**DSC 478 PROJECT PROPOSAL**

**Prof. Mobasher**

**Sentiment Analysis with Tweets**

**Team Members:**

**Rawan Hammad**

**Anisha Patel**

**Saurin Patel**



# Introduction

The growth of social networks like: Twitter, Facebook, LinkedIn, Google+, Pinterest and others, has generated a large amount of information about the preferences and behavior of the users. The majority of the data that is constantly generated in the social networks could contain valuable information like perceptions and tendencies from the users to the objects, personalities or services. To process the data that exist in the social networks is necessary the usage of linguistic mechanisms that contribute to define the correct sense of the words. One of the areas that have surged of the Natural Language Processing (NLP) to analyses the information contained in the web either in sentences or in documents is the sentiment analysis, also denominated opinion mining, which was initially focused mainly to electronic commerce but nowadays has been expanded to other fields like education, medicine, politics and others. Due to increase in the number of users, companies are trying to use the sentiments of users to analyze their opinions. In this project our goal is to determine the sentiment of a user from a tweet.

.

# Process

The specific dataset that will be used can be found in the link below:

<https://www.kaggle.com/datasets/kazanova/sentiment140>

The dataset has 6 attributes and 1.6 million tweets. Briefly, the dataset contains the following 6 fields:

1. **target**: the polarity of the tweet (*0* = negative, *2* = neutral, *4* = positive)
2. **ids**: The id of the tweet ( *2087*)
3. **date**: the date of the tweet (*Sat May 16 23:58:44 UTC 2009*)
4. **flag**: The query (*lyx*). If there is no query, then this value is NO\_QUERY.
5. **user**: the user that tweeted (*robotickilldozr*)
6. **text**: the text of the tweet (*Lyx is cool*)

Using this data and the machine learning algorithm NLP, we can mine the opinion provided by user based on the language used in their tweets.

In this project, we will be following the knowledge discovery process. There are currently no null values in the given dataset. The project will include exploratory analysis of the tweets. We will use Matplotlib to plot the graphs. We will use data mining techniques, clustering, classification, and regression to analyze the dataset. For Supervised methods, we will use Classification and regression. The goal will be to optimize the parameters of the modelling techniques. We will incorporate Ensemble methods such as Random Forest or ADABoost along with basic approaches such as Decision Trees, KNN, and Naïve Bayes. For regression, we will try different types of regularization. Feature selection (PCA) and feature reduction may be used to increase the accuracy and decrease overfitting, as needed. Some of the tools that will be used in this project include Jupyter notebooks, using Python, as well as Python libraries such as Pandas, NumPy, Sklearn, and other libraries as needed.

Logo

Description automatically generated

# Vision and Goals

The goal of this project is primarily to practice the use of NLP machine learning algorithm and to create a solution as a team. Due to the nature of the data provided, the results can only be used for research purposes and cannot be used for commercial purposes, per the disclaimer provided within the data source’s website.

# Resources

The table below contains the resources used or to be used in this project.

|  |  |  |
| --- | --- | --- |
|  |  | **Project Resources** |
| **Type** | **Quantity** | **Notes** |
| Data files | 1 | https://www.kaggle.com/datasets/kazanova/sentiment140 |

# Success Criteria

The success criteria for this project is listed in detail on this page:

<http://facweb.cs.depaul.edu/mobasher/classes/csc478/project-checklist.html>