Scott Bickel, Brian Golden, Steve Styer (Baker’s Kilobyte)

CMPSC 448

Project Milestone 1

1/27/14

# **Datasets**

Two datasets of tweets will be explored and used as learning data, one of 600,000 data points and another of 2 million data points. Datasets were collected by searching for tweets matching various “hashtags” over 5-12 weeks. Each of the data sets has the following schema/attributes for data points:

* Tweet ID – a unique identifier for the tweet
* Source User ID – the user that posted the tweet
* Retweet User ID – the user that posted the original tweet being retweeted
* Website – website referenced in the tweet, if any
* Tweet Time – timestamp for the tweet’s posting time
* Hashtag – the hashtag that the search found a match against, i.e. why the tweet is in the dataset
* Week number – the week number from the start of data collection
* Day number – the day number from the start of data collection

# **Learning Tasks**

Below are several kinds of learning tasks or predictions we may like to make about tweets, based on these datasets:

* Is a tweet informative or conversational?
* What is the informative value of a tweet (on a scale of 1-10)?
* Topical Detection, i.e. given a tweet we will label the tweet with a topic. Then can we uncover that topic up to synonyms?

# **Feature Generation**

Depending on the prediction we would like to make, the feature vector will be generated differently:

1. Informative vs. conversational and informative value on a scale of 1-10
   1. Check if a website is the text of the tweet
   2. User retweet count (i.e. higher retweet counts may indicate a more informative tweet)
   3. Use various text complexity measures
   4. Sentence diagramming structure
   5. location mentioned in the tweet
   6. date mentioned in the tweet
   7. Keywords
2. Topical Detection
   1. Hashtags
   2. Keywords

# **Models**

We will potentially use the following machine learning models:

* For informative vs conversational, we will use a Support Vector Machine, as it is a binary decision.
* To determine the informative value of a tweet, we will use Artificial Neural Networks.
* For our topical detection, we will use a decision tree model for classification.

# **Estimated Timeline**

* 1/27/14 – 2/2/14: Work on feature generation for predictions
* 2/3/14 – 2/11/14: Run diagnostics on models choices of feature vector generation