**Team members:**

*Rahul Mandaloju, 11514828*

*Lokesh Punati, 11514358*

# **Problem Statement**

Twitter is a popular social networking website where members create and interact with messages known as “tweets”. This serves as a means for individuals to express their thoughts or feelings about different subjects. Various different parties such as consumers and marketers have done sentiment analysis on tweets to gather insights into products or to conduct market analysis. Furthermore, with the recent advancements in machine learning algorithms, we are able to improve the accuracy of our sentiment analysis predictions. In this report, we will attempt to conduct sentiment analysis on “tweets” using various different machine learning algorithms. We attempt to classify the polarity of the tweet where it is either positive or negative. If the tweet has both positive and negative elements, the more dominant sentiment should be picked as the final label.

We use the dataset from [Kaggle](https://www.kaggle.com/c/cs5228-project-2/data) which was crawled and labeled positive/negative. The data provided comes with emoticons, usernames and hashtags which are required to be processed and converted into a standard form. We also need to extract useful features from the text such as unigrams and bigrams which is a form of representation of the “tweet”. We use various machine learning algorithms to conduct sentiment analysis using the extracted features. However, just relying on individual models did not give a high accuracy so we picked the top few models to generate a model ensemble. Ensembling is a form of meta learning algorithm technique where we combine different classifiers in order to improve the prediction accuracy. Finally, we report our experimental results and findings at the end.

# **Data Description**

The data given is in the form of comma-separated values files with tweets and their corresponding sentiments. The training dataset is a csv file of type tweet\_id,sentiment,tweet where thetweet\_id is a unique integer identifying the tweet, sentiment is either 1 (positive) or 0 (negative),and tweet is the tweet enclosed in "". Similarly, the test dataset is a csv file of type tweet\_id,tweet.

Positive and negative words collection in a text file.Around 10000 words in which 5000 are positive and 5000 are negative are collected and compared when labeling tweets as positive or negative.

**Feature Extraction**

We extract two types of features from our dataset, namely unigrams and bigrams. We create a frequency distribution of the unigrams and bigrams present in the dataset and choose top N unigrams and bigrams for our analysis.

# **Conclusion**

## **Summary of achievements**

The provided tweets were a mixture of words, emoticons, URLs, hashtags, user mentions, and symbols. Before training we pre-process the tweets to make it suitable for feeding into models. We implemented several machine learning algorithms like Naive Bayes and Decision Tree to classify the polarity of the tweet. We used two types of features namely unigrams and bigrams for classication and observes that augmenting the feature vector with bigrams improved the accuracy. Once the feature has been extracted it was represented as either a sparse vector or a dense vector. It has been observed that presence in the sparse vector representation recorded a better performance than frequency.

Our model achieved an accuracy of 83.4% on the Kaggle dataset.

## **Future directions**

**Handling emotion ranges:** We can improve and train our models to handle a range of sentiments. Tweets don’t always have positive or negative sentiment. At times they may have no sentiment i.e. neutral. Sentiment can also have gradations like the sentence, This is good, is positive but the sentence, This is extraordinary. is somewhat more positive than the rst. We can therefore classify the sentiment in ranges, say from -2 to +2.

**Using symbols:** During our pre-processing, we discard most of the symbols like commas, full-stops, and exclamation marks. These symbols may be helpful in assigning sentiment to a sentence.

References: https://www.kaggle.com/c/cs5228-project-2/data