CIS630 Project 2 Proposal: Twitter Affect

Tao Feng, Yayang Tian, Chun Chen

March 10, 2013

1. **Summary**
   We plan to implement a fine-grained classifier that determines people's sentiment on Twitter.

2. **Challenge**
   Sentiment analysis on microblog has been little studies due to its special format and length-constraint. Although (Go, 2010; Pak, 2010; Bollen 2011) have proposed effective ways to classify sentiment on microblog, They are only limited to thumb-up and thumb-down classification. In other words, no prior research or promising results have effectively classified tweets into more than three classes. In this project, we focus on subjectivity detection and try to classify tweets into five classes. To better elaborate our results, we propose build a web application on top of that.

3. **Data Collection**
   To quickly identify useful data, we propose to use two resources:

   (a) **Twitter Hashtags** (Edinburg Twitter corpus, e.g. #fail, #success)

   (b) **Twitter Emoticons** (EMOT data set created by Go, Bhayani, e.g. :), : ( )

   Beyond the gold standard for training the classifier, we plan to use manually labeled dataset **ISIEVE** to evaluate our results. After collection, we use tokenization, normalization, and POS tagging for **preprocessing**. For instance, stemming, lowercase, URL substitutions, repetition reduction.

4. **Feature Collection**

   (a) Lexicon counts from CIS630 project 1 (b) Unigram (c) Bigram (d) Presence of emoticons

   We don't use pos features because (Go, 2012) shows that it's not useful for affect detection on Twitter.

5. **Machine Learning Methods**
   We try three ways (Pang, Lee 2002) with proportion of emotion words as base line (Alec Go, 2010).

   (a) Naive Bayes (b) Max Entropy (c) SVM

6. **Buiding Web App**
   We plan to build a web application on Google App Engine to provide interactive interface to users. Related work includes Twitratr and tweetfeel, both of which still have some limitations. We plan to improve them to be more fine-grained and domain specified.

7. **Schedule**

March 11 - 20	data collection, preprocessing, and feature extraction
March 21 - 31	machine learning, and web skeleton
April 1 - 10	full-fledged web application
April 10 - 15	results analysis and report

8. **Extras**
   If time permits, we would love to try the followings:

   (a) New lexcion: due to the sparseness of sentiment resources for microblogs (Wilson, 2011)

   (b) Geo Sentiment: it's interesting to have a sentiment distribution of tweets on a world map