**Smart India Hackathon 2020**

**…………………………………………………………………………………………………………**

**Problem Statement-** A complete software web application which provide standard solution handle multiple code-mixed Indian languages English, Hindi, Kannada ,Bengali, Urdu etc and perform context aware Sentiment Analysis give results such as “Positive”, “Negative“ and “Neutral”.

………………………………………………………………………………………………………………

**Problem Analysis**

* **Analysis on Different languages**
* **Identifying Need of Maintenance**
* **Homophonic Word**
* **Noise Detection**
* **Strain in Finding Meaning**
* **Dataset without Sentiment**
* **Emoji Analysis**

**Solution**

* **Analysis on Different languages-** In this approach we train the classifier on English reviews and for testing, we translate the Hindi reviews or any other mixed language into English using Googletrans and then we classify the Sentiment of review and also we have separately worked on hinglish data by making stopwords dictionary for it.
* **Identifying Need of Maintenance-** There is increase in review and comments by users day by day, as users use slangs and words which are not originally present in language. To maintain such large dataset we will use tf-idf or CountVectorizer method to transform text into meaningful representation of numbers.
* **Challenges in finding abusive words-** We will be using machine learning based methods to detect hate speech on user comments from three sentiments Positive, Negative, Neutral.
* **Dataset without Sentiment-** In some dataset, there is no sentiments available. So, there is difficulty to find sentiments of comments from dataset. For that, our model first find the Sentiment Score from our design model and then further dataset is processed.
* **Graphical Analysis-** We will present a detailed analysis by scatter plots, bar graphs, pi-charts etc. using python libraries.
* **Emojis handling-** For analyzing emojis we have imported UNICODE\_EMOJI library and also made a folder of all emojis with their meaning.

Requirements

* **python3** (Anaconda environment is preferred)
* **Scikit-learn** for performing different classification model on dataste
* **Pandas** to read and write csv files
* **Numpy** for performing different operations on dataset
* **NLTK** for text mining in dataset
* **Googletrans** for translating any mixed language to english
* **Pickle** to load the trained model for future use
* **Codecs** to provide stream and file interfaces for transcoding data into program.
* **Matplotlib** for visualization
* **UNICODE\_EMOJI** for analyzing emojis in dataset

We have used three approaches to classify the sentiment of text reviews as positive or Negative.

1. **Resource Based Semantic Analysis using HindiSentiWordnet**.---> In this approach we used Hindi Sentiwordnet to classify the review's sentiment.
2. **IN language Semantic Analysis**. : This approach is based on training the classifiers on the same language as text.
3. **Machine Translation Based Semantic Analysis**. : In this approach we train the classifier on English reviews and for testing, we translate the Hindi reviews or any other mixed language into English using Googletrans api and then we classify the Sentiment of review.