Topic and Opinion Mining on Twitter over Time

MohammedJunaid Hundekar	hundekar@usc.edu	2548 4069-96
Sagar Makwana	smakwana@usc.edu	4640 7504 94
Ankita Jain	jainab@usc.edu	2914 3012 25
Ankit Kothari	ankothar@usc.edu	4138 8544 60

Traditionally, public opinions are generated using surveys. However, we believe opinions can be generated by analyzing the sentiment in the text. We attempt to highlight the potential of text streams as a substitute and supplement for traditional polling.

Up until now, to the best of our knowledge most of studies looking into opinion mining on Twitter, look at subset of data related to the topic of the study and build their model to that specific dataset. The aim of our project is to look at a subset of twitter data not limited to any topic, and attempt to find the most relevant topics in those tweets. After finding these topics we will then measure the opinions on them.

In order to build out sentiment classifier we plan to use the dataset provided by Stanford University (sentiment140) for English Language and the dataset provided by Spanish Society for Natural Language Processing (TASS) for Spanish Language.

Sentiment 140 contains over 1.6 million tweets labeled as positive or negative based on the presence of emoticons. Along with this we will be using hand classified tweets from SemEval which also includes tweets from a variety of topics. The TASS corpus contains over 68,000 Twitter messages, written in Spanish by about 150 well-known personalities and celebrities of the world of politics, economy, communication, mass media and culture, between November 2011 and March 2012.

The above described data will be used to train our model and we will be using twitter API to download tweets related to our languages and time frame to test the performance of our model.

In the preprocessing step we first extract the following features from the tweets : Hashtag, Handles, Urls, Emoticons, Punctuations, Repeating Characters. After extraction we normalize the features. Normalization includes stemming using Porter's Algorithm to reduce the inflected words to its root, stopword removal and case folding.

For modelling topics, we plan to use Latent Dirichlet Allocation (LDA) and Dynamic LDA to keep track of topics over time. We plan to compare it with the trending hashtags to see if relevant topics are discovered. For Opinion mining, we plan to use stanford sentiment140 as a baseline model. Stanford sentiment140 model uses just the presence of positive and negative emoticons to label the tweets. We plan to use not only the emoticons but the words as well to improve on the baseline model. Also we plan to compare our opinions with the polls from various news sources.

Tools to be used for the project include but not are limited to Twitter API, Google Cloud Server, Python, Anaconda

Distribution of Tasks:

Twitter crawling: Ankit Kothari
Opinion Model : Junaid Hundekar
Topic Modeler : Sagar Makwana
Build and normalize corpus,Building a tool to interact with our model : Ankita Jain

References:

1) Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.
2) Bo Pang,Lillian Lee. 2008. Opinion mining and sentiment analysis.
3) Zhunchen Luo, Miles Osborne, Ting Wang. 2012. Opinion Retrieval.

Word Count: 428 Words.