Project 3: Cricket Sentiment Analysis for India vs Australia Cricket World Cup Semifinal 2015

Team Members: Abhishek Agrawal(akagrawa), Karthik Krishna Gadiraju(kgadira), Nirmesh Khandelwal(nbkhande), Nisarg Gandhi(ndgandh2), Rohit Arora(rarora4).

1. Background

In this project we have implemented Cricket Sentiment Classification for World Cup semifinal match between India vs Australia 2015 from Twitter's stream using R and NodeJS. In this project we simulate real-time streaming using MongoDB APIs. This is a supervised learning task, which means that in order to train a classification model we identify the class of each tweet in advance. In this end-to-end analysis task, we initiated with data collection and merging in MongoDB, followed by data pre-processing, feature engineering, dynamic model building, dynamic model update and prediction in R.

In the subsequent sections we discuss each of these steps in detail.

2. Data Collection & Merging

Since cricket matches are scheduled events, and tweets about these events can be collected only during the match hours, one of the challenge was to collect data during the event, in order to use it for future analysis. This data can be simulated as streaming data. We collected nearly 600,000 tweets (covering complete match hours) over the duration of twelve hours and stored it in MongoDB.

NodeJS's nTwitter package was used to retrieve tweets from Twitter containing the words 'CWC2015', 'CWC15', 'INDVSAUS', 'AUSVSIND', 'WorldCup2015', 'IND VS AUS', 'AUS VS IND', 'Worldcup', and 'ICCWorldCup2015'. Please note all these are unbiased words that don't support a particular team.

Since, Twitter API tends to break over the duration of data collection; the data was collected on different machines simultaneously to avoid data loss. Later these tweets from different machines were merged on the basis of Tweet ID in a new MongoDB collection.

3. Data Pre-processing

The data is first split into Training and Test datasets. Initially, 1000 tweets are fetched from the MongoDB for training, these tweets are pre-processed before model creation. The following text mining transformations were made on tweet's text using R package 'tm':

- Converting text into documents
- Removing Twitter handles and punctuations marks
- Removing digits and words of length less than 3.
- Converting tweet's text to lower case
- Filtering stop words based on custom stop word list.
- Striping white spaces.
- Stemming (i.e., retrieving the root of a word)

4. Feature Engineering

In this step, term-document matrix was constructed. To avoid curse of dimensionality we only considered words with frequency higher than configured threshold. This helped in selecting a subset of the words to build our term-document matrix.

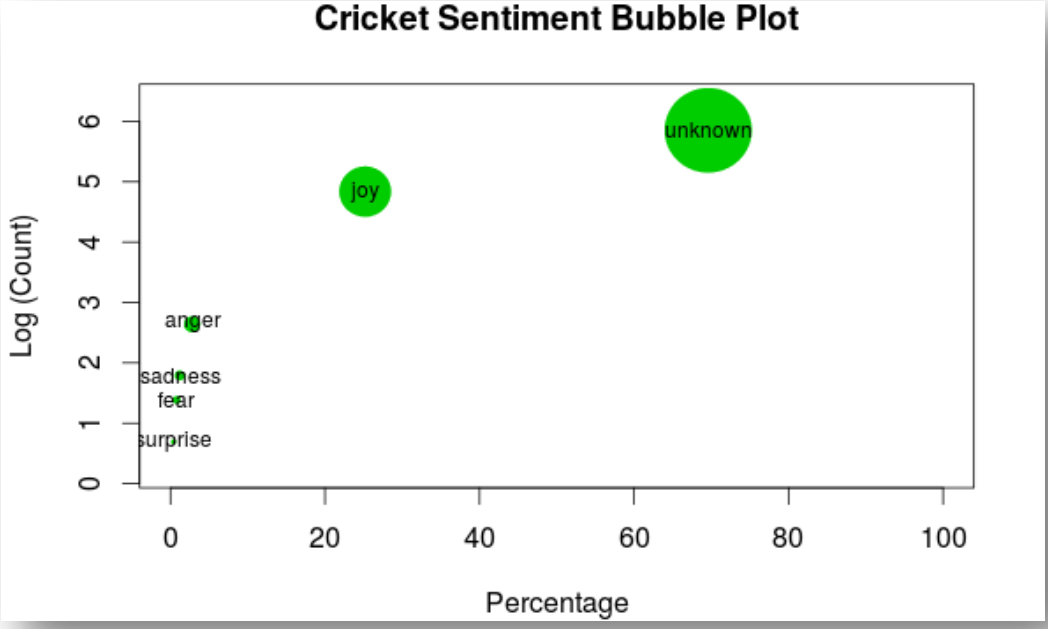
5. Dynamic Model Building

In order to train our classification model we generated the ground truth using R's 'sentiment' package that labels the tweets as 'anger', 'joy', 'sadness', 'fear', 'surprise', 'disgust', and 'NA' (or 'Unknown' in case not recognized).

The trained term document matrix was then converted to a data stream and trained with RMOA package using 'Naïve Bayes Multinomial' algorithm.

6. Test Data Prediction and Dynamic Model Update

After training the model, the same pre-processing steps were performed for the test data set. We extracted the features from the trained model and projected it on the test data. Then, label for the test data was predicted using the trained model. This was validated against the ground truth (emotion label) generated using R's 'sentiment' package.



We then updated the model using the trainMOA function which takes current model as argument and the test dataset (plus ground truth) as training sample.

7. Conclusion

We were successfully able to implement end-to-end Sentiment Analysis for Cricket data. The primary reason for many tweets being classified as ‘unknown’ is because we are getting data from different languages, which the ‘sentiment’ R package is unable to identify. But when the program is run for significant amount of time, we start noticing different emotions such as ‘anger’, ‘joy’, ‘sadness’, ‘fear’, ‘surprise’ following a power distribution.

8. Future work

For the capstone project our idea is to extend the notion of sentiment analysis further, to view how these emotions change with respect to space and time.