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 have moved all the experimental results to Section 6.

3. We have reviewed recently published papers in Section 2.

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

Regards,

Authors

Reviewer #1

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.

Answer: the original number of variables is 130. We have applied a stepwise procedure to Model 4 to reduce number of variables from 130 to 16 (See Section 6.2.2). The square root of VIF indicates how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other predictor variables in the linear model. VIF does not mean the frequency of each keyword. The smaller VIF is, the less severity of [multicollinearity](https://en.wikipedia.org/wiki/Multicollinearity) is. If VIF<10, we reckon that there is no [multicollinearity](https://en.wikipedia.org/wiki/Multicollinearity) (similar to the correlation). In table 3, all VIF values are less than 10, suggesting that correlation between variables is weak.

In terms of variable reduction, we apply stepwise procedure to Model 4. The reasons that we didn’t use PCA are as follows:

1) We want to know the relationship between variables (keywords) and the index. If we use PCA, the variables are replaced by components.

2) We conducted PCA by SPSS but the result is not good. PCA extracts 32 components from 130 variables based on the condition that eigenvalues greater than 1 (See Table 1). 32 variables are more than the result of 16 variables by stepwise procedure. What’s more, we tried to extract meanings of 32 variables but failed. The variables for each principal component have no obvious relationship and there is no unified financial explanation to describe these variables (See Table 2).

Table 1: Total Variance Explained

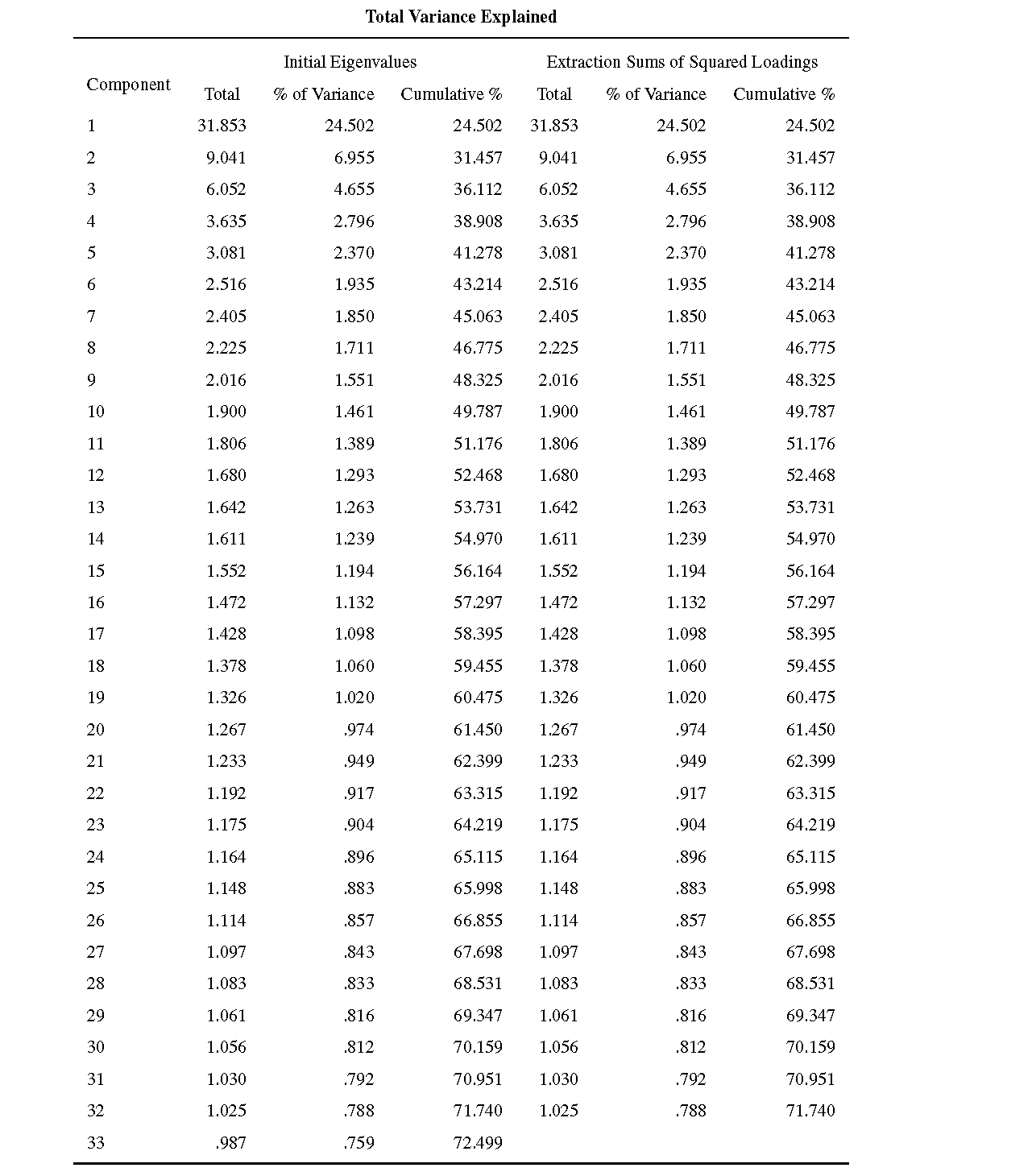


Table 2: Component Score Coefficient Matrix

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Zscore(x\_i)  Factor | Zscore(x1) | Zscore(x2) | Zscore(x3) | Zscore(x4) | …… | Zscore(x130) |
| 1 | .024 | .014 | .018 | .021 | …… | .010 |
| 2 | -.015 | .053 | .049 | -.021 | …… | .020 |
| 3 | -.001 | .004 | -.018 | .077 | …… | .060 |
| 4 | .019 | -.042 | -.054 | -.023 | …… | -.047 |
| 5 | -.018 | .041 | -.047 | .010 | …… | .014 |
| 6 | .029 | .052 | -.061 | -.022 | …… | .013 |
| 7 | -.043 | -.051 | .078 | -.059 | …… | .090 |
| 8 | .009 | .052 | -.002 | -.017 | …… | .063 |
| 9 | -.045 | -.031 | -.049 | -.002 | …… | -.041 |
| 10 | -.033 | .010 | .040 | -.019 | …… | .000 |
| 11 | .044 | .002 | -.047 | .066 | …… | .110 |
| 12 | .000 | .025 | .020 | .018 | …… | -.094 |
| 13 | .023 | .039 | -.014 | -.030 | …… | -.036 |
| 14 | -.029 | -.086 | .062 | .002 | …… | -.051 |
| 15 | .000 | -.096 | .050 | -.007 | …… | -.038 |
| 16 | -.043 | .003 | -.026 | -.011 | …… | .092 |
| 17 | -.086 | .003 | -.059 | -.083 | …… | -.042 |
| 18 | -.006 | -.112 | -.004 | .019 | …… | .045 |
| 19 | -.022 | .014 | .025 | -.068 | …… | -.042 |
| 20 | .005 | -.095 | .005 | -.060 | …… | .024 |
| 21 | .080 | -.181 | .002 | -.021 | …… | -.078 |
| 22 | .006 | .031 | -.082 | -.071 | …… | -.009 |
| 23 | .098 | .083 | .037 | -.066 | …… | -.123 |
| 24 | .042 | .059 | -.096 | .052 | …… | -.036 |
| 25 | .006 | .099 | -.008 | .055 | …… | .037 |
| 26 | .067 | -.053 | .036 | .031 | …… | -.017 |
| 27 | -.009 | .024 | .024 | .061 | …… | -.062 |
| 28 | -.017 | -.020 | .024 | -.039 | …… | -.010 |
| 29 | .022 | .087 | -.026 | .007 | …… | .060 |
| 30 | -.031 | .068 | .044 | .001 | …… | .084 |
| 31 | .011 | .040 | .028 | -.042 | …… | -.019 |
| 32 | -.015 | -.003 | .024 | .004 | …… | .138 |

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.