Dear Editor and Reviewers,

We would like to express our gratitude for the great efforts that you and the anonymous referees have put in this paper. We found the suggestions are constructive. We have strived to address all the issues as thoroughly as possible. The main changes are summarized as follows:

1. The text segmentation algorithm is moved to Section 4.1. Before we introduce how we collect data we introduce the segmentation algorithm as a tool first. We think this way makes reader clearer.

2. We move all the experimental results to Section 6.

We hope our revision meets your satisfaction, and we look forward to your favorable response shortly.

Regards,

Ning Wang

Shanhui Ke

Yibo Chen

Tao Yan

Andrew Lim

Reviewer #1

In this paper, the authors employed text mining and statistical models to explore the relationship between the Shanghai Stock Exchange Composite Index and the collective emotions of individual investors. The emotions of individual investors are quantified by extracting and aggregating investor online posts that contain finance-related keywords. Viewpoints are quantified by analyzing and aggregating the online posts of individual investors that are related to stock markets, in which the major challenge lies in how to automatically identify a set of suitable keywords. To overcome this difficulty, a keyword set that contains k finance-related words is determined semi-automatically. Text segmentation is also a challenge when dealing with Chinese text mining. The authors proposed a specialized procedure for breaking Chinese sentences into words should be developed. The proposed procedure is based on two concepts, namely, stickiness and entropy. This procedure is reasonable.

Statistical results of posts from Sina Weibo suggest a relationship between collective viewpoints reflected on Sina Weibo and movements of the SSECI. This study contributes to the existing literature by demonstrating that the microblog sentiment level reports can be quantitatively incorporated as a proxy to provide valuable support to portfolio decision making.

Considering the number of the keywords is large, the authors showed the VIF for each variable (the frequency of each keyword). For rigorous logic reasoning, the correlation matrix of these variables is suggested to be presented here. And the authors are suggested to using the principal component analysis to the variables before they are incorporated into the models.

Reviewer #2:

1. The only question to be debated is the focus group selected by the authors. How authors selected, or in other words, what were the criteria of selection for the n=522 user sample? How can we know if this population is representative (good enough) for the study? Usually n=522 would be considered as an impressive number of participants yet, it seems there is no limitation when considering people blogging in the social media... I think this issue together with the algorithm proposed for word extraction could be discussed a little bit further in the section 4.1.

Answer: we used the Sina blogs of 522 persons to determine which words are keywords in the stock market. When analyzing the relationship between keywords and the index, we analyzed Sina microblogs of **all** Sina microblog users.

We used 522 persons because: 1) We are not able to download the microblogs of all users. Thus, the demographic characteristics of the overall Sina Weibo, such as language styles and currently hot topics, cannot be obtained by investigating the entire data space of Sina Weibo. 2) The 522 persons are celebrities in the financial industry. They are opinion leaders and Sina blog only lists these 522 persons. We have explained these two reasons in the end of Section keyword selection (Section 4.2).

The text segmentation algorithm is moved to Section 4.1. Before we introduce how we collect data we introduce the segmentation algorithm as a tool first. We think this way makes reader clearer.

2. The 'recent work' section seems to be a thorough review on the topic, yet I think one of the following papers could be additionally referred to, since it is a good example of similar, yet a bit different methodological approach and similar, yet different (US stock market) application as well as similar, yet a bit different data source (twitter).

Answer: we have reviewed recent literature, including the paper figured out by the reviewer.

3. Some minor clarification should be done in regard to the following issues:

page 9.

Model 4 is revoked at the end of section 5.3 and at the beginning of section 5.4. It is difficult to follow if authors refer to modified equation (4)?

Actually all the equations could be referred to within the body of text in order to clarify the logic flow of the paper.

Answer: A stepwise variable selection procedure is applied to Model 4. After this procedure we reduce the number of variables to 16, obtain the coefficients of 16 variables and the residuals (observations of *εd*). *εd* is modeled by Model 6. We fit Model 6 using , obtaining *p, q, d*. Finally, Model 4 and Model 6 are merged and fitted using data of microblogs. In this process, *p*, *q*, *d* are known. We have revised our writing.

page 10 - Model 4 and Model 7 are mentioned but the context is missing. Where are the subsequent Models 5 and 6?

page 14 - we are back with Model 4 (section 6.3) and then with Model 7 (section 6.4)

Answer: Model 7 actually refers to Model 6, we wrongly referred.

page 17 - there are results for Model 7 and 8 (is it referring to Equation (8)????)

Answer: they refer to ARIMA(0,0,1) and ARIMA(3,0,2).

4. I believe the paper could be structured a bit more in a classical way (i.e. step by step - introduction, methods, experimental settings, results, discussion - instead of jumping from results back to more theory of the models again and to new results afterwards again) in order to give the reader chance to follow the results and conclusions easier.

Answer: we have reorganized the paper. Results are moved from each method to a whole section. Discussion is moved to the end of experimental results.

The literature review is broad, but not much papers published in recent years is on the list of references.

Actuall only few (3?) papers cited are published after 2010. Please take a look below and maybe search of other related work from recent years again.

Romanowski A., Skuza M. (2017) Towards Predicting Stock Price Moves with Aid of Sentiment Analysis of Twitter Social Network Data and Big Data Processing Environment. In: Pełech-Pilichowski T., Mach-Król M., Olszak C. (eds) Advances in Business ICT: New Ideas from Ongoing Research. Studies in Computational Intelligence, vol 658. Springer, Cham

<https://link.springer.com/chapter/10.1007/978-3-319-47208-9_7>

Answer: we have reviewed recent literature, including the paper figured out by the reviewer.