

Investment Recommendation using Investor Opinions in Social Media

Wenting Tu, David W. Cheung, Nikos Mamoulis, Min Yang, Ziyu Lu
Department of Computer Science, The University of Hong Kong
Pokfulam, Hong Kong
{wttu, dcheung, nikos, myang, zylu}@cs.hku.hk

ABSTRACT

Investor social media, such as StockTwist, are gaining increasing popularity. These sites allow users to post their investing opinions and suggestions in the form of microblogs. Given the growth of the posted data, a significant and challenging research problem is how to utilize the personal wisdom and different viewpoints in these opinions to help investment. Previous work aggregates sentiments related to stocks and generates buy or hold recommendations for stocks obtaining favorable votes while suggesting sell or short actions for stocks with negative votes. However, considering the fact that there always exist unreasonable or misleading posts, sentiment aggregation should be improved to be robust to noise. In this paper, we improve investment recommendation by modeling and using the quality of each investment opinion. To model the quality of an opinion, we use multiple categories of features generated from the author information, opinion content and the characteristics of stocks to which the opinion refers. Then, we discuss how to perform investment recommendation (including opinion recommendation and portfolio recommendation) with predicted qualities of investor opinions. Experimental results on real datasets demonstrate effectiveness of our work in recommending high-quality opinions and generating profitable investment decisions.

1. INTRODUCTION

In the past few years, social-media sites (e.g., Twitter) have emerged as one of prime platforms where people share their ideas and opinions. Some social media sites attract special categories of users, who exchange specialized information. For instance, StockTwits (www.stocktwits.com) is Twitter-like platform for users who share their investing and trading opinions.

Given the availability of platforms such as Twitter or Stocktwits, an interesting question is how to exploit the opinions of investors posted there in order to properly invest in stock markets. Previous work on the subject extracts [7] and aggregates [3, 8, 5, 4] the sentiments of users in the posted messages into a representative investment opinion about a stock. However, few of them explicitly consider the problem that there could be posts with misleading opinions either by non-experts or by malicious users. Our work

explicitly considers filtering out such noisy posts by evaluating the quality of each single investing opinion. Among previous work, only [1] noticed that different users post opinions of different quality and propose a framework to identify experts by considering the performance of users' past opinions. The main difference between our work and [1] is that [1] identifies the quality of users rather than that of specific opinions. However it is possible that different posts by the same user may have different quality (e.g., the user is expert only for a subset of the investment products). On the other hand in our work, we explicitly predict the quality of specific opinions and use additional information (e.g., content) for this purpose. Moreover, [1] only uses the opinions posted by authors in the past to derive the characteristics of authors, while we consider more features (e.g., social popularity) of authors.

Our work deals with the problem of predicting the quality of investing opinions and employs this knowledge to recommend to users high-quality opinions and portfolios to invest in. We first formulate the opinion quality prediction problem. Then, we propose multiple categories of features related to opinion author, opinion content and stocks that the opinion talks about. We also discuss how to construct a prediction function upon the proposed features using regression theory. After that, we show how to perform investment recommendation based on the predicted opinion qualities. Finally, we conduct experiments for evaluating the performance of investment recommendation using our methodology and competitor approaches to verify the effectiveness of our work.

2. PROBLEM DEFINITION

Consider a collection of opinions \mathcal{O} . For each opinion $o \in \mathcal{O}$, we denote by a^o the author of o , by c^o the content of o , and by \mathcal{E}^o the set of stocks o talks about. Moreover, we formulate a set \mathcal{V}^o of *investment views* in o , which consists of (e, l) pairs, where e is a stock discussed in o and l is the corresponding sentiment label: *bullish* (positive) or *bearish* (negative). For example, Figure 1 shows two investment opinions posted in Stocktwits. The investment views of the first opinion are $\{(GE, bearish)\}$ and those of the second opinion are $\{(BBUX, bullish)\}$.

Next we define the quality of an investing opinion o , denoted as $Q(o)$. In our work we regard o to be of high-quality if the investment views \mathcal{V}^o could help readers to perform profitable trading. Specifically, suppose \mathcal{E}_{bu}^o (\mathcal{E}_{be}^o) includes the stocks corresponding to bullish (bearish) sentiment labels in \mathcal{V}^o . If o is of high-quality, the prices of stocks in \mathcal{E}_{bu}^o (\mathcal{E}_{be}^o) should significantly rise (fall) in the future. Then, if a reader of o longs the stocks in \mathcal{E}_{bu}^o (i.e., buys the stocks in \mathcal{E}_{bu}^o and sells them in the future) and shorts the stocks in \mathcal{E}_{be}^o (i.e., borrows the stocks in \mathcal{E}_{be}^o to sell them and buys them back in the future), the reader could obtain high profit. According

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR '16, July 17–21, 2016, Pisa, Italy.
Copyright 2016 ACM ··· \$15.00.

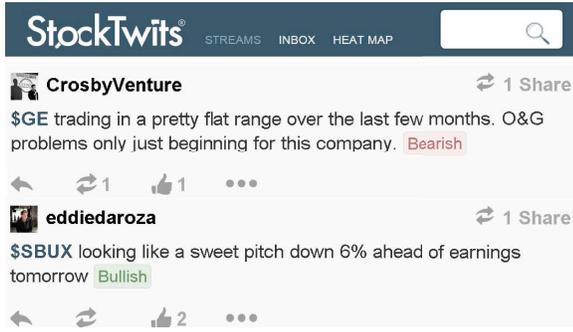


Figure 1: Stocktwits platform.

to the above discussion, $\mathcal{Q}(o)$ can be defined as

$$\frac{1}{|\mathcal{E}^o|} \left(\sum_{e \in \mathcal{E}_{bu}^o} \frac{p^f(e) - p^c(e)}{p^c(e)} + \sum_{e \in \mathcal{E}_{be}^o} \frac{p^c(e) - p^f(e)}{p^c(e)} \right), \quad (1)$$

where $p^c(e)$ and $p^f(e)$ are current price and future price of stock e (in this paper, we set as future price the price after a week). In this formula, $\frac{p^f(e) - p^c(e)}{p^c(e)} \left(\frac{p^c(e) - p^f(e)}{p^c(e)} \right)$ is the increase (decrease) rate of e 's price in the future, which actually equals to the return rate corresponding to the investment view (e, l) , where l is bullish (bearish).

The main goal of our work is to correctly predict the qualities of investment opinions and then use them for investment recommendation. We focus on two recommendation tasks. One is suggesting a set of high-quality investor opinions to users (i.e., *opinion recommendation*). The other is recommending a portfolio consisting of a set of stocks to long or short (i.e., *portfolio recommendation*).

3. QUALITY ESTIMATION FOR INVESTING OPINIONS

3.1 Features

We start by exploring how well different categories of features capture the quality of investing opinions. We consider features related to the opinion author, opinion content and the stocks the opinion talks about. Below we describe each of them in detail.

The author of an opinion is one of the main factors determining its quality; if the author is an expert in the domain, the opinions posted by him/her are more likely to be high quality. The most intuitive information reflecting author expertise is the performance of opinions posted by the author in the past. Suppose we are predicting the quality of opinion o . We first extract the opinions posted by a^o (i.e., the author of o) in the past. Then, we take the average quality of opinions written by a^o in the past as the first feature describing author expertise. This feature is denoted by $A_avgqual$. Moreover, we also take the percentage of high-quality opinions in all opinions posted by a^o in the past as the second feature. Specifically, we regard an opinion to be of high-quality if its quality value is larger than a threshold (e.g., 0.1). Then, the percentage of high-quality opinions in \mathcal{O}_u is used as another feature (denoted by $A_perchigh$) in our model. Besides measuring author expertise from his/her past opinions, we also explore additional information. In investor-based social media, a user typically *follows* other users if he/she values their opinions. Thus, we use social popularity of authors (denoted by A_social) as a third author expertise feature. We model A_social as the PageRank score [2] of the author (i.e., a commonly used measurement for evaluating social popularity of users). The fourth feature of author expertise is the number

of opinions the author posted in the past (denoted as A_onum); i.e., we assume that if a user actively posts opinions, he/she has higher expertise.

The content of an opinion is also potentially related to its quality since it may explain why the authors post their particular view. Sentiment strength in an opinion can also be extracted from the content. Thus, we believe content features may play an important role in predicting opinion quality. In our work, we consider word-level content features. Specifically, for a given opinion, we analyze the words it contains. For each word w , we calculate $w_avgqual$ and $w_perchigh$: the average quality of opinions containing w and the percentage of high-quality opinions in all opinions posted containing w , respectively (similar to $A_avgqual$ and $A_perchigh$). Let $v^a(w)$ and $v^p(w)$ be the $w_avgqual$ and $w_perchigh$ values of word w , respectively. Then, assuming that the content of opinion o consists of words $\{w_1, w_2, \dots, w_n\}$, we define two opinion content features $C_avgqual$ and $C_perchigh$, denoted by $\mathbb{E}(v^a)$ and $\mathbb{E}(v^p)$, as the mean of discrete values $\{v^a(w_i)\}_{i=1}^n$ and $\{v^p(w_i)\}_{i=1}^n$, respectively.

Besides features related to opinion author and content, we also define features for stocks. Similar to $C_avgqual$ and $C_perchigh$, for a given opinion, we examine the stocks it talks about. For each stock e mentioned in o , $e_avgqual$ and $e_perchigh$ are the average quality of opinions talking about e and the percentage of high-quality opinions in all opinions talking about e , respectively. Then, assuming that o mentions stocks $\{e_1, e_2, \dots, e_n\}$, we employ the mean of $e_avgqual$ and $e_perchigh$ values of $\{e_i\}_{i=1}^n$ as two features (denoted as $S_avgqual$ and $S_perchigh$) to describe the stocks mentioned in o .

3.2 Regression-based Quality Prediction

For a new opinion o_i , we extract the values of features $A_avgqual$, $A_prechigh$, A_social , A_onum , $C_avgqual$, $C_prechigh$, $S_avgqual$, $S_prechigh$, as defined in Section 3.1. Suppose the features of o_i are $[x_1^i, x_2^i, \dots, x_n^i]$. The predicted quality of o_i (denoted as $\hat{\mathcal{Q}}(o_i)$) can be obtained by a function f (i.e., $\hat{\mathcal{Q}}(o_i) = f([x_1^i, x_2^i, \dots, x_n^i])$). We can choose regression models [9] for defining f . For this purpose, we can use as training set all $\{o_j, \hat{\mathcal{Q}}(o_j)\}$ pairs, where o_j is an opinion posted in the past and $\hat{\mathcal{Q}}(o_j)$ is the real quality of o_j . After f is trained, we can use it to estimate the quality of o_i .

4. INVESTMENT RECOMMENDATION WITH PREDICTED OPINION QUALITIES

In our work, we utilize the predicted opinion qualities for two investment-recommendation tasks: recommending high-quality opinions and recommending portfolios. The high-quality opinions are selected as those with the highest predicted qualities. For generating portfolios, we use the predicted qualities of opinions to weigh them and then aggregate the sentiment about stocks in the opinions to generate scores for the stocks. This way, we can create daily investment recommendations based on the posted opinions.

4.1 Quality Sensitive Sentiment Aggregation for Recommending Portfolios

In previous work on constructing portfolios based on investor opinions, stocks are selected according to their aggregate sentiment indexes. One of the most popular aggregate bullish and bearish sentiment indexes [3, 8, 5] are:

$$\delta_i^{bu} = \ln \left[\frac{1 + |\mathcal{O}_i^{bu}|}{1 + |\mathcal{O}_i^{be}|} \right], \quad \delta_i^{be} = \ln \left[\frac{1 + |\mathcal{O}_i^{be}|}{1 + |\mathcal{O}_i^{bu}|} \right], \quad (2)$$

where δ_i^{bu} (δ_i^{be}) denotes bullish (bearish) sentiment index for stock

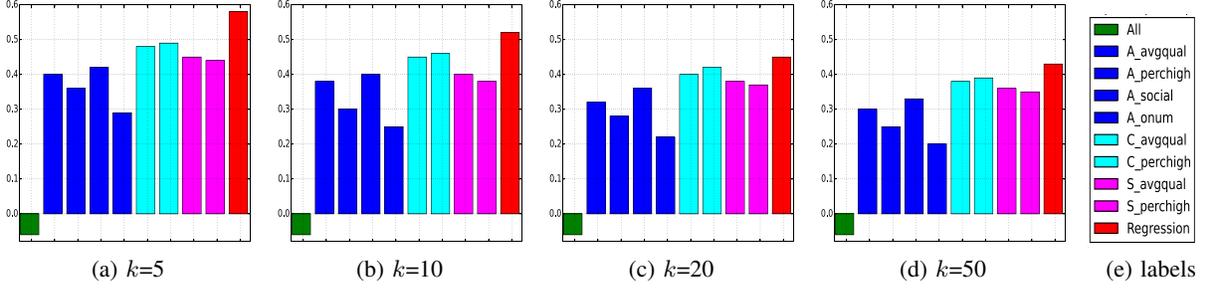


Figure 2: Performance comparison of recommended opinions. From left to right in order, the bars correspond to the average quality of all opinions in \mathcal{O}_d (*All*), the average quality of \mathcal{O}_d^{rec} by feature *A_avgqual*, *A_prechigh*, *A_social*, *A_onum*, *C_avgqual*, *C_prechigh*, *S_avgqual*, *S_prechigh*, and the average quality of \mathcal{O}_d^{rec} by our regression model based on multiple features (*Regression*). The y-axis shows the average value of recommended opinions’ performances (i.e., equation 4)

e_i while \mathcal{O}_i^{be} (\mathcal{O}_i^{be}) is the set of opinions containing bullish (bearish) views on e_i . In other words, for each o in \mathcal{O}_i^{bu} (\mathcal{O}_i^{be}), the investment views \mathcal{V}^o should contain (e_i, l_i) and l_i is bullish (bearish). Obviously, if δ_i^{bu} is larger than δ_i^{be} (i.e., $|\mathcal{O}_i^{bu}| > |\mathcal{O}_i^{be}|$), we should consider to long e_i while if the δ_i^{be} is larger than δ_i^{bu} (i.e., $|\mathcal{O}_i^{be}| > |\mathcal{O}_i^{bu}|$), we should consider to short e_i . Moreover, if the difference between δ_i^{bu} and δ_i^{be} (i.e., $|\delta_i^{bu} - \delta_i^{be}|$) is larger, the views are more consistent and the rising (if $\delta_i^{bu} > \delta_i^{be}$) or falling (if $\delta_i^{be} > \delta_i^{bu}$) probability should also be large. Thus, for generating a portfolio, we could sort stocks in decreasing order of $|\delta_i^{bu} - \delta_i^{be}|$ and take the top k' ones as the portfolio suggestion. Finally, we long (short) the stocks if their bullish indexes are larger (smaller) than their bearish indexes.

In our work, we attempt to utilize the predicted qualities in portfolio recommendation by giving higher weights to views in high-quality opinions when we aggregate sentiment labels in opinions’ investment views. We call this *quality-sensitive* portfolio recommendation. Note that in Equation (2), $|\mathcal{O}_i^{bu}|$ actually equals $\sum_{o \in \mathcal{O}_i^{bu}} 1$. Thus, by weighting views in o with $\hat{Q}(o)$, the quality-sensitive aggregation for sentiment indexes should be

$$\delta_i^{bu} = \ln \left[\frac{1 + \sum_{o \in \mathcal{O}_i^{bu}} \hat{Q}(o)}{1 + \sum_{o \in \mathcal{O}_i^{be}} \hat{Q}(o)} \right], \delta_i^{be} = \ln \left[\frac{1 + \sum_{o \in \mathcal{O}_i^{be}} \hat{Q}(o)}{1 + \sum_{o \in \mathcal{O}_i^{bu}} \hat{Q}(o)} \right]. \quad (3)$$

4.2 Daily Investment Recommendation

Now, we discuss in detail how to perform opinion recommendation and portfolio recommendation on each tradable day. Thus, our framework recommends a set of high-quality opinions before the stock market opening time, in order for these opinions to help users determine their trading strategy. In addition, we recommend a set of stocks to users to trade. As a test case, in our work, we focus on the stocks traded at New York Stock Exchange (NYSE) and Nasdaq stock markets. The opening time of the stock markets is 9:30 am (Eastern Standard Time). We set the time to generate recommendations at 8:00 am. The recommendation procedure for k opinions and k' -stock portfolios is summarized by Algorithm 1 (assuming the recommendation is performed on day d).

5. EXPERIMENTS

5.1 Data

For experiments, we collect all messages posted in 2014 from the investor-based social media StockTwits. StockTwits users post

Algorithm 1 Daily Investment Recommendation

Input:

\mathcal{E}_d = tradable stocks on day d
 \mathcal{O}_d = opinions mentioning stocks in \mathcal{E}_d and posted during from 8am on day $d - 1$ to 8:00 am on day d

Output:

- k recommended opinions \mathcal{O}_d^{rec}
 - a portfolio $\mathcal{E}_d^{rec} \subseteq \mathcal{E}_d$ consisting of k stocks
 - 1: **for all** $o_i \in \mathcal{O}_d$ **do**
 - 2: calculate the features $[x_1^i, x_2^i, \dots, x_n^i]$
 - 3: estimate predicted quality $\hat{Q}(o_i) = f([x_1^i, x_2^i, \dots, x_n^i])$
 - 4: **end for**
 - 5: $\mathcal{O}_d^{rec} = k$ opinions in \mathcal{O}_d according to k highest $\hat{Q}(o_i)$ values
 - 6: **for all** $e_i \in \mathcal{E}_d$ **do**
 - 7: calculate sentiment indexes δ_i^{bu} and δ_i^{be}
 - 8: calculate stock score $s(e_i) = |\delta_i^{bu} - \delta_i^{be}|$
 - 9: **end for**
 - 10: $\mathcal{E}_d^{rec} = k'$ stocks in \mathcal{E}_d according to k' highest stock scores
-

short messages (limited to 140 characters) that include ideas or opinions on specific investments. Stock symbols in messages are preceded by a ‘‘CashTag’’ (\$) (as Figure 1 shows). Here, we only use of StockTwits messages have at least one CashTag. Finally, the dataset we used contains 1,630,914 messages posted by 29,804 users and related to 6,541 stocks traded at NYSE and Nasdaq. Finally, we used the Yahoo! Finance API to crawl historical prices of stocks, used in Equation (1) and for evaluating effectiveness.

5.2 Experiments on Recommending Opinions

To generate \mathcal{O}_d^{rec} , we first predict qualities in \mathcal{E}_d . Specifically, we train a linear regression model [6] with {opinion, quality} pairs corresponding to the opinions posted on other days. After the regression model is trained and the qualities of opinions in \mathcal{E}_d are predicted, \mathcal{O}_d^{rec} is set as the opinions with the top k predicted qualities.

For each \mathcal{O}_d^{rec} , we evaluate its performance as the average quality of opinions in it:

$$\frac{1}{|\mathcal{O}_d^{rec}|} \sum_{o \in \mathcal{O}_d^{rec}} \mathcal{Q}(o). \quad (4)$$

Figure 2 presents the average performance of all \mathcal{O}_d^{rec} s generated on all tradable days in 2014. Besides the results corresponding to our model (multiple features + regression, denoted by *Regression*, we also show the average quality of all opinions in \mathcal{E}_d (to see if the average quality of \mathcal{O}_d^{rec} is larger than it) and the results of using

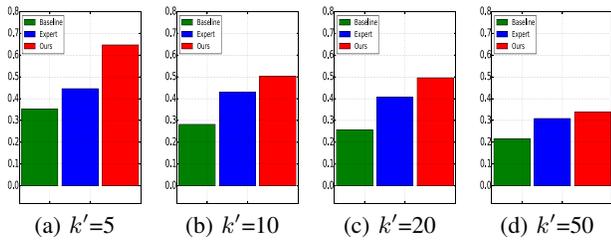


Figure 3: Comparison of recommended portfolios obtained by *Baseline* and *Expert* and our method (*Ours*). The y-axis shows the average value of recommended portfolios' performances (i.e., equation 5).

each of the features alone. To test each feature alone, we generate \mathcal{O}_d^{rec} as opinions corresponding to k highest values based on each single feature. As Figure 2 shows, each of the proposed features could be used retrieve high-quality opinions to a different extent, since the average quality of opinions in \mathcal{O}_d^{rec} corresponding to all features are larger than the average quality of all opinions. The best performance is achieved by our model (multiple features + regression).

5.3 Experiments on Recommending Portfolios

To generate \mathcal{E}_d^{rec} , we use conventional sentiment indexes (denoted as *Baseline*) used in previous work [3, 8, 5] (see Equation (2)) and our (quality-sensitive) sentiment indexes (denoted as *Ours*) corresponding to Equation (3) to construct portfolios (see Section 4.1 and Algorithm 1 for details). Moreover, as another baseline, we employ the author expertise model proposed in [1] to generate expert-sensitive sentiment indexes (denoted as *Expert*, see [1] for details).

For evaluating \mathcal{E}_d^{rec} , we calculate the return rate obtained if we long the stocks for which opinions in \mathcal{E}_d^{rec} give bullish recommendations and short the stocks for which opinions in \mathcal{E}_d^{rec} give bearish suggestions. Specifically, let \mathcal{E}_{dbu}^{rec} (\mathcal{E}_{dbe}^{rec}) be the stocks that attract bullish (bearish) ballots from opinions in \mathcal{E}_d^{rec} . The performance of \mathcal{E}_d^{rec} can be evaluated as

$$\frac{1}{|\mathcal{E}_d^{rec}|} \left(\sum_{e \in \mathcal{E}_{dbu}^{rec}} \frac{p^f(e) - p^c(e)}{p^c(e)} + \sum_{e \in \mathcal{E}_{dbe}^{rec}} \frac{p^c(e) - p^f(e)}{p^c(e)} \right), \quad (5)$$

Figure 3 shows the average performance of all \mathcal{E}_d^{rec} s generated on all tradable days in 2014. We compare the performances corresponding to our method and other competitors (i.e., *Baseline* and *Expert*). The results show that our quality-sensitive method suggests the most profitable portfolios.

Besides, we also study the cumulative return obtained if we just trade stocks in portfolios recommended by our method and perform continuous trading from the beginning to the end of 2014. Specifically, from the first tradable day in 2014, we repeat the following procedure: Suppose money for trading is available on day d . Firstly, we long stocks in \mathcal{E}_{dbu}^{rec} and short stocks in \mathcal{E}_{dbe}^{rec} for a week. After a week passes since day d , we sell stocks in \mathcal{E}_{dbu}^{rec} and buy back the borrowed stocks in \mathcal{E}_{dbe}^{rec} . Thus, money for trading will be available again on day $d+7$. In the next iteration of trading, d becomes $d+7$ and we perform trading again according to \mathcal{E}_{d+7}^{rec} . Figure 4 shows the cumulative gain after 39 such trading sessions. Here, we also include the cumulative return obtained if we trade on Dow Jones Industrial Average (DJIA) index, which simulates a portfolio of 30 significant stocks traded at NYSE and Nasdaq. We can see that, the return rates obtained by our portfolios are much higher than those

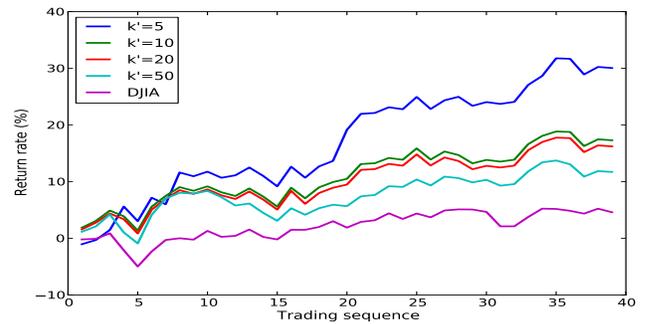


Figure 4: Cumulative return with recommended portfolios (in 2014).

by the DJIA index, proving that our portfolios consisted from s -stocks performing better than average. Also note that the return rate increases and that smaller values of k' give higher return. Specifically, for $k' = 5$, at the end of 2014, we could obtain a revenue of around 30%.

6. CONCLUSIONS

In this paper, we studied the problem of analyzing the qualities of investor opinions posted on social media. Moreover, we explore the use of predicted opinion qualities for opinion recommendation and portfolio recommendation. Experiments on a real-world dataset verify the effectiveness of our work in recommending high-quality investment opinions and profitable portfolios.

7. REFERENCES

- [1] R. Bar-Haim, E. Dinur, R. Feldman, M. Fresko, and G. Goldstein. Identifying and following expert investors in stock microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011.
- [2] J. Heidemann, M. Klier, and F. Probst. Identifying key users in online social networks: A pagerank based approach. In *Proceedings of the International Conference on Information Systems (ICIS)*, 2010.
- [3] C. Oh and O. Sheng. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. In *Proceedings of the International Conference on Information Systems (ICIS)*, 2011.
- [4] N. Oliveira, P. Cortez, and N. Areal. On the predictability of stock market behavior using stocktwits sentiment and posting volume. In *Progress in Artificial Intelligence*, pages 355–365. Springer, 2013.
- [5] T. Rao and S. Srivastava. Analyzing stock market movements using twitter sentiment analysis. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2012.
- [6] G. A. Seber and A. J. Lee. *Linear regression analysis*, volume 936. John Wiley & Sons, 2012.
- [7] W. Tu, D. Cheung, and N. Mamoulis. Improving microblog retrieval from exterior corpus by automatically constructing microblogging corpus. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2015.
- [8] W. Tu, D. Cheung, and N. Mamoulis. Time-sensitive opinion mining for prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2015.
- [9] V. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.