic
* Visualize the results of the sentiments expressed by topic, sentiment, location, number of followers for the user.

## Introduction:

A Topic Model is a type of statistical model for discovering the abstract topics that occur in a collection of documents. It is used to identify the hidden semantic structures in the text body.

Sentiment Analysis of textual data is a process of determining whether the text is positive, negative or neutral. A sentiment analysis system for textual data combines natural language processing (NLP) and machine learning techniques to assign a positive, negative or neutral score to the text. There is rule based, machine learning based or deep learning AI based sentiment analysis techniques.

The global user community is increasingly connected through social networking apps and websites. Twitter is currently the largest microblog website in the world, is extensively used by people, common users and influencers to express their opinions, to raise questions, review brands/products. Opinion Mining from social networking sites help us to understand social behavior, gauge public opinion towards a topic and predict trends. Organizations can understand customer reactions for a product from opinion mining. The challenges for opinion mining from twitter are the short length of the tweets due to the length restrictions and informal language used.

## Research Questions and Methods:

The main research question in this project is to explore the sentiments expressed in tweets within a topic(context) on Twitter data. The plan is to use twitter data collected with a specific hashtag(covid19). Given the pandemic situation there is a need to identify the most talk about topics within the tweets. Others features to be analyzed are what was the sentiment expressed about these topics by users, how the sentiments vary by location, what are the sentiments expressed by the social media influencers (users with large number of followers) on different topics.

The machine learning methods, Natural Language processing techniques to be applied in order to address these questions are as follows:

1. Topic modeling for identifying the top topics from the tweets using

* Latent Semantic Analysis or Latent Semantic Indexing (LSA)
* Latent Dirichlet Allocation (LDA)
* Non-Negative Matrix Factorization (NMF)

1. Clustering the tweets based on the topics
2. Building a rule based or hybrid model for the classification of the tweets in each cluster
3. Visualize the results of the top topics and the sentiments expressed within those topics.
4. Language to be used is R, Python and suitable libraries of Python, R.

## Datasets

The dataset to be used for this project are tweets collected with the hashtag of covid19.

<https://www.kaggle.com/gpreda/covid19-tweets>. The dataset has about 17k records. There are 13 columns: User\_name, User\_location, User\_description, User\_created, User\_followers, User\_friends, User\_favourites, User\_verified, date, text, hastags, source, is\_retweet.