The Tweet Rises

Aleksander Bello, Alexandru Cioc, Victor Duan, Archan Luhar, Louis O'Bryan

Motivation and Audience

- How are people feeling right now?
  - Interesting question for both sociological research and for corporate gauging of reactions
    - How do discussions on a product spread?
    - How do people react to specific events?
    - How quickly does sentiment change?

Twitter

- Microblogging
  - 140-character “tweets”
- Very popular
  - Has affected numerous world changes

Goal

- Categorize tweets based on emotional content
- Visualize the sentiments on a heatmap in a web browser in real-time
- Allow for filtering of Tweets based on topic

Literature

Go *et al.*

- Emoticons can be used to gauge sentiment
  - Accuracy > 80%
- Part-of-speech has a NEGATIVE effect

Kouloumpis *et al.*

- Expanded on Go *et al.*
- Lexicon features are good
  - Words that indicate positive or negative
- Micro-blogging features good
  - Emoticons, abbreviations, internet lingo

Other Notes

Bifet *et al*.

- Bag-of-words is good
  - Every word is independent
  - Ignore grammar and solely look at number of occurrences
- Learning models need a higher learning rate in order to guarantee convergence

Sakaki *et al*.

- i.i.d. exponential distribution can be assumed for *natural* events but NOT for the spread of media events

Challenges

- Correctly analyze Tweets
    - Use known data sets for verification
- Live Tweets via the Twitter Firehose API
    - Provides 1% of all real-time Tweets
- Visualization the Tweets
    - Can use map of USA and color changes to indicate the sentiment of a region

Project Plan

- Begin by working on the front and backend
  - 3ish weeks
  - Get basic visualization working
  - Get Tweet streaming working
    - Don’t worry about classification at first
- Get accurate classification
  - Remaining time
  - Improve, improve, improve
  - Try various methods according to the literature

Project Plan contd.

- Naive Bayes
    - Feature selection based upon Mutual Information
    - > 80% accuracy
- SGD
    - Vanilla implementation of SGD with fixed learning rate
    - >85 % accuracy on Edinburgh data set

Division of Labor

- In first few weeks, parallelize tasks
    - Retrieve twitter data and parse
    - NLP
    - Backend
    - Frontend
- Then everyone works on NLP for the remainder