Information Systems 78 (2018) 189–198

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

Investment recommendation by discovering high-quality opinions in investor based social networks

Wenting Tu[a], Min Yang[b,*], David W. Cheung[c], Nikos Mamoulis[c]

[a] School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai, China
[b] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[c] Department of Computer Science, The University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 15 April 2017
Revised 15 February 2018
Accepted 24 February 2018
Available online 24 February 2018

Keywords:
Investment recommendation
Investor based social network
Data mining

ABSTRACT

Investor based social networks, such as StockTwist, are gaining increasing popularity. These sites allow users to post their investment opinions 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. A typical way is to aggregate 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 our work, we study how to estimate qualities of investment opinions in investor based social networks. To predict the quality of an investment opinion, we use multiple categories of factors generated from the author information, opinion content and the characteristics of stocks to which the opinion refers. With predicted qualities of investment opinions, we perform two types of investment recommendation. The first is recommending high-quality opinions to users and the second is recommending portfolios generated by sentiment aggregation in a quality-sensitive manner. Experimental results on real datasets demonstrate the effectiveness of our work in recommending high-quality investment opinions and profitable portfolios.

1. Introduction

In recent years, social networks (e.g., Twitter) have become one of prime places where web users present their ideas and opinions. In addition, there have been lots of social networks attracting special groups of users, who share specialized opinions. As an example, StockTwits[1] is an **I**nvestor **B**ased **S**ocial **N**etwork (**IBSN**) where investors post their investment views and investment opinions.

With the availability of IBSNs, we could collect a large volume of investment opinions posted by real investors. A challenging task is making use of opinions from massive investors to help users invest. Most of previous work [1–4] utilize the idea of "wisdom of crowd" that aggregates the investment views of lots of users into a single investing decision. Specifically, after extracting sentiment (bullish or bearish) in each investment opinion, previous work integrates sentiments in all opinions about a stock into an investing decision (e.g., long or short) on it. For example, if the ratio of bullish votes compared to bearish votes are larger than a threshold or in the top list, we will buy or sell the related stock.

However, the effectiveness of sentiment aggregation is likely to be influenced by the low-quality opinions. IBSNs are open to nearly all web users. Thus, there may be many low-quality opinions posted by non-experts or by malicious users. Sentiments from low-quality investment opinions will reduce the performance of sentiment aggregation. Unfortunately, there is limited previous work that explicitly considers this problem.

For reducing the negative influence of noisy investment opinions, our work models the qualities of investment opinions posted on ISBNs. In previous work, only Bar-Haim et al. [5] has a similar goal to ours. Bar-Haim et al. [5] noticed that different users post opinions of different quality and propose a framework to identify experts by considering the performance of users' past opinions. Compared to Bar-Haim et al. [5], our work infers the quality of specific opinions rather than users. For estimating qualities of investment opinions, we not only consider author's past performance but also explore other information (e.g., the social popularity of opinion author, words the author used or which stock the author talks about). Moreover, our work not only studies gen-

* Corresponding author.
E-mail address: minyang.ai@gmail.com (M. Yang).
[1] www.stocktwits.com .

https://doi.org/10.1016/j.is.2018.02.011
0306-4379/

erating portfolios from investment opinions in different qualities but also discusses opinion recommendation, which aims to recommend high-quality investment opinions to users. Compared with a single trading strategy (e.g., a portfolio) generated by an algorithm, investment opinions are more explainable since they include more information (e.g., the reasons why opinion authors suggest to long a stock) for users as references.

In the rest of this paper, we define some notations and formulate the quality of an investment opinion in Section 3. After that, we describe our methodology for predicting qualities of investment opinions and recommending investment advice in Section 4. Experiments are showed in Section 5 and related work is given in Section 6. Finally, we conclude and give directions for future work in Section 7.

2. Summary of contributions

In summary, the contributions of this paper are as follows:

- We bring in the idea of predicting qualities of investment opinions for improving investment recommendation. We formulate the qualities of investment opinions, accordingly.
- We study several factors (related to opinion author, words used in the opinion and which stocks the opinion talks about) for predicting opinion qualities. Hypotheses upon these factors are also presented.
- We build a prediction model for estimating qualities of investment opinions. With the prediction model, we are able to estimate the qualities of new investment opinions.
- We study how to perform investment recommendation (including opinion recommendation and portfolio recommendation) with predicted qualities of investment opinions.
- We conduct an experimental evaluation based on real data that evaluates our methodology in terms of the ability to recommend high-quality investment opinions and profitable portfolios.

This article is a substantial extension of Tu et al. [6], which was presented at SIGIR'16 as a short paper. The differences between the conference version and this manuscript are listed as follows.

- In this paper, we explicitly present hypotheses upon the factors used for predicting opinion qualities (in Section 4.1).
- In [6], we formulated estimating opinion qualities as a problem of linear regression. In this article, we extend it into a problem of non-negativity least squares. The comparison between these two methods is conducted and the result demonstrates the improvement caused by our extension (see Section 5.5 for details).
- In the experiment, we use a larger Stocktwits dataset. Moreover, more details about the experimental setup and evaluation metrics are given. Particularly, we add hypothesis testing (in Section 5.4) for understanding the effectiveness of the factors more comprehensively.

3. Qualities of investing opinions

Our work deals with the problem of estimating qualities of investment opinions. In this section, we formulate the quality of an investment opinion.

3.1. Notations

We first define some notations. We use o to denote an investment opinion containing investment views of the opinion author. For each investment opinion o, we denote by a^o the author of o, and by $\mathcal{E}^o$ the set of stocks a^o talks about in o. 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 (i.e., $e \in \mathcal{E}^o$) and l is the corresponding sentiment label: *bullish* (positive) or *bearish* (negative). For example, Fig. 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)}. Note that some platforms allow users to express their sentiments in an explicit manner. For example, users in the stocktwits platform could mark their opinions as bullish or bearish (as Fig. 1 shows). However, if investment opinions are extracted from other platforms such as twitter or sinaweibo, that are not specifical for investors, we may need to employ sentiment analysis approaches to extract authors' sentiments.

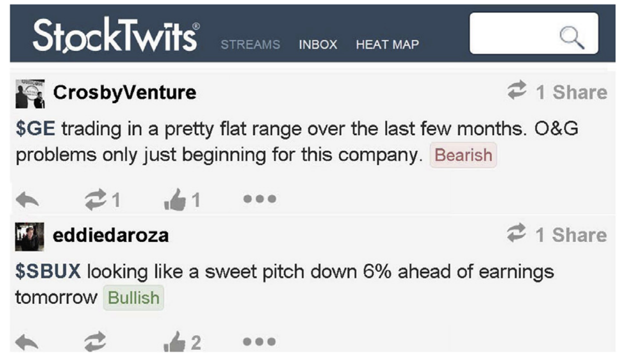


Fig. 1. Two investment opinions in the Stocktwits platform.

3.2. Ground truth of investment-opinion quality

Next, we define the quality of an investment opinion o, denoted as $\mathcal{Q}(o)$. According to the purpose of bringing users high profits, we formulate the *quality of o* (denoted by $\mathcal{Q}(o)$) as **how much profit a user could earn by trading according to the investment views in *o***. Given $\mathcal{V}^o$ of an investment opinion o, we denote $\mathcal{E}^o_{bu}$ ($\mathcal{E}^o_{be}$) as the stocks corresponding to bullish (bearish) sentiment labels in $\mathcal{V}^o$. If the user trade according to $\mathcal{V}^o$, he/she should *long* the stocks in $\mathcal{E}^o_{bu}$ (i.e., buy the stocks in $\mathcal{E}^o_{bu}$ and sell them in the future) and *short* the stocks in $\mathcal{E}^o_{be}$ (i.e., borrow the stocks in $\mathcal{E}_{be}$ to sell them and buy them back in the future). Then, for an investment opinion published on day d, the profit the user will earn by trading according to o is

$$\frac{1}{\mathcal{N}(\mathcal{E}^o)}\left(\sum_{e\in\mathcal{E}^o_{bu}}\frac{p^d_\triangle(e)-p^d(e)}{p^d(e)}+\sum_{e\in\mathcal{E}^o_{be}}\frac{p^d(e)-p^d_\triangle(e)}{p^d(e)}\right), \tag{1}$$

where $p^d(e)$ and $p^d_\triangle(e)$ are current price (i.e., price on day d) and future price of stock e (i.e., price on day $d+\triangle$). In our work, we set as the future price the price after a week (i.e., $\triangle = 7$). Note that, in some stock markets (e.g., Chinese stock market), people are not allowed to short stocks. Under this circumstance, the profit is calculated as:

$$\frac{1}{\mathcal{N}(\mathcal{E}^o_{bu})}\sum_{e\in\mathcal{E}^o_{bu}}\frac{p^d_\triangle(e)-p^d(e)}{p^d(e)}. \tag{2}$$

According to our definition of investment opinions' qualities (i.e., *how much profit a user could earn by trading according to the investment views in the opinion*), the formula (1)[2] can be used for quantifying the real opinion quality $\mathcal{Q}(o)$.

[2] Or formula (2), if people are not allowed to short stocks.

4. Proposed methodology

The main goal of our work is to correctly predict the qualities of investment opinions and then use them for investment recommendation. We concentrate on two recommendation tasks. One is suggesting a set of high-quality investment 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*). In this section, we start by proposing several related factors and corresponding hypotheses. Then, we discuss how to predict opinion qualities by combining the effectiveness of the factors. Finally, details about performing opinion recommendation and portfolio recommendation with predicted opinion qualities are given.

4.1. Quality related hypotheses and factors

This part explores how well different categories of factors capture the quality of investment opinions. We consider factors related to the expertise of opinion author (author level), words used in the opinion (content level) and stocks the opinion talks about (stock level). Below we describe each of them in detail.

4.1.1. Author level

One of the most noticeable factors determining the quality of an opinion is the expertise of the author. An intuitive factor measuring author expertise is whether the opinions he/she posted in the past are always correct. Here, an opinion o is seen as “correct” if it can bring users positive profit (i.e., $\mathcal{Q}(o) > 0$). Then, the percentage of correct opinions in historical opinions posted by the opinion author is used as our first factor (denoted by A_CP):

$$\text{A\_CP}(o) = \frac{\mathcal{N}(\mathcal{O}^c_{a^o})}{\mathcal{N}(\mathcal{O}_{a^o})}, \tag{3}$$

where $\mathcal{N}(\cdot)$ indicates the number of elements in a set, $\mathcal{O}_{a^o}$ contains the past opinions (i.e., opinions published before the releasing time of o) posted by a^o and $\mathcal{O}^c_{a^o}$ includes the correct opinions (i.e., opinions of which qualities are positive) in $\mathcal{O}^c_{a^o}$. The hypothesis about the correlation between A_CP(o) and $\mathcal{Q}(o)$ could be described as:

$\mathcal{H}^{\text{a}}_{\text{cp}}$**: The quality of an opinion o is high if the opinion author a^oalways posted correct opinions in the past (i.e., $\mathcal{Q}(o) \propto$ A_CP(o)).**

The above factor A_CP actually uses the percentage of opinions with positive qualities to reflect author expertise. However, A_CP ignores amplitude of qualities. Thus, we use the **average quality of opinions written by a^oin the past** as the another factor describing author expertise. We denote this factor as A_AQ:

$$\text{A\_AQ}(o) = \frac{1}{\mathcal{N}(\mathcal{O}_{a^o})} \sum_{o' \in \mathcal{O}_{a^o}} \mathcal{Q}(o'). \tag{4}$$

The hypothesis between A_AQ(o) and $\mathcal{Q}(o)$ is:

$\mathcal{H}^{\text{a}}_{\text{aq}}$**: The quality of an opinion o is high if a^oalways posted high-quality opinions in the past (i.e., $\mathcal{Q}(o) \propto$ A_AQ(o)).**

Besides measuring author expertise from his/her past opinions, we also explore additional information. In an investor-based social network, a user typically *follows* other users if he/she values their opinions. Thus, we use the social popularity of authors (denoted by A_SP) as a third author-level factor. We model A_SP as the natural logarithm of the number of the author's followers (i.e., a commonly used measurement for evaluating the social popularity of users) [7–9]:

$$\text{A\_SP}(o) = \ln(\mathcal{F}(a^o)), \tag{5}$$

where $\mathcal{F}(a^o)$ denotes as the number of the a^o's followers. The hypothesis under factor A_SP is that:

$\mathcal{H}^{\text{a}}_{\text{sp}}$**: The quality of an opinion o is high if $a^o$is popular in the investor based social networks (i.e., $\mathcal{Q}(o) \propto$ A_SP(o)).**

The fourth factor we use for reflecting author expertise is how much energy the author puts in investment. To measure the injected energy, we assume that if a user posts more investment opinions, he/she puts more time and energy into investment. We denote A_IE as this factor and the normalized number of a^o's past opinions as the factor value. Specifically, we have

$$\text{A\_IE}(o) = \frac{\mathcal{N}(\mathcal{O}_{a^o})}{\sum_{a' \in \mathcal{A}} \mathcal{N}(\mathcal{O}_{a'})}, \tag{6}$$

where $\mathcal{A}$ contains all authors in the investor based social network, and a hypothesis:

$\mathcal{H}^{\text{a}}_{\text{ie}}$**: The quality of an opinion o is high if a^oinjects a lot of energy (evaluated by the number of his/her past opinions) in investment (i.e., $\mathcal{Q}(o) \propto$ A_IE(o)).**

4.1.2. Content level

The words in opinion content are 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 factors may play an important role in predicting opinion quality. In our work, we consider content factors about words used by the author of the opinion. Similar to factors A_CP and A_AQ, for each word w, we calculate two scores corresponding to correctness percentage and average quality of opinions containing w, respectively. Specifically, we use the percentage of high-quality opinions in all opinions posted containing w as correctness percentage score of w (denoted as W_CP(w)):

$$\text{W\_CP}(w) = \frac{\mathcal{N}(\mathcal{O}^c_w)}{\mathcal{N}(\mathcal{O}_w)}, \tag{7}$$

where $\mathcal{O}_w$ includes the past opinions containing w and $\mathcal{O}^c_w$ contains the correct opinions in $\mathcal{O}^c_w$. We also calculate average quality score of w (denoted as W_AQ(w)) as the average quality of all opinions posted containing w:

$$\text{W\_AQ}(w) = \sum_{o' \in \mathcal{O}_w} \frac{1}{\mathcal{N}(\mathcal{O}_w)} \mathcal{Q}(o'). \tag{8}$$

Typically, there is a set of words in one opinion. Assuming that the content of opinion o consists of words $\{w_1, w_2, \cdots, w_n\}$, we define two factors C_CP(o) and C_AQ(o), as the mean of discrete values $\{\text{W\_CP}(w_i)\}_{i=1}^n$ and $\{\text{W\_AQ}(w_i)\}_{i=1}^n$, respectively.:

$$\begin{aligned} \text{C\_CP}(o) &= \frac{1}{\mathcal{N}(\mathcal{W}_o)} \sum_{w \in \mathcal{W}_o} \text{W\_CP}(w), \\ \text{C\_AQ}(o) &= \frac{1}{\mathcal{N}(\mathcal{W}_o)} \sum_{w \in \mathcal{W}_o} \text{W\_AQ}(w), \end{aligned} \tag{9}$$

where set $\mathcal{W}_o$ contains all words appearing in o. Corresponding to factors C_CP(o) and C_AQ(o), we have two hypotheses about their correlations with opinion qualities:

$\mathcal{H}^{\text{c}}_{\text{cp}}$**: The quality of an opinion o is high if the words in $o$always appeared in past correct opinions (i.e., $\mathcal{Q}(o) \propto$ C_CP(o)).**

$\mathcal{H}^{\text{c}}_{\text{aq}}$**: The quality of an opinion o is high if the words in $o$always appeared in past high-quality opinions (i.e., $\mathcal{Q}(o) \propto$ C_AQ(o)).**

4.1.3. Stock level

We also wish to consider whether the stocks an opinion talks about will determine its quality. We assume that the predictability of online investors to different stocks may be different. For example, if the prices of a stock are controlled by some institutions behind the scenes, it is hard for individual investors to predict the stock's future price, and thus the qualities of investment opinions talking about this stock are difficult to be ensured. Similar to C_CP(o) and C_AQ(o), we also characterise an opinion by using S_CP(o) and S_AQ(o):

$$\mathrm{S\_CP}(o) = \frac{1}{\mathcal{N}(\mathcal{E}_o)} \sum_{e\in\mathcal{E}_o} \mathrm{E\_CP}(e), \qquad \mathrm{S\_AQ}(o) = \frac{1}{\mathcal{N}(\mathcal{E}_o)} \sum_{e\in\mathcal{E}_o} \mathrm{E\_AQ}(e), \tag{10}$$

where $\mathcal{E}_o$ includes all stocks mentioned in o, and

$$\mathrm{E\_CP}(e) = \frac{\mathcal{N}(\mathcal{O}_e^c)}{\mathcal{N}(\mathcal{O}_e)}, \qquad \mathrm{E\_AQ}(e) = \sum_{o'\in\mathcal{O}_e} \frac{1}{\mathcal{N}(\mathcal{O}_e)} \mathcal{Q}(o'), \tag{11}$$

where $\mathcal{O}_e$ includes the past opinions talking about stock e and $\mathcal{O}_e^c$ contains the correct opinions in $\mathcal{O}_e^c$. The two hypotheses corresponding to S_CP(o) and S_AQ(o) are:

- $\mathcal{H}^s_{cp}$**: The quality of an opinion o is high if the stocks mentioned in $o$were always talked about in past correct opinions (i.e., $\mathcal{Q}(o) \propto \mathbf{S\_CP}(o)$).**
- $\mathcal{H}^s_{aq}$**: The quality of an opinion o is high if the stocks mentioned in $o$were always talked about in past high-quality opinions (i.e., $\mathcal{Q}(o) \propto \mathbf{S\_AQ}(o)$).**

4.2. Predicting qualities of investment opinions

Now, we have eight hypotheses based on eight factors. This section discusses how to fit qualities of investment opinions by utilizing eight factors simultaneously. To combine the effectiveness of the factors, we consider the linear fitting model [10]:

$$\mathcal{Q}(o) = \mathbf{w} \cdot \mathbf{x_o} + b, \tag{12}$$

where $\mathbf{x_o}$ is a vector consisted from A_CP(o), A_AQ(o), A_SP(o), A_IE(o), C_CP(o), C_AQ(o), S_CP(o), S_AQ(o), vector $\mathbf{w}$ contains weights of the factors, and b is the bias term.

Suppose we have a set of historical investment opinions $\mathcal{O}^{tr}$. For each opinion in $\mathcal{Q}^{tr}$, its real quality is known. As an instance, supposing today is 2015-03-16, real qualities of opinions posted on or before 2015-03-09 are known (since $p^d_\triangle(e)$ in formula (1) is set as the price after a week). For each o in $\mathcal{O}^{tr}$, we have an observation $(x_o, \mathcal{Q}(o))$. Given a sufficiently large number of such observations, one can reliably estimate the true underlying parameters (i.e., $\mathbf{w}$ and b), by employing the technique of least squares [11]. Moreover, according to the meanings of each factor and related hypothesis, it is hard to trust negative correlation between each factor and opinion quality. For instance, A_CP (i.e., the percentage of author's correct opinion in the past) is not likely negatively related to the quality of opinions posted by the author. For this reason, non-negativity constraint upon $\mathbf{w}$ should be included. Thus, we formulate modeling prediction function as a problem of non-negative least squares:

$$(\mathbf{w}^*, b^*) = \arg\min_{\mathbf{w}\geq 0} \sum_{o\in\mathcal{O}^{tr}} ||\mathbf{w} \cdot \mathbf{x_o} + b - \mathcal{Q}(o)||_2, \tag{13}$$

where $||\cdot||_2$ denotes the $\mathbf{L}_2$ norm and $\mathbf{w}{\geq}0$ means that each component of $\mathbf{w}$ should be non-negative. The optimal parameters $\mathbf{w}^*$, b^* can be solved by a number of different ways in the research of non-negative least squares [11–14]. In our work, we use the technique proposed in [14]. Once we have $\mathbf{w}^*$ and b^*, the predicted quality of a new opinion o' is $\mathbf{w}^* \cdot \mathbf{x_{o'}} + b^*$. With these predicted qualities, we could rank candidate opinions for finding high-quality opinions.

4.3. Investment recommendation with predicted opinion qualities

The purpose of investment recommendation is giving users investment advice. A typical task of investment recommendation is recommending a set of stocks (i.e., a portfolio) to users. Besides portfolio recommendation, we also consider recommending users a set of high-quality investment opinions. Compared with a single portfolio, investment opinions include more information (e.g., the reasons why opinion authors suggest to long a stock) for users as references.

4.3.1. High-quality opinion recommendation

Once we have collected a set of training data (including historical investment opinions and their real qualities), we can learn the prediction function by formulation (13). Then, we recommend investment opinions with high predicted qualities to users to help them invest. The procedure for recommending high-quality opinions is summarized by Procedure (1) (assuming the recommendation is performed at 8am each day).

Procedure 1 Daily opinion recommendation (on day d).

Input:

$\mathcal{O}^{tr}_d$ = historical investment opinions of which real qualities are known

$\mathcal{O}^{te}_d$ = investment opinions posted during from the last recommendation time (i.e., 8am on day $d-1$) to current recommendation time (i.e., 8am on day d)

Output:

k recommended opinions $\mathcal{O}^{rec}_d$

Construct training set $\{(\mathbf{x_o}, \mathcal{Q}(o))\}$ according to $\mathcal{O}^{tr}_d$
Learn $f : \mathcal{X} \to$ y (by formulation 13)
for all $o' \in \mathcal{O}^{te}_d$ **do**
  Construct $\mathbf{x_{o'}}$ according to o'
  Estimate predicted quality $\hat{\mathcal{Q}}(o')$ as $f(\mathbf{x_{o'}})$
end for
$\mathcal{O}^{rec}_d = k$ opinions in $\mathcal{O}^{te}_d$ according to k highest $\hat{\mathcal{Q}}$ values

4.3.2. Portfolio recommendation by quality-sensitive sentiment aggregation

Portfolio recommendation suggests a set of stocks and related actions (long or short) to users. Conventional methodologies for generating a portfolio based on investment opinions always aggregate a large amount of investors' sentiments into a score for each stock. However, few of them consider reducing noise viewpoints by taking opinion qualities into account. In this section, we first briefly introduce traditional approaches to recommend portfolio, and then discuss how to improve portfolio generation by merging investment sentiments in a quality-sensitive manner.

In previous work, profitable stocks are selected according to their aggregate sentiment indexes. One of the most popular aggregate bullish and bearish sentiment indexes [2–4] are:

$$\delta_i^{bu} = \ln\left[\frac{1+\mathcal{N}(\mathcal{O}_i^{bu})}{1+\mathcal{N}(\mathcal{O}_i^{be})}\right], \qquad \delta_i^{be} = \ln\left[\frac{1+\mathcal{N}(\mathcal{O}_i^{be})}{1+\mathcal{N}(\mathcal{O}_i^{bu})}\right], \tag{14}$$

where δ_i^{bu} (δ_i^{be}) denotes bullish (bearish) sentiment index for stock e_i while $\mathcal{O}_i^{bu}$ ($\mathcal{O}_i^{be}$) is the set of opinions containing bullish (bearish) views on e_i. In other words, for each o in $\mathcal{O}_i^{bu}$ ($\mathcal{O}_i^{b}$), 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 recommending stocks, we could sort stocks in decreasing order of $|\delta_i^{bu} - \delta_i^{be}|$ and take the top k' ones as the recommended stocks. Finally, we suggest to long (short) the stocks if their bullish indexes are much larger (smaller) than their bearish indexes.

In our work, we attempt to utilize the predicted qualities in stock recommendation by giving higher weights to sentiments in high-quality opinions. We call this **quality-sensitive sentiment aggregation**. Note that in Eq. (14), $\mathcal{N}(\mathcal{O}_i^{bu})$ actually equals $\sum_o \mathbb{I}\{o \in \mathcal{O}_i^{bu}\}$, where $\mathbb{I}$ is the indicator function (i.e., $\mathbb{I}\{\text{a true statement}\} = 1$, and $\mathbb{I}\{\text{a false statement}\} = 0$). Thus, by weighting views in o with $\hat{\mathcal{Q}}(o)$, the quality-sensitive aggregation for sentiment indexes should be

$$\delta_i^{bu} = \ln\left[\frac{1 + \sum_o \hat{\mathcal{Q}}(o)\mathbb{I}\{o \in \mathcal{O}_i^{bu}\}}{1 + \sum_o \hat{\mathcal{Q}}(o)\mathbb{I}\{o \in \mathcal{O}_i^{be}\}}\right],$$
$$\delta_i^{be} = \ln\left[\frac{1 + \sum_o \hat{\mathcal{Q}}(o)\mathbb{I}\{o \in \mathcal{O}_i^{be}\}}{1 + \sum_o \hat{\mathcal{Q}}(o)\mathbb{I}\{o \in \mathcal{O}_i^{bu}\}}\right]. \tag{15}$$

The recommendation procedure for k'-stock portfolios is summarized by Procedure (2).

Procedure 2 Daily portfolio recommendation (on day d).

Input:

$\mathcal{E}_d$ = stocks tradable and mentioned on day d
$\mathcal{O}_d^{tr}$ = historical investment opinions of which real qualities are known
$\mathcal{O}_d^{te}$ = investment opinions posted during from the last recommendation time (i.e., 8am on day $d-1$) to current recommendation time (i.e., 8am on day d)

Output:

k recommended opinions $\mathcal{O}_d^{rec}$

```
Construct training set {(x_o, Q(o))} according to O_d^tr
Learn f : X → y (by formulation 13)
for all o' ∈ O_d^te do
    Construct x_o' according to o'
    Estimate predicted quality Q̂(o') as f(x_o')
end for
for all e_i ∈ E_d do
    Calculate sentiment indexes δ_i^bu and δ_i^be according to formulation 15
    Calculate stock score s(e_i) = |δ_i^bu − δ_i^be|
end for
E_d^rec = k' stocks in E_d accordingto k' highest stock scores
```

5. Experiments

5.1. Data preparation

For experiments, we collect all messages posted during from 2014-01-01 to 2015-05-31 from the investor-based social media StockTwits. In StockTwits, users post short messages (limited to 140 characters) that include ideas or opinions on specific investments. Stock symbols in messages are preceded by a "CashTag" ($) (as Fig. 1 shows). Here, we only taking StockTwits messages having at least one CashTag and explicit sentiment label as investment opinions. Finally, the dataset of StockTwits investment opinions contains 2,325,858 messages posted by 40,737 users and related to 5723 stocks traded at NYSE and Nasdaq. We used the Yahoo! Finance API[3] to crawl historical prices of stocks.

[3] www.finance.yahoo.com.

5.2. Preliminary data analysis

To better understand the structure of the dataset, we performed some preliminary data analysis. Fig. 2(a) represents the percentage of opinions having at least one "CashTag" $ in the Stocktwits platform. We can find that more than half of opinions talk about at least one stock, which verifies that the stocktwits platform is a good source for collecting investor opinions about individual stocks. Fig. 2(b) shows that most opinions talk about stocks traded in NYSE and NASDAQ market. As shown by Fig. 2(c), among the opinions containing sentiment tags, the number of bullish opinions is much more than the number of bearish opinions. From Fig. 2(d), we can see that most opinions posted in Stocktwits Platform have 2–30 words. Fig. 2(e) and (f) represent two cases of "long tail"[4] [15]. Specifically, a large proportion of opinions are posted by a small proportion of users and related to a small proportion of stocks. Fig. 2(g)–(i) show the number of opinions, users and stocks appeared on the first day of each month. On 2015-01-01, around 2600 users posted more than 90,000 investment opinions related to more than 1000 stocks. Moreover, it seems that the numbers would continually grow in the future. Obviously, it is very hard for a user or even an investment institution to through out all of the investment opinions. Thus, identifying high-quality investment opinions and high-profit stocks are of great use, which confirms the practical value of our work.

5.3. Experimental setup

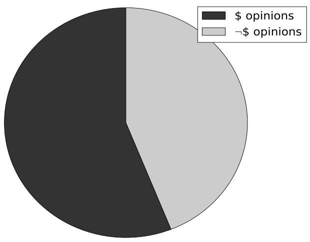
We note that the opening time of stock markets is 9:30 am. Thus, in our experiments, we set the time to generate recommendations at 8:00 am on each tradable day. If recommendation results are given to users at 8:00 am, they may have adequate time to consider whether or not trade according to our recommendation. We start the investment recommendation from 2014-06-01 (rather than 2014-01-01) since we need a collection of historical data for calculating values of factors A_CP, A_AQ, A_IE, C_CP, C_AQ, S_CP, S_AQ. During the progression of the experiment, we will continuously update the historical data. As an instance, when the recommendation is simulated on 2015-03-16, the historical opinions (i.e., opinions of which real qualities are known) includes opinions posted on or before 2015-03-09 (since $p_\triangle^d(e)$ in formula (1) is set as the price after a week in our work).

5.4. Hypothesis testing

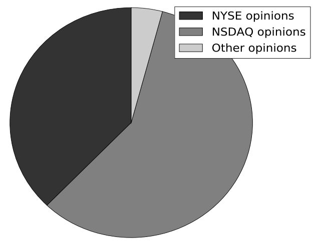
Before presenting the experimental results of opinion and portfolio recommendation, we first test hypotheses introduced in Section 4.1. We have proposed eight factors A_CP, A_AQ, A_IE, C_CP, C_AQ, S_CP, S_AQ. Each factor corresponds to a hypothesis. Given a factor, the corresponding hypothesis actually assumes the factor values of high-quality opinions are high while factor values of low-quality opinions are low. For instance, $\mathcal{H}_{\mathrm{cp}}^{\mathrm{a}}$ assumes A_CP values of high-quality opinions are higher than A_CP values of low-quality opinions. Thus, for each factor v, we calculate **A**verage **F**actor **V**alue of opinions in **H**ighest m **Q**ualities (denoted as **AFV-HQ**) and **A**verage **F**actor **V**alue of opinions in **L**owest m **Q**ualities (denoted as **AFV-LQ**):

$$\text{AFV-HQ}(v) = \frac{1}{m}\sum_{o \in \mathcal{O}_h^q} \mathcal{V}(o), \quad \text{AFV-LQ}(v) = \frac{1}{m}\sum_{o \in \mathcal{O}_l^q} \mathcal{V}(o), \tag{16}$$

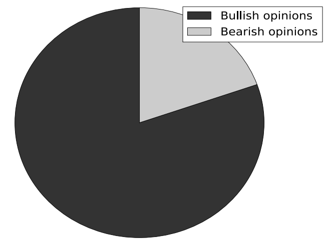
[4] A long tail of some distributions of numbers is the portion of the distribution having a large number of occurrences far from the "head" or central part of the distribution.



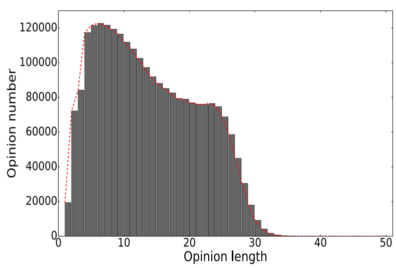
(a) The percentage of $ opinions.



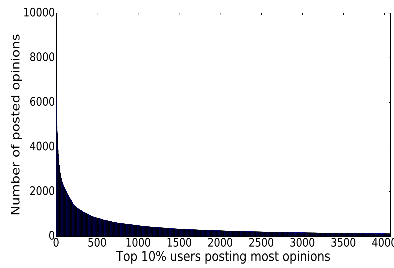
(b) The percentages of stocks traded in different markets.



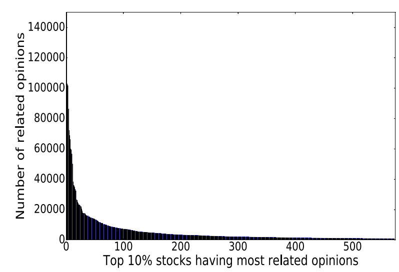
(c) The percentages of bullish and bearish opinions.



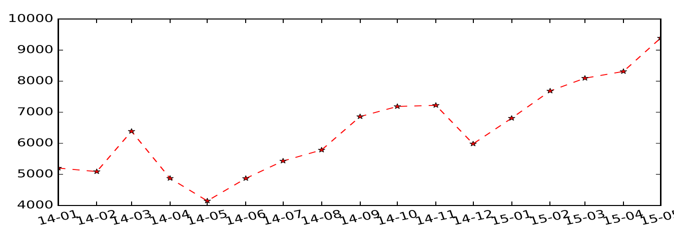
(d) The distribution of opinion length.



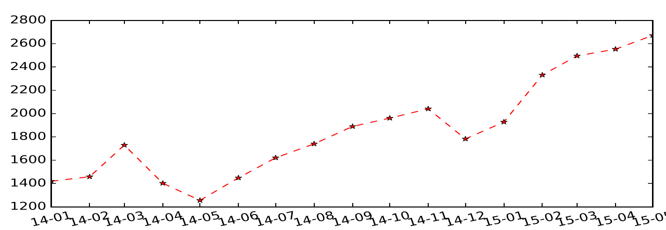
(e) The distribution of the number of opinions posted by each user.



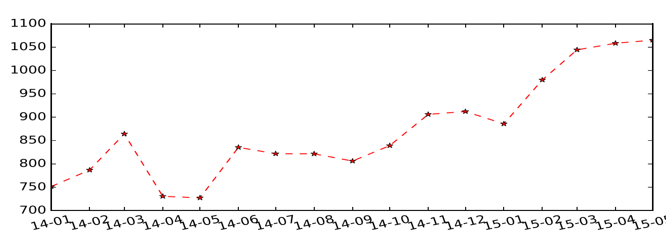
(f) The distribution of the number of opinions related to each stock.



(g) The number of opinions posted on the first day of each month.



(h) The number of users appeared on the first day of each month.



(i) The number of stocks mentioned on the first day of each month.

Fig. 2. Preliminary data analysis. For more details and a discussion of the results, please see Section 5.2.

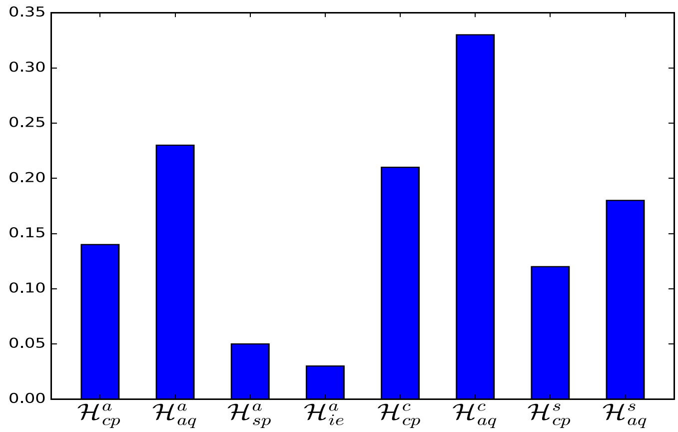


Fig. 3. Hypothesis testing.

where $\mathcal{V}(o)$ denotes o's value of factor v and $\mathcal{O}^q_h$ ($\mathcal{O}^q_l$) includes opinions in top (bottom)-m qualities. Then, the reliability of a hypothesis can be measured by:

$$\text{Reliability}(\mathcal{H}) = \frac{\text{AFV-HQ}(v) - \text{AFV-LQ}(v)}{|\text{AFV-HQ}(v) + \text{AFV-LQ}(v)|}, \tag{17}$$

where $\mathcal{H}$ is the hypothesis upon factor v (e.g., $\mathcal{H}$ is $\mathcal{H}^{\text{a}}_{\text{cp}}$ if v is A_CP).

Fig. 3 shows reliability values (when $m = 100$) of hypotheses $\mathcal{H}^{\text{a}}_{\text{cp}}, \cdots, \mathcal{H}^{\text{s}}_{\text{aq}}$ introduced in Section 4.1. The result supports all of hypotheses, to different extent. Moreover, the fact that reliability values of all hypotheses are positive verifies the rationality behind the non-negativity constraint $\mathbf{w}\succeq 0$ used in formulation (13).

5.5. Comparison on recommending high-quality opinions

5.5.1. Evaluation metrics

In our experiment, we recommend an opinion set $\mathcal{O}^{rec}_d$ to users on each tradable day d during the test session (i.e., from 2014-06-01 to 2015-05-31). For each $\mathcal{O}^{rec}_d$, we evaluate its performance (denoted by $\mathcal{P}(\mathcal{O}^{rec}_d)$) as the average quality of opinions in it:

$$\frac{1}{\mathcal{N}(\mathcal{O}^{rec}_d)} \sum_{o \in \mathcal{O}^{rec}_d} \mathcal{Q}(o). \tag{18}$$

Finally, the performance of an opinion-recommendation strategy is quantified as the average of performances of all generated opinion sets on all tradable days $\mathcal{D}^{te}$:

$$\frac{1}{\mathcal{N}(\mathcal{D}^{te})} \sum_{d \in \mathcal{D}^{te}} \mathcal{P}(\mathcal{O}^{rec}_d). \tag{19}$$

5.5.2. Baselines

Random Recommendation: The simplest method to recommend opinions is random selection. Specifically, we randomly selected k opinions as $\tilde{\mathcal{O}}^{rec}_d$, and repeated this process 10,000 times. Finally, we use the average value of all selected $\tilde{\mathcal{O}}^{rec}_d$s' performances as the performance of $\mathcal{O}^{rec}_d$ generated by **Random** recommendation.

Expert based Recommendation: We use the author expertise model (denoted as **Expert**) proposed in [5] to recommend opinions, by recommending k opinions posted by top experts.

Hypothesis based Recommendation: As introduced in Section 4, we have eight hypotheses: $\mathcal{H}^{\text{a}}_{\text{cp}}, \cdots, \mathcal{H}^{\text{s}}_{\text{aq}}$. Each hypothesis corresponds to a recommendation strategy that ranks opinions according to their values of the factor. Then, the high-quality opinions are selected as ones with top factor values. For example, hypothesis $\mathcal{H}^{\text{a}}_{\text{cp}}$ (i.e., $\mathcal{Q}(o) \propto \text{A\_CP}(o)$) ranks investment opinions according to their A_CP values and consider opinions with highest A_CP values as high-quality opinions. Thus, according to $\mathcal{H}^{\text{a}}_{\text{cp}}, \cdots, \mathcal{H}^{\text{s}}_{\text{aq}}$, we have eight compositors for recommending opinions.

Predicted-quality based Recommendation: As introduced in Section 4, our methodology predicts opinion qualities by combining the effectiveness of multiple factors. Finally, recommended opinions are ones in top-k qualities. In our work, searching optimal prediction function is formulated as a problem of non-negative least squares (see formulation (13) for details). If we remove the non-negative constraint $\mathbf{w}\succeq 0$, it will become a linear-regression problem [16]. We will also compare performances of opinion recommenders using **Pred**icted **Qual**ities estimated by Linear Regression (**LR**) and Non-Negative Least Squares (**NNLS**), denoted as **PredQual(LR)** and **PredQual(NNLS)**, respectively.[5]

5.5.3. Recommendation performance

Fig. 4 represents recommendation performances of competitors introduced in Section 5.5.2. As Fig. 4 shows, each of the proposed hypotheses could be used for retrieving high-quality opinions to a different extent, since the recommendation performances corresponding to all hypotheses are larger than the random selection. The experimental result indicates that the methods (**Expert**, $\mathcal{H}^{\text{a}}_{\text{cp}}, \mathcal{H}^{\text{a}}_{\text{aq}}, \mathcal{H}^{\text{a}}_{\text{sp}}, \mathcal{H}^{\text{a}}_{\text{ie}}$) only rely on author information perform worse than our methods (**PredQual(LR), PredQual(NNLS)**), which verifies the value of combining various information such as content and related stocks. The best performance is achieved by **PredQual(NNLS)**, which demonstrates the effectiveness of adding the non-negative constraint $\mathbf{w}\succeq 0$ in formulation (13).

5.6. Comparison on recommending portfolios

5.6.1. Evaluation metrics

Given a recommended portfolio $\mathcal{E}^{rec}_d$ generated on the test date d, let $\mathcal{E}^{rec}_{d_l}$ ($\mathcal{E}^{rec}_{d_s}$) denoted stocks in $\mathcal{E}^{rec}_d$ and of which bullish (bearish) sentiment index is larger. For evaluating $\mathcal{E}^{rec}_d$, we use the profit earned by a user if he/she long the stocks in $\mathcal{E}^{rec}_{d_l}$ (i.e., stocks in $\mathcal{E}^{rec}_d$ and should be longed) and short the stocks in $\mathcal{E}^{rec}_{d_s}$ (i.e., stocks in $\mathcal{E}^{rec}_d$ and should be shorted). Thus, the performance of $\mathcal{E}^{rec}_d$ can be evaluated as (denoted by $\mathcal{P}(\mathcal{E}^{rec}_d)$):

$$\frac{1}{\mathcal{N}(\mathcal{E}^{rec}_d)} \left(\sum_{e \in \mathcal{E}^{rec}_{d_l}} \frac{p^d_{\triangle}(e) - p^d(e)}{p^d(e)} + \sum_{e \in \mathcal{E}^{rec}_{d_s}} \frac{p^d(e) - p^d_{\triangle}(e)}{p^d(e)} \right). \tag{20}$$

Similar to performance of investor-opinion recommendation, the final performance of stock recommendation is formulated as

$$\frac{1}{\mathcal{N}(\mathcal{D}^{te})} \sum_{d \in \mathcal{D}^{te}} \mathcal{P}(\mathcal{E}^{rec}_d). \tag{21}$$

5.6.2. Baselines

Traditional Sentiment Aggregation (Trad-Aggr): As a baseline method for portfolio recommendation, we employ traditional sentiment indexes (see Eq. (14)) used in previous work [2–4] to sort stocks and generate portfolios (as Section 4.3.2 discusses).

[5] In [6] (i.e., the conference version of our work), we used PredQual(LR) for opinion recommendation. In this manuscript, we extend it into PredQual(NNLS) by adding the non-negative constraint $\mathbf{w}\succeq 0$.

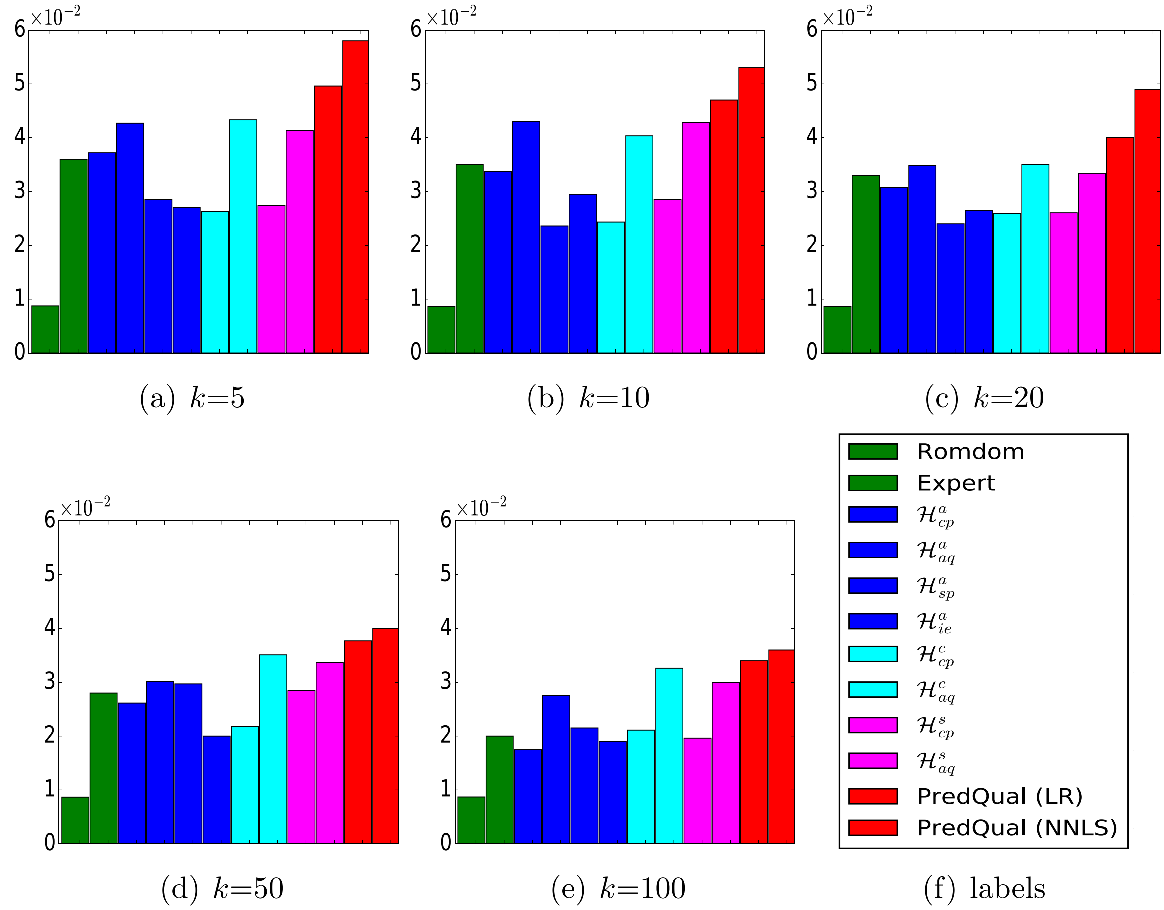


Fig. 4. Performance comparison for opinion recommendation. From left to right in order, the bars correspond to methods **Random, Expert**, $\mathcal{H}^{a}_{cp}$, $\mathcal{H}^{a}_{aq}$, $\mathcal{H}^{a}_{sp}$, $\mathcal{H}^{a}_{ie}$, $\mathcal{H}^{c}_{cp}$, $\mathcal{H}^{c}_{aq}$, $\mathcal{H}^{s}_{cp}$, $\mathcal{H}^{s}_{aq}$, **PredQual(LR)** and **PredQual(NNLS)** (see Section 5.5.2 for details). The y-axis shows the performance of opinion recommendation (i.e., formulation (19)).

Expert Identification: As another competitor, we use the author expertise model proposed in [5] to generate expert-sensitive sentiment indexes (denoted as **Expert**, see [5] for details).

Quality-sensitive Sentiment Aggregation (QS-Aggr): Our work improves traditional sentiment indexes by aggregating opinion sentiments with opinion qualities as weights (see formulation (15)).

5.6.3. Recommendation performance

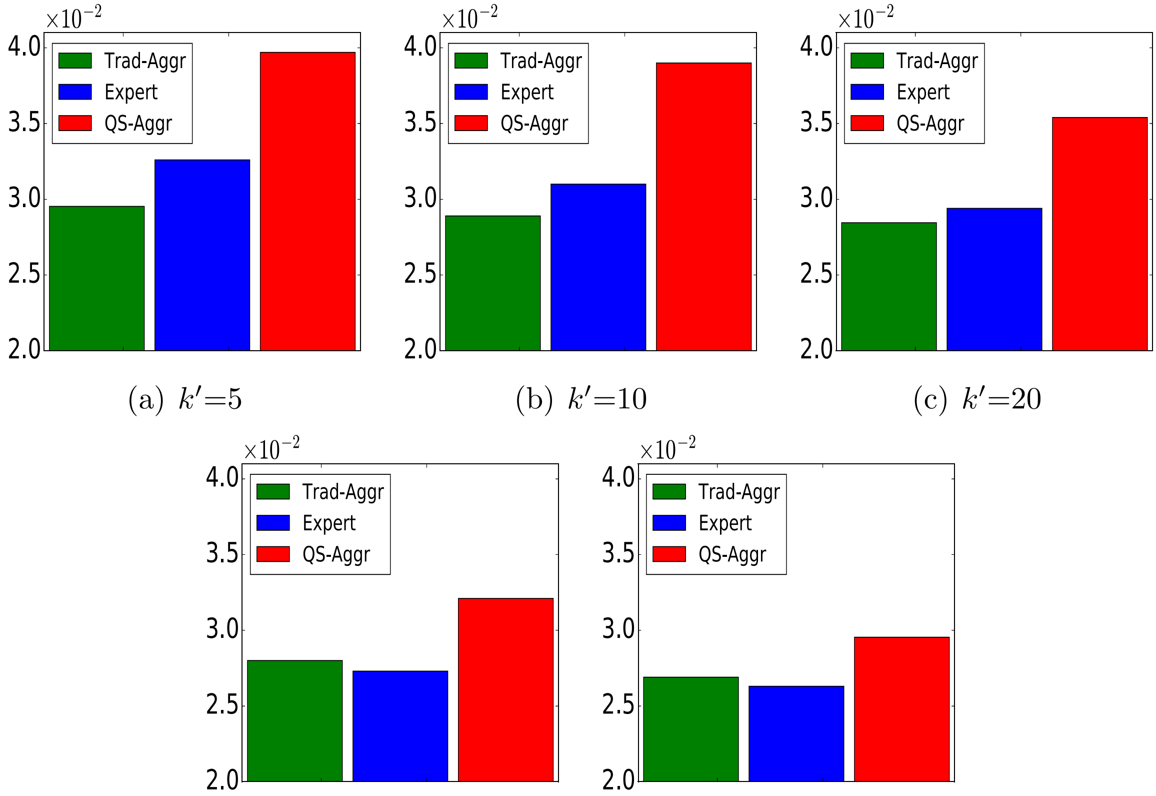
In Fig. 5, we compare the performances (i.e., formulation (21)) of our method (i.e., **QS-Aggr**) and other competitors (i.e., **Trad-Aggr** and **Expert**). Both of **Expert** and **QS-Aggr** outperform **Trad-Aggr**, which verifies the effectiveness of giving high-quality opinions larger weights when combining opinion sentiments. Our method **QS-Aggr** suggests the most profitable stocks and demonstrates the effectiveness of quality-sensitive sentiment indexes. This result demonstrates that qualities of investor opinions should be measured by considering various factors (rather than only considering authors' previous performance).

6. Related work

Recall that our methodology aims to recommend a set of stocks by aggregating views in high-quality opinions which are posted on investor-based social networks. In this section, we summarize work related to ours. First, we review approaches studying financial recommendation systems. Then, recent methods that predict stock market with social-media data are summarized.

6.1. Recommendation systems in finance

Recommendation systems, as an edge tool to combat information overload [17], have been advocated by both academia and industry. The objective of recommendation systems is automatically delivering to users things they are interested in [18]. In the financial domain, recommendation systems have gained an increasing attention [19]. Users in financial recommendation systems are always real-world investors having investment intentions and items are finance-related products such as stocks, loans and real estates. Among various kinds of financial products, stocks have gained most attention. In recent years, lots of approaches use intelligent algorithms to predict stock markets and identify stocks which might bring to users long-term or short-term returns. The intelligent algorithms that have been successfully applied in this direction include support vector machines (SVM), genetic algorithm, deep learning and so on [20–22]. Another direction is to perform stock recommendation with social-media data [23]. A detail review of studies in this direction will be given in the later subsection (i.e., Section 6.2). Moreover, personalization criteria is considered in some studies for improving stock recommendation. Liu and Lee [24] attempts to realize personalized recommendation by designing several features to select stocks according to preferences of a specific user. Chalidabhongse and Kaensar [25] describes investor profiles by a set of dimensions, and then proposes a system to give personalized recommendation to the investors based on their personal profiles and their historical system interactions. Taghavi et al. [26] proposes a hybrid recommender which includes collaborative



(a) $k'=5$ (b) $k'=10$ (c) $k'=20$ (d) $k'=50$ (e) $k'=100$

Fig. 5. Comparison of portfolio recommendation by **Trad** and **Expert** and our method **PredQual (NNLS)**. The y-axis shows the performance of portfolio recommendation (i.e., formulation (21)).

and content-based filtering to provide personalized stock recommendation. Yang et al. [27] divides investors into several groups, and then recommends stocks preferred by the users who are in the same group with the target investor. Since our work recommends stocks by discovering high-quality investor opinions, the recommendation service provided by our method is non-personalized. One of our future work is to extend our approach to realize personalized investment recommendation (see Section 7 for more details).

6.2. Stock prediction with social media

Note that our work is based on investor opinions from social-media platforms. In this subsection, we review previous work which also studies the use of social media for predicting stock markets.

Search queries are used for capturing investor attention and predict stock movements according to the theory of behavioral economics. Antweiler and Frank [28] collected a large volume (i.e., 1.5 million) messages from Yahoo Finance and Raging Bull platforms to study the predictive power of online messages for the stock market. Finally they found that stock turnovers could be predicted by the volumes of messages. The analysis in [29] is performed on a collection of queries submitted to a popular search engine. Their work shows that dynamics of query volume can identify early warnings of financial risk.

Besides searching behavior, the public emotion (e.g., joy, sadness) revealed in social-media platforms is another indicator commonly used for stock prediction. Bollen et al. [30] investigated whether collective mood states on Twitter are related to the value of the Dow Jones Industrial Average (DJIA). They find that certain mood states are indeed predictive of the DJIA closing values. Zhang et al. [31] explored the relationship between hope and fear on the one hand and the Dow Jones, NASDAQ, and S&P 500 on the other hand. Their results indicate that the level of tweet emotionality was significantly related to all three aggregated indicators. Recently, in [32], through study on over 10 million stock-relevant tweets from Weibo, both correlation analysis and causality test show that five attributes of the stock market in China can be competently predicted by various online emotions, like disgust, joy, sadness and fear. López-Cabarcos et al. [33] studied the impact technical and non-technical investors have on the stock market. However, separating Stocktwits users into technical and non-technical investors might suffer from a lack of accuracy since the users do not explicitly show whether they are technical or non-technical. Sul et al. [23] performed experiments to analyze the cumulative sentiment (positive and negative) in 2.5 million Twitter postings about individual S&P 500 firms and compared this to the stock returns of those firms. The results showed that the sentiment in tweets about a specific firm from users with less than 171 followers had a significant impact on the stock's returns.

In this paper, we choose investment opinions as the resource for realizing investment recommendation. Compared with search queries, investment opinions always contain much richer information to enable us to infer the motivation behind investors' attentions. Moreover, sentiments in investment opinions reveal attitudes of investors to stock prediction. Thus, compared to public emotion, they are likely to be correlated to the stock market more strongly. Rao and Srivastava [3] presented the strong correlation between stock prices and sentiments generated from 4 million tweets related to stocks belonging to DJIA, NASDAQ-100 and 13 other big enterprises. Mao et al. [1] forecasted the stock market by the aggregation of bullish and bearish sentiments containing in tweets. In their experiments, both of sentiment and volume indicators have a strong predictive power of daily market returns. Oh and Sheng [2] employed a J49 classifier to infer sentiment tags of investment opinions collected from the stocktwits platform. They also proved the predictive ability of their sentiment indicators for future market directions. Tu et al. [4] improved sentiment classification on bullish/bearish attitudes expressed in investment opinions and obtains a more predictable sentiment index. Bing et al. [34] focused on studying sentiment analysis on tweets about specific listed companies.

In the above work, investment recommendation is made by aggregating the views (e.g., buy or sell) contained in all the opinions, and buying most buy stocks that get most buy votes and short-selling the stocks with the strongest bearish sentiment. These methods aggregate the views in opinions to reflect the wisdom of the whole but ignore the variability in user expertise. Thus, finding a way to evaluate the quality of investment opinions and treat opinions with different qualities differently is essential to improve the predicting effectiveness. The purpose of our work is to analyze the opinion quality from multiple perspectives. The most closely related work to ours is the work of Bar-Haim et al. [5]. It improves previous work by modeling user expertise to reduce the noise. Specifically, it presents a framework for identifying experts. First, they define the expertise as the percentage of correct opinions in the historical opinions of a user. Then, they will use the Pearson's Chi–square test to detect experts if they significantly outperform others. The limitation of this work is that it uses only author factor to model the quality variance. In their methodology, different opinions written by the same author will have the same quality. 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). To detect the quality difference among opinions from the same author, we need to explore factors. Thus, in our work, we explore many additional factors to explicitly predict the quality of each specific opinion rather than the expertise of a single user. Moreover, in [5], author expertise only relies on the past performance of the author while our work utilizes more factors of authors to derive their characteristics.

7. Conclusion and future work

In this paper, we studied the problem of estimating the qualities of investment opinions posted on investor based social networks. We first proposed several related factors and then predicted qualities of investment opinions by combining the factors. Moreover, we have studied to use predicted opinion qualities to retrieve high-quality opinions and generate profitable portfolio more effectively. Experiments on a real-world dataset verify the effectiveness of our work in recommending high-quality investment opinions and improving conventional portfolio recommendation.

As future work, we plan to explore more factors to predict qualities of investment opinions more effectively. For an instance, the propagation characteristic of investment opinions may be useful in estimating the opinion quality by assuming high-quality opinions will gain more attentions and be forwarded more frequently. Moreover, motivated by recent progress (such as [35]) in personalized opinion recommendation, we intend to also consider personal investment preferences of users for further improving the investment recommendation, using personalization criteria. Finally, we plan to use the theory of topic models [36] to improve the accuracy of discovering high-quality investor opinions.

References

[1] H. Mao, S. Counts, J. Bollen, Predicting financial markets: comparing survey, news, twitter and search engine data, arXiv:1112.1051 (2011).
[2] C. Oh, 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.
[3] T. Rao, 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.
[4] W. Tu, D. Cheung, N. Mamoulis, Time-sensitive opinion mining for prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2015.
[5] R. Bar-Haim, E. Dinur, R. Feldman, M. Fresko, G. Goldstein, Identifying and following expert investors in stock microblogs, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011.
[6] W. Tu, D. Cheung, N. Mamoulis, M. Yang, Z. Lu, Investment recommendation using investor opinions in social media, in: Proceedings of the international ACM SIGIR conference on Research and development in information retrieval (SIGIR), 2016.
[7] P. Analytis, A. Delfino, J. Kämmer, M. Moussaïd, T. Joachims, Ranking with social cues: integrating online review scores and popularity information, in: Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), 2017.
[8] E. Pancer, M. Poole, The popularity and virality of political social media: hashtags, mentions, and links predict likes and retweets of 2016 us presidential nominees tweets, Soc. Influ. 11 (4) (2016) 259–270.
[9] D. Garcia, P. Mavrodiev, D. Casati, F. Schweitzer, Understanding popularity, reputation, and social influence in the twitter society, Policy Int. 9 (3) (2017) 343–364.
[10] W. Wheaton, Linear least squares, Stat. Chall. Modern Astron. II (2012) 441.
[11] C. Lawson, R. Hanson, Solving least squares problems, Math. Comput. 30 (135) (1995) 665.
[12] A. Dax, On computational aspects of bounded linear least squares problems, ACM Trans. Math. Software 17 (1) (1991) 64–73.
[13] R. Bro, D. J., A fast non-negativity-constrained least squares algorithm, J Chem. 11 (5) (1997) 393–401.
[14] M. Slawski, M. Hein, Non-negative least squares for high-dimensional linear models: consistency and sparse recovery without regularization, Electron. J. Stat. 7 (0) (2012) 3004–3056.
[15] B. Alpheus, S. Dwayne, The Long Tail of Expertise, Pearson Education, 2011.
[16] S. George, L. Alan, Linear Regression Analysis, 936, John Wiley & Sons, 2012.
[17] M. Aljukhadar, S. Senecal, C. Daoust, Using recommendation agents to cope with information overload, Int. J. Electron. Commerce 17 (2) (2012) 41–70.
[18] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, Knowl. Data Eng. IEEE Trans. 17 (6) (2005) 734–749.
[19] D. Zibriczky, Recommender systems meet finance: a literature review, in: Proceedings of the International Workshop on Personalization & Recommender Systems in Financial Services (FinRec), 2016.
[20] W. Huang, Y. Nakamori, S. Wang, Forecasting stock market movement direction with support vector machine, Comput. Operat. Res. 32 (10) (2005) 2513–2522.
[21] H. Kim, K. Shin, A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets, Appl. Soft Comput. 7 (2) (2007) 569–576.
[22] E. Chong, C. Han, F. Park, Deep learning networks for stock market analysis and prediction: methodology, data representations, and case studies, Expert Syst. Appl. 83 (2017) 187–205.
[23] H. Sul, A. Dennis, L. Yuan, Trading on twitter: using social media sentiment to predict stock returns, Dec. Sc. 48 (3) (2017) 454–488.
[24] N. Liu, K. Lee, An intelligent business advisor system for stock investment, Expert Syst. 14 (3) (1997) 129–139.
[25] T. Chalidabhongse, C. Kaensar, A personalized stock recommendation system using adaptive user modeling, in: Proceedings of the International Symposium on Communications and Information Technologies (ISCIT), 2006, pp. 463–468.
[26] M. Taghavi, K. Bakhtiyari, E. Scavino, Agent-based computational investing recommender system, in: Proceedings of the ACM Conference on Recommender Systems (RecSys), ACM, 2013, pp. 455–458.
[27] Y. Yang, J. Li, Y. Yang, An efficient stock recommendation model based on big order net inflow, Math. Problems Eng. 2016 (2016).
[28] W. Antweiler, M. Frank, Is all that talk just noise? the information content of internet stock message boards, J. Finance. 59 (3) (2004) 1259–1294.
[29] I. Bordino, S. Battiston, G. Caldarelli, M. Cristelli, A. Ukkonen, I. Weber, Web search queries can predict stock market volumes, PLoS ONE 7 (7) (2012) e40014.
[30] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1) (2011) 1–8.
[31] X. Zhang, H. Fuehres, P. Gloor, Predicting stock market indicators through twitter I hope it is not as bad as I fear, Procedia-Social and Behavioral Sciences 26 (2011) 55–62.
[32] Z. Zhou, J. Zhao, K. Xu, Can online emotions predict the stock market in China? in: International Conference on Web Information Systems Engineering (WISE), 2016.
[33] M. López-Cabarcos, J. Piñeiro-Chousa, A. Pérez-Pico, The impact technical and non-technical investors have on the stock market: evidence from the sentiment extracted from social networks, J. Behav. Exp. Finance 15 (2017) 15–20.
[34] L. Bing, K. Chan, C. Ou, R. Sun, Discovering public sentiment in social media for predicting stock movement of publicly listed companies, Inf. Syst. 69 (2017) 81–92.
[35] W. Tu, D. Cheung, N. Mamoulis, More focus on what you care about: personalized top reviews set, Neurocomputing 254 (1) (2017) 3–12.
[36] D. Blei, Probabilistic topic models, Commun. ACM 55 (4) (2012) 77–84.