Proposal for the Course Project

(GMU Fall 2013, CS 780, Data Mining for Multimedia Data, Prof Jessica Lin)

Team Members

Talha Oz and Venkat Tadakamalla

Partly Sunny with a Chance of Hashtags

a kaggle Competition

http://www.kaggle.com/c/crowdflower-weather-twitter

1. Problem

We will be taking up an interesting problem involving a real world dataset (a micro blog data involving weather). We will be competing in the kaggle competition [3] organized by CrowdFlower [4], a crowdsourcing company, for sentiment analysis of tweets related to weather. The goal is to predict:

- whether a tweet has a positive, negative or neutral statement on the weather;
- whether if it talks about current, past or future;
- What kind of weather information is mentioned, wind, snow or some other.

To be more precise, given tweets and location information, confidence scores of 24 labels in three categories: *sentiment (5)*, *when (4)* and *kind (15)* are to be predicted.

Sentiment	When	Kind
I can't tell	current (same day) weather	clouds
Negative	future (forecast)	cold
Neutral / author is just sharing information	I can't tell	dry
Positive	past weather	hot
Tweet not related to weather condition		humid
		hurricane
		I can't tell
		ice
		other
		rain
		snow
		storms
		sun
		tornado
		wind

Table 1: 24 labels to predict their confidence scores

2. Data

Tweets are crowd-sourced by the customers of the competition organizer and each tweet is labeled by multiple graders with different confidence scores. The graders and their confidence scores are not revealed with data; instead their confidence scores are embedded into the training data. A grader can label a tweet with only one *sentiment* and one *when* category but allowed for multiple choices of *kind*. Below are two examples given from the training dataset.

```
"1","Jazz for a Rainy Afternoon: {link}","oklahoma","Oklahoma", //id, tweet and location
"0","0","1","0","0", //sentiment
"0.8","0","0.2","0", //when
"0","0","0","0","0","0","0","0","0","1","0","0","0","0","0" //kind
"30","Today was windy AF, I don't wanna go meet up w/ @mention and get my sis from school
:/","arizona","Phoenix (Prescott)",
"0","0.613","0.387","0","0", //sentiment
"0.793","0","0","0.207", //when
"0","0","0","0","0","0","0","0","0","0","0","0","0","0","0","1" //kind
```

Table 2: A few samples from the training dataset

Training dataset	Test dataset
77946	42157
17.1 MB	4.88 MB

Table 3: Number of records in the training and test datasets

3. Solution

We are not experienced in sentiment analysis. Hence, we will be reading publications on the topic. Based on our limited current exposure to the subject, we listed below some approaches to the problem; however, it is likely that we might follow a very different path after reviewing the literature.

We will probably be building different models for each of the three categories. Since the data is noisy we first do some preprocessing such as removing stop words and spell correcting (like [1]). Then as feature extraction we might do part-of-speech (POS) tagging to identify verbs, nouns and adjectives where we will later can check the tense of the verbs, compare noun with a hand curated bag of weather related words dictionary (mapping each word to a *kind*), and assess positivity of adjectives (like [2]). We might employ our Naïve Bayes and k-NN algorithms that we implemented in our second homework by using TF-IDF. To reduce the number of features (number of unique words in all of the tweets) we might put a threshold on TF-IDF values or alternatively try LDA or LSA for topic modeling. We will be tuning our NB by giving different weights to different features we created. To evaluate our success we will be doing 10-fold cross validation.

4. Evaluation Methods

In the evaluation section of the competition, organizers explained their evaluation metric as RMSE. So, we will also be using the Root Mean Squared Error (RMSE) to measure the accuracy of our predictions:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (p_i - a_i)^2}{n}}$$

Where:

- n is 24 times the total number of tweets
- p_i is the predicted confidence rating for a given label
- a_i is the actual confidence rating for a given label

References

[1] http://norvig.com/spell-correct.html
[2] Subjectivity Lexicon http://mpqa.cs.pitt.edu/
[3] http://www.kaggle.com/c/crowdflower-weather-twitter
[4] http://crowdflower.com/