**NATURAL LANGUAGE PROCESSING PROJECT PROPOSAL**

**TITLE: SENTIMENT ANALYSIS ON TWITTER MESSAGES**

**GitHub Link:** <https://github.com/NitheeshSama/NLP_project.git>

**Team Members:**

**Nitheesh Sama**

**Lokesh Bachu**

**Rajasri Lekkala**

**Sai Krishna Bussa**

**MOTIVATION:**

Twitter has been a stage where people share their views and opinion on a particular product, thing, movie or an issue. The amount of data that is produced from this is so large and an organization that wants to know the public opinion on their products can make use of this data and analyze it to improve their products or services. Sentiment analysis technique can be applied to get the required analysis from twitter dataset. As part of our project, we are using a twitter dataset and perform sentiment analysis to gain insights into the data. This technique can be further modified according to different data to gain insights into them.

**OBJECTIVE:**

In this project we will classify twitter messages into Positive, Negative, Neutral and Irrelevant based on the content of the messages. We will be using python programming with packages like panadas, nltk, ski-kit learn, etc accordingly to perform sentiment analysis on the dataset. After performing the sentiment analysis, we would apply Machine Learning classifiers, evaluate the model and find the accuracy of the model. At the end all the messages will be classified as positive, negative, neutral or irrelevant and the accuracy of the classification can be determined.

**SIGNIFICANCE:**

This type of projects plays very important role in fields of Market Analysis, Brand Monitoring, Business Intelligence Build up and Customer Service. Sentiment analysis is very important to understand and detect customer feelings. It is important to accomplish the project purpose using sentiment analysis because to understand real user opinions, complaints and suggestions, we must again filter the unrelated Tweets (Spam, junk, marketing, news and random) and this method is very helpful in doing so. These types of analysis are also used for strategic purpose by organizations to know their competition.

**FEATURES:**

Some of the important features of our project are

1. **Data Collection**: In this project we are using data collected from twitter and saved in a csv file as a data source. The dataset is obtained from Kaggle.com and contains data collected from twitter on games like **Borderlands**, **Dota-2** and **World of Craft** etc.
2. **Data Processing:** In this project, the data processing is done at different levels like removal of punctuation, removal of stop words, conversion to lower case and Lemmatization etc. These steps are helpful in making the analysis part easy.
3. **Data Analysis:** As part of data analysis, we will be classifying the messages into one of the 4 categories using sentiment analysis.
4. **Data Visualization:**  After the analysis part, we will be visualizing the number of messages in each category, most frequent words in that category and so on (As this is the early stage of the project, we did not come a conclusion on which visualization to use)

After the sentiment analysis part is done, we are planning to apply decision tree classifier on the and then evaluate the resulting model to find the accuracy, important features etc.

**WORKFLOW OR STEPS:**

The steps that will be performed as per the project are

1. **Data Preparation**
2. **Removing punctuation and replacing short words.**
3. **Lower Case Conversion.**
4. **Stop words Removal.**
5. **Lemmatization.**
6. **Analysis**
7. **Machine Learning Model fitting**
8. **Evaluation of model**

**REFERENCES:**

**Data Set reference:** [**https://www.kaggle.com/**](https://www.kaggle.com/)

**DATASET DETAILS:**

**Name: twitter\_training.csv**

**Number of Features: 4**

**Number of rows: 74002**

**Data Source: Twitter**

**Data Set Source: Kaggle.com**