The Performance Evaluation of Machine Learning Classifiers on Financial Microblogging Platforms

Tianyou Hu
University of Auckland
Owen G Glenn Building, 12 Grafton
Road, Auckland, New Zealand
t.hu@auckland.ac.nz

Arvind Tripathi
University of Auckland
Owen G Glenn Building, 12 Grafton
Road, Auckland, New Zealand
a.tripathi@auckland.ac.nz

Abstract

As technological advancements facilitate democratization of knowledge, Microblogging platforms are vying to become the premier source of knowledge and are competing with news outlets. A huge number of messages is generated on different microblogging platforms. In financial markets, microblogging websites, such as StockTwits, have become a rich source for amateur investors, which make them ideal sources for market sentiment analysis. Indeed, StockTwit[1] has been widely used by researchers for sentiment analytics and market predictions. However, the quality of the sentiment analysis is highly dependent on the machine learning classifiers used as well as the preprocessing of data. In this study, we compare the performance efficiency of different machine learning classifiers on the user-generated content on StockTwits. We find that Logistic Regression Classifier performs best in a 2-way classification of StockTwits data. Our results report better classification accuracy than a similar research using data from Twitter. We have discussed managerial impications of our results.

Keywords: Social Media, Microblogging, Machine Learning, Sentiment Analysis, User-Generated Content, Stock Market

Introduction

Social media, especially microblogging services are becoming popular sources for information in almost all domains. For example, millions of Tweets are generated on Twitter everyday. Users create, share and discuss information on various topics, from personal life, and healthcare problems to societal issues and politics. Financial analysis and investment strategies, which used to be the limited to domain experts, is now provided by retail investors on social media (Chen, De, Hu, & Hwang, 2014). The quality of information available on social media platforms is comparable to expert opinions. In fact, many studies have established connections between sentiments on social media platforms and market returns (Oh & Sheng, 2011; Chen et al., 2014; Leung & Ton, 2015). Many studies have analyzed tweets from Twitter but since Twitter covers a very broad range of topics, it's difficult to filter and choose the right Tweets concentrating on the desired topic. We argue that domain specific microblogging platforms, such as StockTwits for stock market provide a better data source to study discussions and analyze market sentiments.

In recent years, using researchers have shown the effect of sentiments derived from microblogging platforms on stock markets (Bollen, Mao, & Zeng, 2011; Leung & Ton, 2015). Social media users use microblogging services to share their opinion about stock markets. This huge amount of data on microblogging platforms like StockTwits, is a treasure trove for market analysts and becomes a new market sentiment indicator and competes with the one based on traditional sources (newspaper, online news media or blogs written by experts). Furthermore, the short length of each message (maximum 140 characters per message) and the use of cashtags (an identifier like hashtag but starts with '$') make it a less noisy and easier to analyze. Furthermore, high frequency of content creation by users also allows analysts to track user behaviour at different level, in real-time, during trading.

Given the untrusted content, it's very challenging for an average person to process the huge amount of data and estimate market sentiments. These shortcomings can be addressed by using machine learning techniques. There has been increasing interest in stock market predictions using various machine learning techniques. Different machine learning algorithms have been used to classify messages into different sentiment groups. However, we are yet to understand classification efficiency of these algorithms for analysing messages from a microblogging platform. In this research, we compare the classification performance of different classifiers used for classifying posts on a microblogging platform StockTwits. .

Section 2 reviews the related literature on feature selection and sentiment analysis methods. Section 3 describes the data used in this research. Section 4 explains the machine learning classifiers used in this research. Section 5 presents the results. We conclude with discussion in section 6.

Literature Review

In literature, many approaches have been used to conduct sentiment analysis in social media.

Researchers have used various pre-defined dictionaries and machine learning classifiers to extract user sentiments from social media messages and articles in different context. To deal with this issue, Loughran and McDonald (2015) compared the four most widely used dictionaries, which are Henry (2008), Harvard's General Inquirer (GI), DICTION, and L&M (Loughran & Mcdonald, 2011). Each dictionary has its expertise, but the L&M is better than the rest three dictionaries in financial context for the following two reasons. First, the L&M dictionary does not miss common positive and negative words, which makes it more comprehensive than the rest. Second, the L&M dictionary was created for financial context analysis. It has been shown that L&M does really poor in short message classification in comparison with machine learning classifiers (Hu & Tripathi, 2015a). Thus, we will only compare machine-learning classifiers in this study.

Regarding the state of the art for machine learning classification in financial markets, Antweiler and Frank (2004) came up with a novel idea to compute bullishness index using computational linguistics method and showed that stock messages can predict market volatility. Bollen et al. (2011) measured collective mood state in term of two states (positive vs negative) and 6 dimension (Cal, Alert, Sure, Vital, Kind and Happy) from Twitter data using OpinionFinder and Google Profile of Mood States, and found an accuracy of 86.7% in predicting the directional changes in the closing price of Dow Jones Industrial Average. Sprenger, Tumasjan, Sandner, and Welpe (2014) collected Twitter messages containing cashtags of S&P 100 companies and classified each message using Naïve Bayes (NB) trained with a set of 2,500 tweets. Results demonstrated that bullishness index is correlated with the abnormal return and message volume is associated with trading volume. Oh and Sheng (2011) collected data from StockTwits for three months. The messages were classified by a bag of words approach which applied a machine learning algorithm J48 classifiers. They argued that the sentiments appear to have strong forecasting power over the future market directions. Tirunillai and Tellis (2012) collected data from consumer reviews and classified the reviews using NB and Support Vector Machine (SVM). Results showed that negative UGC has a significant negative effect on abnormal returns with a short "wear-in" and long "wear-out" effects, positive UGC has no significant effect on these metrics. Oliveira, Cortez, and Areal (2013) collected data from StockTwits for 605 trading days. Messages were counted as "bullish" if they contain the words "bullish", same logic was applied to messages containing "bearish" words. In contrast with previous studies, they found no evidence of return predictability using sentiment indicators, and of the information content of posting volume for forecasting volatility. Leung and Ton (2015) collected 2.5 million messages from Hotcopper (the biggest Australian stock discussion forum). The messages were classified using NB with a manually classified training set of 10,000 messages. They found that the number of board messages and message sentiment significantly and positively relate to the contemporaneous returns of underperforming (low ROE, EBIT margin, EPS) small capitalization stocks with high market growth potential.

The goal of this paper is to overcome the limitation of previous studies. Prior studies have used varied machine learning classifiers, but no comprehensive comparison has been made between different classifiers. Also, the nature of microblogging (short in length, use of slangs and typo errors) calls for sophisticated pre-processing before the messages could be fed to machine learning algorithms. Finally, many metadata from messages could be used to increase the performance of these algorithms.

Data

We have focused on top ten US stocks based on market capitalization: Apple (AAPL), Alphabet (GOOG, GOOGL), Microsoft (MSFT), Amazon (AMZN), Berkshire Hathaway (BRK.A, BRK.B), Exxon Mobil (XOM), Facebook (FB), Johnson & Johnson (JNJ), General Electric (GE), Wells Fargo (WFC). For each stock, we have collected messages posted on StockTwits from January 01, 2016 to June 31, 2016. We have randomly selected 20,000 tweets for this research.

StockTwits (http://stocktwits.com/) was selected as our data source for this study. StockTwits is a social media platform designed for sharing ideas between various stakeholders, such as, investors, traders and entrepreneurs, etc., and it is a popular platform, which had 230,000 active users in June 2013. Messages are limited to 140 characters but may contain links, charts or even video, similar to Twitter. However, in contrast to Twitter, StockTwits only focuses on the stock market and stock investment, which makes it a less noisy data source than other general microblogging services, such as Twitter. Each message contains at least one $cashtag (i.e., $AAPL, $AMZN, $GOOG). Since September 2012, users are able to disclose their sentiment for each message (post) as "Bullish" or "Bearish". Since this data contains self-disclosed sentiments, it can be used to test machine-learning algorithms, without manual classification.

Pre-processing of data

To remove noise from messages, We have pre-processed all the messages (Agarwal, et. al., 2011) as following: 1) replace all URLs with a tag ||U||, 2) replace all targets (e.g. "@Sam") and all cashtags (e.g. "$AAPL") with tag ||T|| 3) replace all negations (e.g. not, no, never, n't, cannot) with notation "NOT", and 4) replace a sequence of repeated characters by three characters, for instance, convert goooood to good.

Afterwards, we have processed the tweets using natural language processing tools: 1) use Stanford tokenizer (Klein & Manning, 2003) to tokenize the tweets. 2) use a port-of-speech tagger to process tokenized message and attach a part of speech tag to each word. 3) use the stopword list in Python NLTK to identify and remove stopwords from each message. 4) punctuations are also removed from messages. 5) Then we use WordNet (Miller & Fellbaum, 1998) to find English words. 6) get the stem of each word using Porter stemmer.

Prior polarity scoring

We based some of our features on the prior polarity of words (Agarwal et al., 2011). In this case, Dictionary of Affect in Language (DAL) is used and extended by WordNet. DAL contains about 8000 English words with a pleasantness score between 1 to -3 (negative to positive) for each word. We normalise the scores by dividing all the scores by 3. Words with polarity less than 0.5 are treated as negative, while words with polarity higher than 0.8 are treated as positive and the rest is treated as neutral. When a word is not found in the DAL dictionary, all synonyms are retrieved from WordNet. We then search for each of the synonyms in DAL. If any synonym is from DAL, the same pleasantness score of the original word in DAL is assigned to its synonym. If none of the synonyms appears in DAL, then the word is not linked with any prior polarity.

Features

Following Agarwal et al. (2011), the features that we use could be divided into four classes: first, a list of words from the training set, and the occurrence of these words for each tweets as Boolean values. Second, counts of primary features, which result in a natural number (∈ N). Third, features whose value is a real number (∈ R). Fourth, features whose values are Boolean (∈ B). Each of these general classes is further divided into two subclasses: Polar features VS Non-polar features. We classify a feature as polar if we find it prior polarity by searching DAL (extended by WordNet). All the other features, which do not have any prior polarity fall in the Non-polar category. Finally, Each of Polar and Non-Polar features are divided into two subclasses: POS and Other. POS is features which are parts-of-speech (POS) of words, with types of JJ (Adjective), RB (Adverb), VB (Verb), NN (Noun).

Same as Agarwal et al. (2011), row f_1 belongs to class Polar POS and is the count of the number of positive and negative POS in messages. f_2, f_3, f_4 all belongs to class Polar Other. f_2 is the number of negation words, and positive and negative prior polarity. f_3 is the number of (+/-) hashtags, capitalised words, and words with exclamation marks. f_4 belongs to Non-Polar POS and is the number of different part of speech tags. f_5, f_6 belong to Non-Polar Other. f_5 is other words without polarity; f6 is the number of hashtags, URLS, targets and cashtags. f_7 belongs to Polar POS and is the sum of prior polarity scores of words with POS of JJ, RB, VB, and NN. f_8 belongs to Polar Other and is the sum of prior polarity scores of all words. f_9 refers to class Non-Polar Other and is the percentage of tweets that is capitalised. Finally, f_{10} belongs to class Non-Polar Other and is the presence of exclamation and presence of capitalised words. The descriptions are shown in Table 1.

Table 1. Summary Statistics

N	Polar	POS	# of (+/-) POS (JJ, RB, VB, NN)	f_1
		Other	# of negation words, positive words, negative words	f_2
			# of (+/-) hashtags, capitalised words, exclamation words	f_3
	Non-Polar	POS	# of POS (JJ, RB, VB, NN)	f_4
		Other	# of words without prior polarity	f_5
			# of hashtags, URLs, targets, cashtags	f_6
P	Polar	POS	For POS, ∑ prior polarity score of words that POS	f_7
		Other	∑ prior polarity scores of all words	f_8
	Non-Polar	Other	Percentage of capitalised text	f_9
B	Non-Polar	Other	Exclamation, capitalised text	f_{10}

Machine Learning Models

In this research, we use three different classifiers: Naïve Bayes (NB), Logistic Regression (LR), and Support Vector Machine (SVM). We choose these three classifiers, as NB and SVM are two most widely used classifiers in the social media sentiment analytics in a financial context and LR is a good approach for 2-way classification (classify dataset into two groups), while has not been explored in comparison with other two classifiers in the social media sentiment analytics in a financial context. Each classifier is tested using a 10-fold cross-validation, which is a common practice with machine-learning classifiers. For Naïve Bayes, we use Multinomial NB and Bernoulli NB. For SVM, we use three different kernels, which are linear, poly, and rbf kernels.

Naïve Bayes

NB is based on Bayes' theorem with the naïve assumption of independence between every pair of features. Given *C* stands for a class and W_1 to W_n are the feature vector, Bayes' theorem states the following:

$$P(C|W_1, \dots, W_n) = \frac{P(C)P(W_1, \dots, W_n|C)}{P(W_1, \dots, W_n)} \tag{1}$$

The naïve assumption gives that:

$$P(W_1, \dots, W_n|C) = \prod_{i=1}^{n} P(W_i|C) \tag{2}$$

The relationship of Equation 1 is then simplified to:

$$P(C|W_1, \dots, W_n) = \frac{P(C)\prod_{i=1}^{n} P(W_i|C)}{P(W_1, \dots, W_n)} \tag{3}$$

As $P(W_1, \dots, W_n)$ is always a constant value given the input (W_1 to W_n), we can apply the following classification rule:

$$P(C|W_1, \dots, W_n) \propto P(C)\prod_{i=1}^{n} P(W_i|C) \tag{4}$$

Finally, the classification with the highest posterior probability is chosen.

$$\hat{C} = argmaxP(C)\prod_{i=1}^{n} P(W_i|C) \tag{5}$$

The main difference between NB classifiers is the assumptions that they make regarding the distribution of $P(W_i|C)$.

Multinomial NB uses the NB algorithm for multinomial distributed data. $P(W_i|C)$ is estimated by a smoothed version of maximum likelihood:

$$P(W_i|C) = \frac{N_{Ci} + \alpha}{N_C + \alpha n} \tag{6}$$

Where N_{Ci} is total number of times feature W_i falls in a sample of class *C* in the training set, and N_C is the total number of all features for class *C*. α is the smoothing parameter and prevent zero probabilities.

Bernoulli NB uses the NB classifier for multivariate Bernoulli distributed data. The decision rule for Bernoulli NB is based on:

$$P(W_i|C) = P(i|C)W_i + (1 - P(i|C))(1 - W_i) \tag{7}$$

Which penalised the non-occurrence of a feature *i* that is an indicator for class *C*.

Logistic Regression

The logistic function σ(t) is defined as follows:

$$\sigma(t) = \frac{1}{1+e^{-t}} \tag{8}$$

Let’s presume that t is a function of the independent variables $(W_1, \dots, W_n)$, where:

$$t = f(W_1, \dots, W_n) \tag{9}$$

And the logistic function could be written as:

$$F(W_1, \dots, W_n) = \frac{1}{1+e^{-f(W_1,\dots,W_n)}} \tag{10}$$

F(x) is described as the probability of the dependent variable *(C)* is a “Bullish” or “Bearish”.

Support Vector Machine

SVMs are a group of supervised learning algorithms widely used for classification. To have an overview of SVMs, SVMs provide a separation boundary (linear or non-linear) in the dataset. Let us consider a training set with n observations (x_i). Each of the observations is a p-dimensional vector of features. Each training set has a self-disclosed label (y_i) in this research. Then a hyperplane or a hypersurface is constructed that could separate the training dataset with respect to the labels. To balance the problem of over-fitting and under-fitting, a parameter is introduced into the model: penalty parameter *C* of the error term. The lower your *C* value, the smoother and more generalised your decision boundary is going to be. But if you have a large *C* value, the classifier will attempt to do whatever is in its power to perfectly separate each sample to correctly classify it.

Kernels methods enable SVMs to be functional in a higher dimensional, implicit feature space, without calculating the coordinates of data in that space, but rather by calculating the inner products between all pairs of data.

Measures

We measure the accuracy, precision, recall and F1 measures for all the classifiers.

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} (11) \quad Precision = \frac{tp}{tp+fp} \ (12)$$

$$Recall = \frac{tp}{tp+fn} (13) \qquad F1 = 2 * \frac{Precision * Recall}{Precision + Recall} (14)$$

where *tp* is true positive, *tn* is true negative, *fp* is false positive, and *fn* is false negative.

Results

We use 20,000 messages for this research. Each message has a self-disclosed sentiment, which is “Bullish” or “Bearish”. This provides good training sets as well testing sets for the classifiers. We do a 10-cross validations for this study. The original dataset is partitioned in 10 equal size subsamples. In the ten subsamples, a single subsample is used as the testing dataset, while the rest nine subsamples are used as the training set. Then we report the average accuracy, precision, recall, and F1 measure for all the experiments with different size of data. Figure 1 shows the learning curve for the 2-way classification. “MNB” is Multinomial NB, “BNB” is

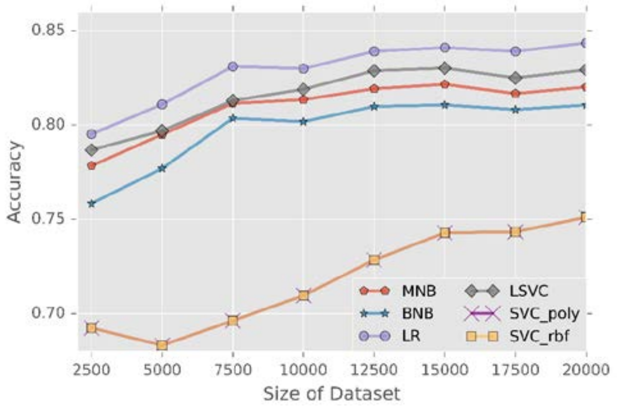

Figure 1. Learning Curve

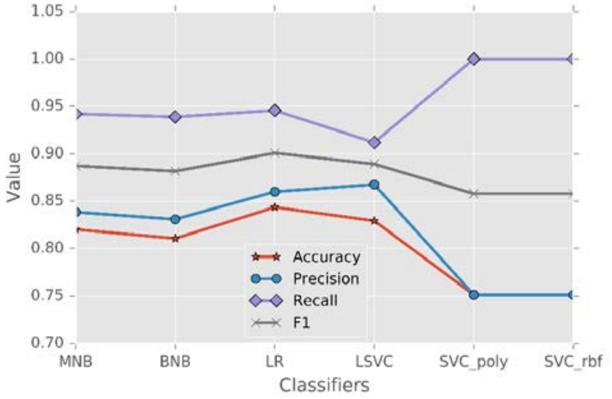

Figure 2. Accuracy, Precision, Recall and F1 Measure

Bernoulli NB, "LR" is Logistic Regression, "LSVC" is Linear SVM, "SVC_poly" is SVM using poly kernel, and "SVC_rbf" is SVM using radial basis function kernel.

It is clear that Logistic Regression Classifier outperforms all the other classifiers in this 2-way classification. However, Logistic Regression is not widely used in the classification of messages from social media in literature. In this case, we encourage researchers to use more classifiers and compare the accuracy of the classifiers, instead of only focusing on one or two classifiers with one kind of kernel. Figure 1 also shows that there is a quite sharp increase in accuracy when the size of dataset moves over 7,500. Thus, we encourage researchers to use a training set of more than 7500 to have a good accuracy in classification.

Overall accuracy, precision, recall and F-Measure are summarised in Figure 2. There is a trade-off between recall and precision, thus researchers have used F-Measure to determine which method is superior to others. Logistic Regression classifier has the highest value for accuracy (0.844) and F-Measure (0.901). This means that LR out-performs other classifiers in the social media sentiment analytics in a financial context. We also notice that SVMs with poly and rbf kernel have a recall of value 1 and the lowest precision among all the classifiers. This means that these two classifiers have no false negative classification and have a great amount of false positive classification, which makes these two classifiers have really poor performance.

Previous research on Twitter has used SVM to classify tweets from Twitter into two sentiment groups and got an accuracy of 75.39%. They streamed the data in real-time. No language, location or any other kind of restriction was made during the streaming process. Tweets in foreign languages are converted it into English using Google translate before the annotation process. They manually annotated 11,875 tweets. In comparison, our research comes up with an accuracy of 81.9% using Linear SVM, with a dataset of 10,000 tweets. Using almost the same method (unigram and metadata features), the accuracy for StockTwits outperform Twitter. The reasons could be: 1) StocksTwits is focusing on the financial market, which has less noise. 2) There is a great portion of users in StockTwits who are investors or traders. These people use more formal and accurate words than average users in Twitter. In this case, StockTwits is considered as a better data source to conduct sentiment analysis, especially in a financial context.

Conclusion

In this study, we achieve the following: First, we find that among the three classifiers, Logistic Regression performs the best in classifying messages on StockTwits. Though prior research studies analyzing financial microblogging services have been using NB or SVM, we report a superior performance of Logistic Regression in this environment. Second, we get a better accuracy using messages from StockTwits than from Twitter as a data source. When we want to find the correlation between social media sentiment and stock market variables, we want to include as many messages from social media platforms as we could. This gives rise to the need to classify all messages (with or without a self-disclosed sentiment) from a social media platform. Thus, we posit that StockTwits could be a better data source than Twitter to analyze sentiments in financial markets.

Reference

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media* (pp. 30–38). Association for Computational Linguistics.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of Internet stock message boards. *Journal of Finance*, *59*(3), 1259–1294. http://doi.org/10.1111/j.1540-6261.2004.00662.x

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. http://doi.org/10.1016/j.jocs.2010.12.007

Chen, H., De, P., Hu, Y. Y. (Jeffrey), & Hwang, B.-H. B.-H. (2014). Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *Rev. Financ. Stud.*, *27*, hhu001–. http://doi.org/10.1093/rfs/hhu001

Henry, E. (2008). Are Investors Influenced By How Earnings Press Releases Are Written? *Journal of Business Communication*, *45*(4), 363–407. http://doi.org/10.1177/0021943608319388

Hu, T., & Tripathi, A. K. (2015a). The Performance Evaluation of Textual Analysis Tools in Financial Markets. Retrieved from http://papers.ssrn.com/abstract=2661064

Hu, T., & Tripathi, A. K. (2015b). The Effect of Social Media on Market Liquidity. Proceedings of Thirty Sixth International Conference on Information Systems (ICIS), Fort Worth, Texas, USA.Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1* (pp. 423–430). Association for Computational Linguistics.

Leung, H., & Ton, T. (2015). The impact of internet stock message boards on cross-sectional returns of small-capitalization stocks. *Journal of Banking & Finance*, *55*(July 1998), 37–55. http://doi.org/10.1016/j.jbankfin.2015.01.009

Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65. http://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T., & McDonald, B. (2015). Textual Analysis in Finance and Accounting: A Survey. *SSRN Electronic Journal*. http://doi.org/10.2139/ssrn.2504147

Miller, G., & Fellbaum, C. (1998). Wordnet: An electronic lexical database. MIT Press Cambridge.

Oh, C., & Sheng, O. (2011). Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. Proceedings of Thirty Second International Conference on Information Systems (ICIS), Shanghai, China.

Oliveira, N., Cortez, P., & Areal, N. (2013). On the Predictability of Stock Market Behavior using StockTwits Sentiment and Posting Volume, 355–365.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, *20*(5), 926–957.

Tirunillai, S., & Tellis, G. J. (2012). Does Chatter Really Matter? Dynamics of User-Generated

Content and Stock Performance. *Marketing Science*, *31*(November 2014), 198–215. http://doi.org/10.1287/mksc.1110.0682