Collecting and Performing Sentimental Analysis on Tweet

Project Proposal

Kabilan Tamilmani, Karthik Sundararaj & Vishwas Sharma

Overview:

Sentiment analysis is a process to identify and classify the message in text data. The message is subjective information about a particular topic. The most common type of classification in sentiment analysis is polarity detection (positive, negative, or neutral). Data available in non-traditional sources such as reviews and social media posts can be valuable if we could perform sentiment analysis. With this analysis, an organization can gather feedback about their new products and services or measure their marketing impact and make decisions in real-time.

Our idea is to develop a project which can collect tweets about the given topic from Twitter, pre-process them and perform sentiment analysis over the pre-processed data. This could give us valuable information such as what the people feel about the particular topic and the most common messages (an appreciation or critic) from the social media community.

Dataset:

- We'll use the 'Sentiment140' dataset [1] from Kaggle as our train dataset. It has contains 1.6m tweets labeled as either positive or negative or neutral.
- We'll use Twitter Rest API or Streaming API to collect tweets about a particular topic.

Evaluation:

The dataset 'complete-tweet-sentiment-extraction-data' [2] from Kaggle is our test dataset. It has 40,000 tweets labeled as either positive or negative or neutral. We will predict the sentiment label for this dataset and compare the actual and predicted sentiment labels. By this comparison, we could estimate the accuracy of our model.

Workflow:

With the help of Twitter APIs, we could collect tweets about a particular topic (user desired topic). The collected tweets are usually unstructured text data. This needs to be structured into key fields and stored. This will be our unlabelled dataset. Then sentiment analytics is performed on this unlabelled dataset. Then similar tweets have to be Clustered with the tf-idf method. This would avoid the repetition of the same messages in the data. Finally, the results are visualized to show the trend of sentiments about the topic.

Reference:

1. https://www.kaggle.com/kazanova/sentiment140
2. https://www.kaggle.com/maxjon/complete-tweet-sentiment-extraction-data