What do you Tweet?
An Analysis of Twitter using Support Vector Machines

Richard Sung, Class of 2012
Advisor: Dr. Ralph A. Morelli

Abstract

In the past few years, Twitter has become a major social networking service with over 200 million tweets made every day. With this newfound source of expanding information, can people stay up to date with what others are posting? Along with the increasing processing power of computers, is there a way computing can analyze tweets on a large scale? Moreover, can computers understand what people think based on what they post? This senior project explores this question by determining the positive or negative sentiment of twitter posts by using a machine learning algorithm called *Support Vector Machines*.

Based on a labeled dataset of tweets, a parser then extracts present features in the text to create a vector. Once a collection of vectors is compiled, data is trained and tested to create a working model, which can then be evaluated to determine the effectiveness of the classifier. Based on a dataset of 359 tweets and 329 features, a model can accurately classify tweets as high as 74.84% using a linear classifier.

Methodology

Bag of Words Model
- Unigrams and Bigrams

Assumptions
- Presence-based features
- Emoticons stripped
- Words are spaced, punctuation omitted
- No twitter specific terms (hashtags, @user, links)

- Data Set CSV format

"Label", "ID", "Date", "Topic", "User", "Message"

"I like his personality… He has good character."

- Libsvm format

Label 0:0 1:0 2:0 3:1 4:0 5:1 6:0 7:0 8:0 [.....] 328:0

Training Set: 2/3 of data	Testing Set: 1/3 of data

Evaluation Set

Data Set and Feature Space

- Dataset consists of 359 tweets

Source: http://twittersentiment.appspot.com
- Vector contains 329 features

(152 positive, 175 negative)

Based on data from: http://twitrratr.com/
- Training/Testing sets created randomly
- SVM Linear Classifier

Positive Features	Negative Features
woo	ftl
quite amazing	irritating
thks	suck
looking forward	lying
kinda impressed	too slow
…	…
…	…
interesting	gross

Results

Dataset	Testing Accuracy	Evaluation set	Evaluation Accuracy	Highest Eval. Accuracy
20	57.5%	339	54.35%	59%
50	60%	309	55.44%	60.52%
100	57.65%	259	63.1%	66.41%
200	60.3%	159	68.25%	74.84%

Conclusions

- Making classifications can be done effectively using Support Vector Machines
- More vectors, better predictions and stronger model
- Many features also more beneficial

Reflections

- What makes Twitter different from other media such as articles and journals?
- Are words the only relevant feature?

References

Go, Alec, Bhayani, Richa, Huang, Lei. Twitter Sentiment Classification using Distant Supervision. http://twittersentiment.appspot.com/

Pak, Alexander and Paroubek, Patrick. Twitter as a Corpus for Sentiment Analysis and Opinion Mining

Pang, Bo and Lee, Lilian. Opinion Mining and Sentiment Analysis.

Pang, Bo and Lee, Lilian, Vaithyanathan, Shivakumar. Thumbs up? Sentiment Classification using Machine Learning Techniques

Libsvm software http://www.csie.ntu.edu.tw/~cjlin/libsvm/

http://blog.twitter.com/2011/08/your-world-more-connected.html