Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis

Stuart Colianni, Stephanie Rosales, Michael Signorotti
Stanford Center for Professional Development, Stanford University

Project Goals

The primary goal for this project is to explore how Twitter data can be utilized to develop advantageous crypto coin trading strategies. The objectives are as follows:

- Outline machine learning pipelines with the objective of identifying cryptocurrency market movement through supervised learning techniques.
- Apply supervised learning algorithms such as logistic regression, Naive Bayes, and support vector machines with unique and novel feature vectors.
- Achieve a test set accuracy greater than 80% for predicting when digital currency prices will go up or down

Data

Examples of raw tweets collected

2015-11-12 00:26:38.626000 bitcoin has no future, says ubs chairman axel weber at bank of england's open forum

2015-11-12 00:21:38.362000 uk chancellor sees a big future for digital currencies #bitcoin #bitcoinnews #btc #cryptocurrency

2015-11-12 00:26:45.489000 #bitcoin, sideways today, but then back down to just above $300 again.

Examples of prices per hour collected

2015-11-15 00:49:09.189000 332.65712695 {u'timestamp': 1447577341, u'ticker': {u'volume': u'41988.81016457', u'price': u'332.65712695', u'base': u'BTC', u'target': u'USD', u'change': u'-0.96702316'}, u'success': True, u'error': u''}

2015-11-15 01:49:21.136000 333.57270673 {u'timestamp': 1447580942, u'ticker': {u'volume': u'43053.76322028', u'price': u'333.57270673', u'base': u'BTC', u'target': u'USD', u'change':

Let's examine the frequency at which people posted tweets about Bitcoin.

Median:	1
Mean:	14
Standard Deviation:	140
Max:	20,000
Min:	1

Pre-processed tweets (before cleaning)

Median:	1
Mean:	8.2
Standard Deviation:	130
Max:	11,000
Min:	1

Processed tweets (after cleaning)

*Note: After 20 day collection, rounded to 2 significant figures

Method

Step 1: Collecting the Data
- Keywords "etherum", "eth", "bitcoin", "litecoin", "ltc", "xrp" are searched in Tweets
- Data is pulled via the Tweepy API
- Prices for cryptocurrency collected every hour via Cryptonator API
- All placed in separate text files

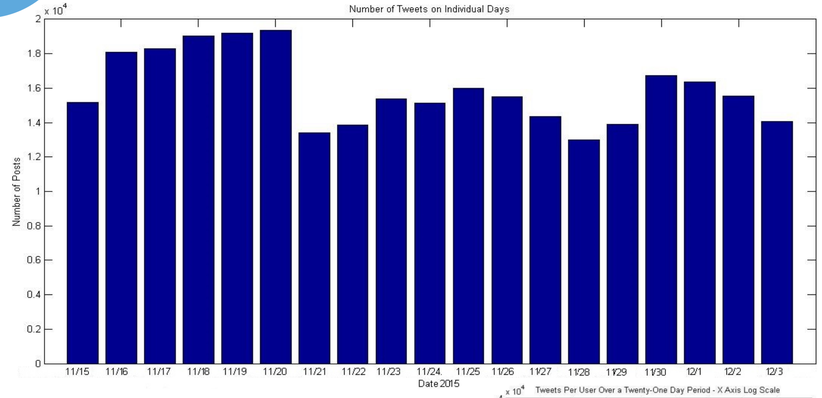
Step 2: Cleaning the Data
- Excess whitespace is removed from tweets
- Text is changed to lowercase
- Duplicate tweets, non-alphanumeric characters, and invalid words are removed
- Porter Stemming algorithm used for remaining words in tweets

Step 3: Creating Models

All are applied to hourly and daily models:
- Multinomial Naïve Bayes
- Bernoulli Naïve Bayes
- Logistic Regression
- SVM with linear kernel
- Sentiment analysis which labels the tweets with a positive and negative score

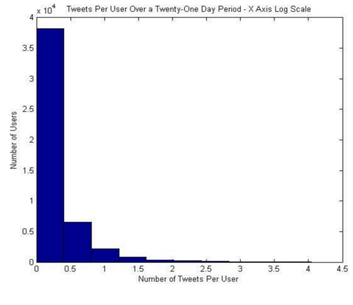
Step 4: Predicting Prices
- Determine whether to buy, sell, or keep
- Analyze accuracy of our prediction
- Perform error analysis and improve models

More Tweet Metrics

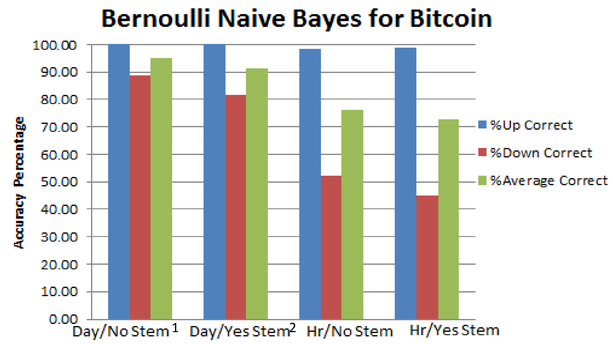


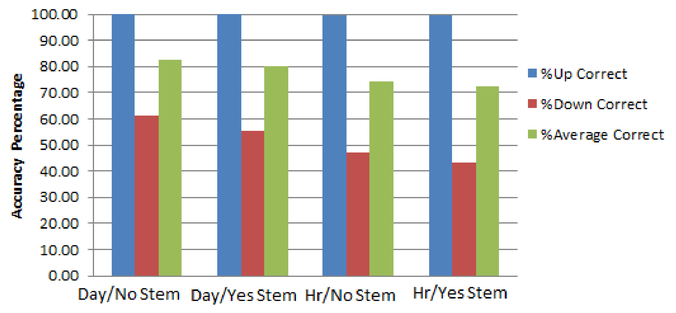
Number of tweets per day about Bitcoin.

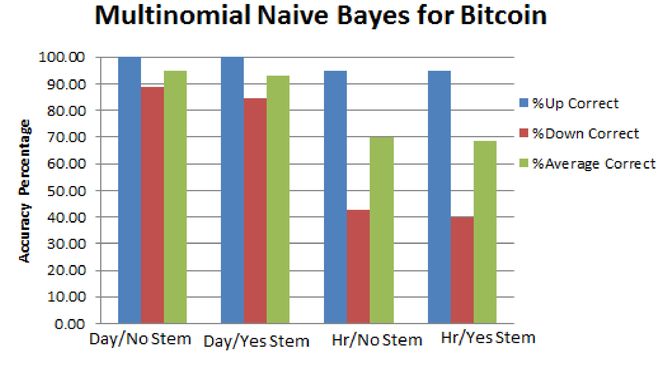
Median:	16,000
Mean:	16,000
Standard Deviation:	2,000
Max:	19,000
Min:	13,000



Results







1. No Stem is not utilizing Porter Stemming algorithm. 2. Yes Stem implements Porter Stemming algorithm.

Note on Sentiment Analysis:

Our team is utilizing the Text-Processing.com API to analyze tweets in order to formulate feature vectors for the learning algorithms. The feature vector will consist of floating point numbers which quantify the positivity, negativity, and neutrality of a particular tweet. This will be done for all tweets in the dataset, and the resulting feature vectors will be used to train and test classifiers such as logistic regression and support vector machines. We will also be incorporating the Profile of Mood States (POMS) to map tweets to mood states. These mood states will serve as separate features. Alternatively, the classifiers can also be trained on data for each mood state individually. The results can be compared to one another to determine which moods most closely predict market movement. Performance of the classifiers trained on the most promising mood states will be compared to the performance of previously used classification methods.