Proposal Project 15 Draft 1

Team members:

Zhenyu Wu, Yiran Luo, Fan Bai, Hongsen Shi

Topic:

Predict What is the Next Viral Trend Using Sentiment Analysis

Abstract:

Social media, esp. Twitter, Youtube, and Facebook, have long been hosts for viral phenomena. Although the appearance of hot topics looks random, we believe we can grasp some latent prediction of future trends by mining out the sentiment information behind the user posts with machine learning techniques.

Data source:

1. Existing Twitter corpora
2. In-time tweets via Twitter's API *preferred
3. Or some others...

Concepts breakdown:

1. Sentiment score dictionary. Every meaningful word from a tweet has a score on different kinds of sentiments. (e.g Happy is 10 on joy, 1 on sadness, 1 on disgusted, etc) The score itself is a vector with multiple attributes.
2. The score of an entire tweet is the summation of scores of all meaningful words.
3. Scaling down the data source. May only use tweets from peak hours (e.g. 8AM~9AM 12PM~1PM and 8PM~10PM). Only select the representative ones.
4. Predict what is the next hot spot. Use the trending keywords along with the aggregated sentiment scores.
5. Comparative performance study using different classifiers. Linear, SVM, kNN, Naive Bayes, Max Entropy, Deep Neural Network, etc. Cross-validation for tuning the parameters.
6. Visualize the verdicts. The overall sentiment report will be presented using statistical graphs. Scatter dot plot could be the major one.

Predicted difficulties:

1. Typos, mis-spelling, foreign words and emoji's are all distractions.
2. Still, huge data source with high dimensionality.
3. Is only using positive/neutral/negative over simplifying the problem?

References:

1. Stanford CS229
2. Twitter Sentiment Analysis Project http://twittersentimentproject.blogspot.com/
3. Name: discover the sentiment of a brand, product, or topic on Twitter. http://help.sentiment140.com/home