Stock Trend Analysis

FONG CHUN MAN (DENNIS)
ADVISED BY PROF. LEE DIK LUN

Overview

Language is a form of communication between humans. It can be transmitted as messages. Messages can be in forms of texts, audios, images and videos. Many people love to express their opinion towards financial topics in form of text. Twitter, a social networking company, has over 500 million active users who generate more than 500 million tweets per day and 5,700 tweets per second on average. It is impossible for someone to look at all tweets generated and make their decision on buying or selling stock. Our system is to analyze the tweets in order to give a prediction on the stock trend providing users with a reference.

Sentiment analysis, usually called opinion mining, is adopted to put those abundant and valuable tweets in good use. It is a technique that analyzes people's sentiments, opinions, evaluations, appraisals, attitudes and emotions towards entities. Collecting and analyzing users' tweets with the technique can easily and efficiently analyze the opinion of how people think or feel about a product or a brand.

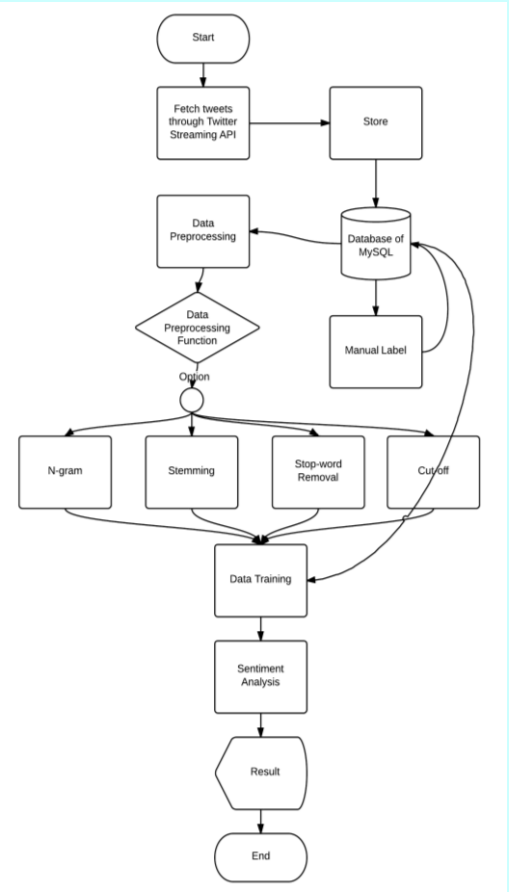
Objectives

The final year project aims to mine and acquire the sentiment from the tweets provided by Twitter to predict the stock price trend. The sentiment analysis is based on probability. Through using a probabilistic method, namely Naïve Bayes algorithm, we can analyze the polarity of tweets. The programme can analyze the information in near real time. Since many researches are focusing on how machine learning algorithms performance, little attention has been paid on how prepossessing methods affect the prediction. We try to create an inspirational research here. In order to evaluate the performance of the prediction, we use and compare n-gram, stop-word removal and stemming approaches based on the Naïve Bayes algorithm.

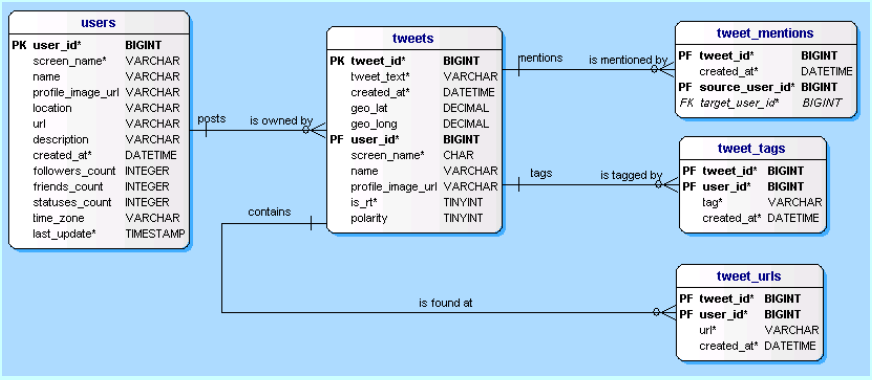
The goal of this project is to gather information from Twitter and analyze them. Users need not acquire professional financial knowledge. Our project mainly focuses on the following objectives:

1. Crawler: Develop a system that automatically and regularly collects tweets for stock trend prediction
2. Sentiment Analysis: Evaluate which preprocessing approach(es) can obtain a more accurate sentiment analysis result(s)

System Design

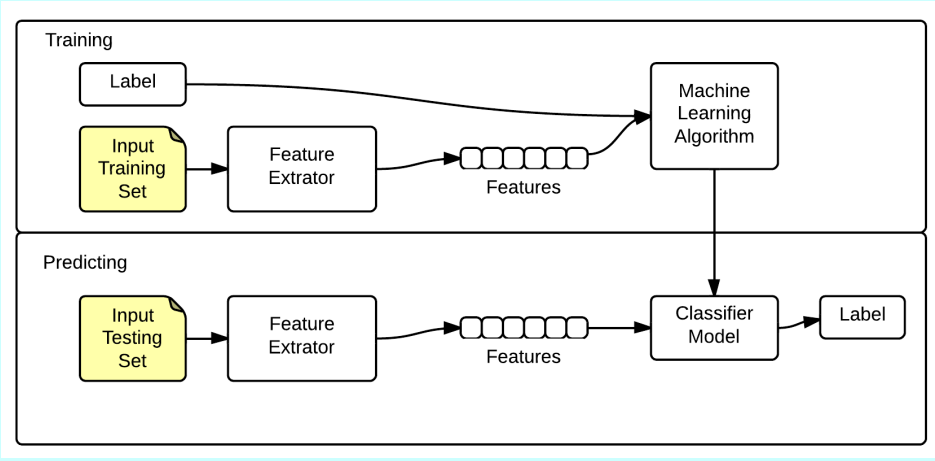


As illustrated in the figure at left hand side, the tweets are first fetched through Twitter Streaming API and stored in the database. Then some tweets' polarities are manually labeled for building a training dataset. After that, user can choose which data prepossessing method, like n-gram, stemming and stop-word removal. The prepossessed tweets are then training in the Naïve Bayes classifier. Other tweets collected in the database are then undergo sentiment analysis. ER Diagram is show below.



Methodology

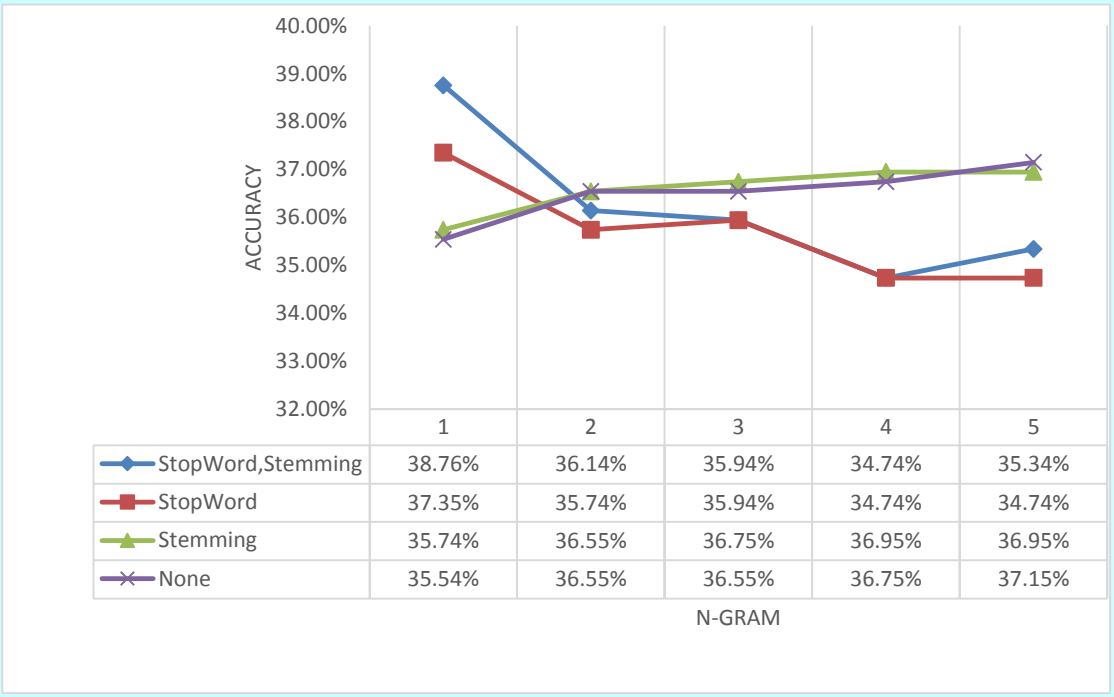
Extracting the sentiment from tweets is the main part of our project. The biggest challenges of the project are from this part because of the sentence complexity, variety of sentence structures and use of words. The project used data mining techniques to find out the polarity of the tweets. The training datasets are from the web, which are already classified by the data provider, and some of the training datasets are from our database, which tweets are crawled and manually classified by me.



1. Some tweets are labeled making up the training dataset.
2. After user has selected the prepossessing techniques, tweets are input into the programme.
3. Features of the tweets are extracted and inserted into the machine learning algorithm.
4. After training, the prediction is made. Tweets are served as input. They are then passed to the feature extractor, which the features are extracted as same as the training dataset.
5. The features are then processed by the classifier model. Finally, the tweets are labeled with polarity.

Results

As shown below, for public dataset, by using only stop-word removal or stop-word plus stemming, there is a general trend of decrease in accuracy when n-gram increases. However, if only stemming or none prepossessing method is used, the accuracy increases when n-gram increases. While for in-house dataset, the accuracies are all 80.4597701149%, which is larger than the average accuracy of public dataset. One of reasons of such a high accuracy is that all labeled data are closely related and much of the information is the same (same words and same words arrangement) since the data are fetched from streaming. For the reason that the results of the same accuracy, the data in test dataset is small.



	1	2	3	4	5
StopWord,Stemming	38.76%	36.14%	35.94%	34.74%	35.34%
StopWord	37.35%	35.74%	35.94%	34.74%	34.74%
Stemming	35.74%	36.55%	36.75%	36.95%	36.95%
None	35.54%	36.55%	36.55%	36.75%	37.15%

Accuracy Results (Public Dataset) of different combinations on prepossessing and n-grams

Conclusion

In this project, a report generating system is developed to evaluate accuracy of sentiment analysis. In general, if only stemming or none prepossessing method is used, the accuracy increases when n-gram increases. Even though there is a same accuracy for in-house dataset, the public dataset gives us an inspiring insight, which stemming is an important factor affecting the accuracy. It is suggested that stemming algorithms apart from Porter's stemming algorithm can be used to compare the accuracy. The result from in-house dataset tells us that it is possible to achieve more than 80% correctness for polarity prediction through Naïve Bayes classifier. Future work includes implement more machine learning algorithms to evaluate the results, implement more preprocessing methods, implement Semi-supervised machine learning, label for each company is needed (semantic labeling), build a system with GPUs to evaluate the performance in real-time trading, evolve to other social networks and add more languages support.