Method

Materials

We have used Twitter as our data source to build a corpus of Hindi tweets with the help of Twitter's Streaming API. We were able to download live tweets from 09 April 2017 until 20 April 2017 and create 12 files for each day. The program that downloaded Hindi tweets ran on Google Cloud Platform. The API parameters were updated to filter tweets in Hindi language and the geo location was set to India.

A total of 169922 tweets were collected during a period of 12 days. During this period, some of the major events that were tweeted about in hindi were "Sonu Nigam's comment on Azan", "Gaurakhsaks of UP", "CM of UP", "Problems with EVM", "Bail for Asaram", "Romeo squad", "Indian Soldiers", "Kulbhushan Jadhav" etc.

The preprocessing step includes removal of punctuations, stop words removal using a list of stopwords published by IIT-Bombay[6] and University of Neuchatel[17], extract meaningful words using tf-idf and using transliteration API to convert hinglish data to Hindi[4]. We have retained emoticons and hashtags in our dataset as they contribute to the sentiment of the topic.

Since there was no annotated corpus for sentiment analysis, we manually annotated transliterated tweets for our model. We classified 2000 tweets into 3 classes: positive (1), negative (-1), neutral (0). This hand annotated data is used to train our random forest model. Apart from this, we use Hindi WordNet to label the tweets as positive, negative neutral depending on the words in our tweets which forms our baseline model.