Fine-Grained Analysis of Financial Tweets

Chung-Chi Chen[1], Hen-Hsen Huang[1], and Hsin-Hsi Chen[1,2]
[1] Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
[2] MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
{cjchen, hhhuang}@nlg.csie.ntu.edu.tw; hhchen@ntu.edu.tw

ABSTRACT

This paper decribes our experimental methods and results in FiQA 2018 Task 1. There are two subtasks : (1) to predict continuous sentiment score between -1 to 1, and (2) to determine which aspect(s) are related to the content of financial tweets. First, we propose a preprocessing procedure for decomposing financial tweets. Second, we collect over 334K labeled financial tweets to enlarge the scale of the experiments. Third, the sentiment prediction task is separated into two steps in this paper, i.e., (1) bullish/bearish and (2) sentiment degree. We compare the results of the CNN, CRNN and Bi-LSTM models. Besides, we further combine the results of the best models in both steps as the model of subtask 1. Finally, we make an investigation of aspects in depth, and propose some clues for dealing with the 14 aspects.

CCS CONCEPTS

• **Information systems→ Information retrieval→ Retrieval tasks and goals→ Sentiment analysis**

KEYWORDS

Financial tweet; sentiment analysis; opinion mining

ACM Reference format:

C.C. Chen, H.H. Huang, and H.H. Chen. 2018. Fine-Grained Analysis of Financial Tweets. In *The 2018 Web Conference Companion (WWW 2018), April 23-27, 2018, Lyon, France*, ACM, New York, NY, 7 pages. DOI: https://doi.org/10.1145/3184558.3191824

1 INTRODUCTION

FinTech (financial technology) is one of the hot topics recently. Adopting the maturity technology to solve the problem or improve the service is one of the popular trends in finance domain. For the NLP challenges, there are plenty of savage resources in this domain. The resources can be classified into official documents, financial statements, news, and social media messages. The social media data are different from the others in the informal writing style. It adds noise and increases the difficulty in data analytics.

WWW '18 Companion April 23-27, 2018, Lyon, France.

ACM ISBN 978-1-4503-5640-4/18/04.
DOI: https://doi.org/10.1145/3184558.3191824

On the other hand, we can still find some clues that are useful to analyze financial social media data. For example, cashtag is a common tag in these data. The cashtag marks the mentioned instrument's ticker by putting "$" in the front of ticker like $AAPL, which is the cashtag of Apple Inc.'s stock. This convention saves us from named entity recognition challenge.

Sentiment analysis is popular in finance domain in this decade, and has been considered as one of useful signals for predicting the price movement of instruments. However, till SemEval-2017 Task5 dataset [3], none of researches had discussed the continuous sentiment scores of financial social media data, showing that there are many unvanquished challenges with financial data.

Classifying opinions into several aspects is one of the important tasks in opinion mining. In order to understand the topic of each comment, it is necessary to define some important aspects. The classical example of this task is hotel comments. Customer's comments may focus on different aspects like service, environment, and so on. We will make further discussion in Section 4.

Our constributions are three-fold as follows. First, we propsoe a fine-grained preprocessing procedure for financial social media data. Second, we propose a two-step model to predict the continuous sentiment score. Third, we investigate some rules for different aspects.

The remainder of the paper is organized as follows. Section 2 describes the data. Section 3 presents our methods in detail. Section 4 proposes some rules for aspects. Section 5 shows the experimental results and makes further discussions. We discuss the predictability of the proposed model in a news headline dataset in Section 6. In Section 7, we compare the experimental results to the related work. Finally, Section 8 concludes the remarks.

2 DATA

2.1 Training Data

Total 675 training instances are provided in FiQA 2018 Task 1. Fig. 1 shows the distribution of sentiment scores. There are 440 positive, 1 neural, and 234 negative instances. Since there is only one neural instance, we just classify the data into bullish and bearish in this paper. Fig. 2 shows the distribution of the degrees of sentiments.

The average of the sentiment degrees is about 0.41. There are total 83 fine-grained aspects in the training data. These aspects are further classified into 4 levels. Since the goal of this open

challenge is to predict the aspect at level 2, we show the frequency information of level 2 aspects in Table 1. There are 21 aspects at level 2. Total 56.15% of instances are annotated as "Price Action" aspect.

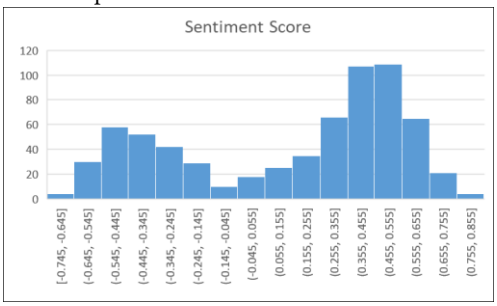

Figure 1: Distribution of sentiment scores.

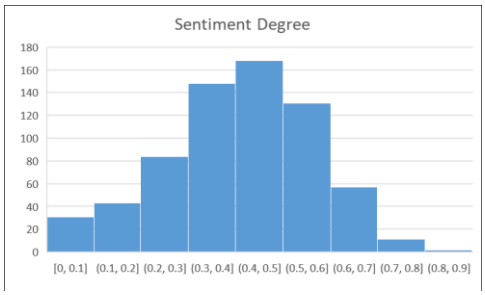

Figure 2: Distribution of degrees of sentiment.

Table 1: Frequency of aspects at level 2.

Aspect	Freq.	Aspect	Freq.	Aspect	Freq.
Price Action	379	Dividend Policy	13	Fundamentals	5
Technical Analysis	94	Options	12	Market	3
Coverage	41	M&A	11	Volatility	3
Risks	28	Rumors	6	Insider Activity	2
Financial	23	Strategy	6	Reputation	2
Sales	19	Stategy	6	Conditions	1
Signal	15	Legal	5	Regulatory	1

2.2 Collected Data

We collected more than 334K labeled financial tweets from StockTwits, a Twitter-like financial social media for investors sharing their ideas. All tweets were annotated as bullish or bearish by the original writers. This dataset avoids the misunderstanding circumstance between annotator and original writer, and could be assumed as a high quality dataset.

Furthermore, we create a sentiment degree for each instance in this dataset with the provided training data. If an instance in the collected data has the same token with an instance in the provided training data, the sentiment degree of this training data becomes a candidate of the sentiment degrees of the instance in the collected dataset. Finally, we average all candidates of a same instance as its sentiment score. Fig. 3 shows the distribution of the created sentiment degrees. About 299K instances get the imitated sentiment degrees.

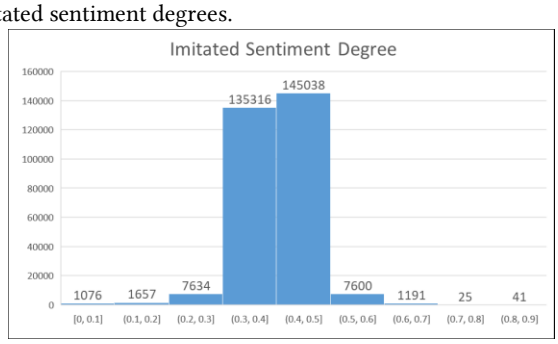

Figure 3: Distribution of degrees of sentiment in the collected data.

3 METHODS OF SENTIMENT ANALYSIS

The fine-grained sentiment analysis task is composed of two steps in this paper. We first predict the bullish/bearish sentiment of a tweet, and then predict its sentiment degree. The preprocessing procedure for financial tweets, the models for bullish/bearish prediction, and the models for sentiment degree prediction will be described in the following subsections.

3.1 Preprocessing Procedure

One financial tweet may be composed of words, cashtags, user id, numbers, URL, hashtags and emojis. Both hashtags and emojis are regarded as words in this paper. Only 3.12% of financial tweets contain at least one hashtag in our collected dataset and 4.67% of financial tweets contain at least one emoji. First, we replace user ids, numbers, cashtags and URLs by "ID", "NUM", "TICKER" and "URL". Second, we remove stop words and punctuation marks. Finally, we transform the remaining tokens into lowercase.

Due to the 140-character limitation of a Twitter post, users focus on a few points in one tweet. With this preprocessing procedure, only the opinionative part will be remained sometimes. (T1-1) is the original tweet, and (T1-2) is the result after preprocessing. The words "looking good" in (T1-2) could help us determine the sentiment of the tweet, i.e., bullish or bearish, and the word "calls" could provide the clue for "Options" aspect. Furthermore, for some tweet like (T2-1), only "longed" will be remained as the opinionative word.

(T1-1) *$MOS looking good here at $58.65. Calls are active in this month and weekly*
(T1-2) *'TICKER', 'looking', 'good', 'NUM', 'calls', 'active', 'month', 'weekly'*
(T2-1) *longed $AMZN 300 @ 189.82*
(T2-2) *'longed', 'TICKER', 'NUM', 'NUM'*

3.2 Bullish/Bearish Sentiment Models

Convolution neural network (CNN), bidirectional long short-term memory (Bi-LSTM), and convolution recurrent neural network (CRNN) are adopted for classifying a tweet into bullish or bearish. The first three layers of the CNN model are one

embedding layer (300-dimensions), one convolution layer (filter and kernel size are 64 and 8), and one max pooling layer. The difference between CNN and CRNN is that max pooling layer in CNN model is replaced by bidirectional LSTM layer with 32-dimension output. The first layers of Bi-LSTM is embedding layer as used in CNN, and the second layer is bidirectional LSTM layer with 32-dimension output. The remainder of each model is the same as follows: One densely-connected layer (200-dimension), one dropout layer (dropout rate is 0.5), one rectified linear unit layer, and the softmax output layer.

3.3 Sentiment Degree Models

The same models used in bullish/bearish classification task are also adopted to predict the degree of sentiment except that we replace softmax output layer by sigmoid.

4 RULES OF ASPECTS

First, we attempt to sort out some meaningful words for "Price Action" aspect. The chi-squared test is adopted to compute the significance [7]. Table 2 shows some clues to classify "Price Action" aspect from the other aspects. Only the words appearing more than 5 times are taken into account. Some words in "Price Action" class are about the action of investor (long and short), and some words are used to describe the price action (back, bounce, and breaking). Individual investors often share the URL of news or analysis report as the reference of their tweets. However, only 22.43% of instances annotated as "Price Action" aspect contain URL, and 55.41% of instances in the other aspects contain URL.

Table 2: Clues to separate "Price Action" from the other aspects.

Price Action		Others	
Word	Chi-squared	Word	Chi-squared
long	77.95	URL	234.20
short	68.05	dividend	46.18
higher	53.96	buy	37.10
shorts	43.96	chart	36.15
close	40.24	rsi	33.86
back	39.65	sales	33.86
hod	32.23	rating	30.78
high	31.81	breakout	27.83
bounce	29.64	technical	27.70
looks	28.68	sell	27.55
breaking	27.74	target	25.36
highs	27.74	revenue	24.62
lows	26.37	growth	22.42

Second, two main tools, i.e., technical indicators and chart, are usually used by investors in technical analysis. Thus, the name list of the technical indicators like moving average (MA) and relative strength indicator (RSI) and the name list of chart pattern like "double top" and "head and shoulders" could be the clues of "Technical Analysis" aspect.

Third, 37 instances in "Coverage" aspect are about the analysis rating and credit rating. The keywords sort out by chi-squared test are shown in Table 3. Furthermore, because 23 instances of "Risks" aspect describe the same news (recall of Tesla Model X), there does not exist a general rule for this aspect. For "Financial" aspect, only the word, revenue, shows the significant tendency toward this aspect. Call, put and option are the keywords of "Options" aspect. Merger and M&A are the keywords of "M&A" aspect. The keywords of "Sales", "Signal", "Dividend Policy", "Rumors", "Signal", "Conditions" and "Regulatory" aspect are the same as their aspect names, particularly, the keyword of "Dividend Policy" is dividend.

In sum, we use the keywords sort out by chi-squared test to classify a tweet into 21 aspects. Because there are some crucial clue(s) for the aspects with few instances, we check the keyword(s) of these aspects first. In other words, the instances that do not satisfy any rules will be classified into "Price Action" aspect.

Table 3: Keywords for the "Coverage" aspect.

Word	Chi-squared	Word	Chi-squared
rating	283.43	outperform	141.66
pt	198.36	ratings	141.66
analyst	170.01	upgrades	141.66
downgraded	141.66	research	116.4

5 EXPERIMENTS

5.1 Experiment Setup

For the bullish/bearish classification task, we use 40,000 bullish instances and 40,000 bearish instances in the collected dataset as the training set, and use 5,000 bullish and 5,000 bearish instances as the test set. The validation set is 10% of instances in the training set. The experimental results are the average of 100 times bootstrapping. The 675 instances provided by FiQA-2018 are used as the second test set.

To keep the distribution as the training data, we separate the collected data into ten parts by the nth percentile of the provided training data, where n is 10, 20, ..., and 90. Fig. 4 shows the number of instances of each part. We use 1,000 instances of each part as the training data, i.e., total 10,000 training data, and 100 instances of each part as the test data. The setting of validation and bootstrapping is the same as the bullish/bearish classification task.

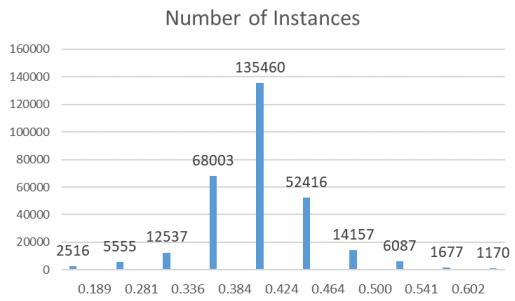

Figure 4: Number of instances in the nth percentile.

Accuracy is used to evaluate the result of the bullish/bearish classification task, and mean-squared error (MSE) and r-squared (R2) are used to evaluate the results of the sentiment degree prediction and final prediction. Precision, recall, and F1-score are adopted to evaluate the experimental results of the aspect classification task. Keras (https://github.com/keras-team/keras) is adopted in our experiments.

Table 4: Accuracy of the bullish/bearish prediction task. (%)

		CNN	CRNN	Bi-LSTM
CP	Set 1	49.94	49.99	49.86
	Set 2	50.33	49.57	48.56
FP	Set 1	71.43	71.47	71.58
	Set 2	75.12	74.86	74.39
OP	Set 1	71.23	71.14	71.45
	Set 2	76.86	76.39	76.85

5.2 Experimental Results of Sentiment Analysis

First, the results of the bullish/bearish sentiment prediction task are shown in Table 4. We compare not only the results of the models, but also those of different preprocessing procedures. Set 1 is the 10,000 test instances of the collected dataset, and Set 2 is the 675 instances in the provided dataset.

- Coarse-preprocessing (CP): we only remove the punctuation of a tweet.
- Fine-preprocessing (FP): we use the preprocessing procedure shown in Section 3.1.
- Opinion-preprocessing (OP): we remove "ID", "NUM", "TICKER", and "URL" from the results of FP.

The Bi-LSTM model with FP and the CNN model with OP get the best in Set 1 and Set 2, respectively. In Set 2, the CNN models with different preprocessing procedures outperform the other models. Thus, we adopt the CNN model with OP as our final model for the bullish/bearish classification task. With the proposed preprocessing procedure, FP, we can get more than 24% improvement, compared with CP, when the CNN, CRNN and Bi-LSTM models are adopted. Comparing OP with FP, the CNN, CRNN, and Bi-LSTM models with OP get about 1.5% improvement in accuracy. Therefore, OP is adopted as our preprocessing procedure for final submission.

Table 5 shows the results of the sentiment degree prediction task. The CNN model performs the best in Set 1, and Bi-LSTM is the best model for Set 2 with the lowest MSE and the highest R2. According to the experimental results, we use the Bi-LSTM model to predict the sentiment degree of final submission.

Finally, we compare our two-step model with one-step models. Our two-step model (CNN-Bi) predicts bullish/bearish by the CNN model with OP, and predicts sentiment degree by the Bi-LSTM model. Table 6 shows the experimental results. The test data in this experiment is all 675 tweets in the provided dataset. CNN-Bi outperforms others with more than 15% improvement in MSE.

Table 5: Results of the sentiment degree prediction task. (%)

		CNN	CRNN	Bi-LSTM
MSE	Set 1	1.42	2.22	1.57
	Set 2	2.05	2.31	1.94
R2	Set 1	41.17	8.16	35.21
	Set 2	19.06	8.55	23.31

Table 6: Results of fine-grained sentiment prediction. (%)

	CNN-Bi	CNN	CRNN	Bi-LSTM
MSE	30.67	47.39	48.67	46.22
R2	-79.05	-176.63	-184.06	-169.77

5.3 Experimental Results of Aspects

Table 7 shows the precision, recall, and F1-score of each aspect. The micro- and macro-averaged F1-score are 75.41% and 50.38%, respectively. If only top 10 frequent aspects are taken into consideration, the micro- and macro-averaged F1-score are boosted to 78.74% and 63.23%, respectively.

Table 8 shows the confusion matrix among aspects. Some instances in "Price Action" aspect are classified into "Technical Analysis" and "Options" aspects. Some instances for error analysis are shown in Table 9. As mentioned in Section 4, technical analysis is one of the popular methods used to predict the movement of price. Most of investors using technical analysis usually establish their views based on two kinds of indicators, i.e., technical indicators and chart patterns.

Table 7: Results of aspect classification task. (%)

Aspect	P	R	F1	Aspect	P	R	F1	Aspect	P	R	F1
Price Action	0.76	0.85	0.80	Dividend Policy	1.00	1.00	1.00	Fundamentals	0.00	0.00	0.00
Technical Analysis	0.68	0.74	0.71	Options	0.33	1.00	0.50	Market	0.00	0.00	0.00
Coverage	0.94	0.73	0.82	M&A	1.00	0.18	0.31	Volatility	1.00	1.00	1.00
Risks	1.00	0.82	0.90	Rumors	0.67	0.67	0.67	Insider Activity	0.00	0.00	0.00
Financial	0.71	0.22	0.33	Strategy	0.00	0.00	0.00	Reputation	0.00	0.00	0.00
Sales	0.91	0.53	0.67	Stategy	0.00	0.00	0.00	Conditions	1.00	1.00	1.00
Signal	0.87	0.87	0.87	Legal	0.00	0.00	0.00	Regulatory	1.00	1.00	1.00

Table 8: Confusion matrix of aspect classification task.

True \ Predict	Price.	Tech.	Cov.	Risks	Fin.	Sales	Sig.	Div.	Opt.	M&A	Rum.	Vol.	Con.	Reg.
Price Action	322	32	1	0	0	0	1	0	22	0	1	0	0	0
Technical Analysis	24	70	0	0	0	0	0	0	0	0	0	0	0	0
Coverage	10	0	30	0	0	0	1	0	0	0	0	0	0	0
Risks	3	0	1	23	0	0	0	0	1	0	0	0	0	0
Financial	17	0	0	0	5	1	0	0	0	0	0	0	0	0
Sales	6	1	0	0	2	10	0	0	0	0	0	0	0	0
Signal	2	0	0	0	0	0	13	0	0	0	0	0	0	0
Dividend Policy	0	0	0	0	0	0	0	13	0	0	0	0	0	0
Options	0	0	0	0	0	0	0	0	12	0	0	0	0	0
M&A	8	0	0	0	0	0	0	0	0	2	1	0	0	0
Rumors	2	0	0	0	0	0	0	0	0	0	4	0	0	0
Volatility	0	0	0	0	0	0	0	0	0	0	0	3	0	0
Conditions	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Regulatory	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 9: Instances for error analysis.

	Tweet	Truth	Prediction
T3	$AAPL **double bottom** could be in, keep in mind	Price Action	Technical Analysis
T4	$RAD All my **charts** are flashing oversold.	Price Action	Technical Analysis
T5	$CRM Sep 40 **calls** are +35% since entry #BANG http://stks.co/deDm	Price Action	Options
T6	$UVXY **Put** the chum out there at key support then next level down - careful	Price Action	Options
T7	BULLISH Engulfing of COCA COLA: http://stks.co/fYCo $KO	Technical Analysis	Price Action
T8	$GOOG Testing the **200 day** after some consolidation http://stks.co/h0cPJ	Technical Analysis	Price Action
T9	$AAPL AAPL: Gundlach Slams iPad mini, **Sees Downside to** $425.	Coverage	Price Action
T10	$ISRG PT **raised** to $700 from $640 at Leerink - keeps Outperform rated	Coverage	Price Action
T11	$AAPL Beat the estimates. Will still go down on lack of new products.	Financial	Price Action
T12	Daily Mail owner considering Yahoo bid $yhoo ,up 2,05% https://t.co/extZr1riyP	M&A	Price Action

Double bottom in (T3) is the name of one chart pattern, and the writer of (T4) concluded his/her analysis result for $RAD based on charts, which could be seen as one kind of technical analysis. Furthermore, in (T5), writer described the "Price Action" of the "Options" of $CRM. Accordingly, we think some of our prediction could be seen as the correct answers.

(T6) shows the necessity of disambiguation, because the "put" in (T6) is not the option of $UVXY. (T7) shows that referring to outside reference is a necessary procedure, because we cannot get the technical analysis information from the context of this tweet. "200 day" in (T8) may be the abbreviation of 200-days moving average. It shows the challenge of analyzing the informal social media data. (T9) and (T10) indicate the need to understand the meaning of the sentence.

Because some aspects have less than 10 instances, we could just sort out one or two keywords by chi-squared test. (T11) is one of the instance for "Financial" aspect. (T12) shows the link could provide some information again.

5.4 Experimental Results in Official Test Set

There are 99 instances in the test set of this open challenge. The following results are provided by organizers. In continuous sentiment score prediction task, the CNN-Bi model gets the MSE of 30.58%, which is lower than the results shown in Table 6 (30.67%), and the R2 is -166.69%. In the aspect prediction task, we achieve the accuracy of 75.76%, and get the second place. That shows the usefulness of our statistics-based keyword extraction. Since some aspects have few training data, we could not get enough information for these aspects. Therefore, these aspects get 30.07%, 26.78%, and 28.32% in precision, recall and F1-score respectively.

6 DISCUSSION

In this section, we use the same models for financial tweet to predict the sentiment score of news headline. The reason why we are wondering the performance of the proposed model in news headline is that some writers of the tweets may copy the news headline into their posts, and add some comments about them. (T13) is an example for this case.

(T13) *Reuters: Green Mountain revenue misses, shares plunge http://stks.co/13mW > $GMCR prints 43.80, market in a foul mood, bad day to disappoint*

Thus, we assume the information of news headline could be learned when the model is trained with financial social media data. There are 438 headlines as the training data in this open

challenge task 1. The distribution of sentiment scores is shown in Fig. 5, and the distribution of sentiment degrees is shown in Fig. 6. Because the news is expected to describe an event objectively, the average of the sentiment degree is 0.34, which is lower than the sentiment degree of tweets. Furthermore, the sentiment degrees of 79.45% of news headlines are lower than 0.5.

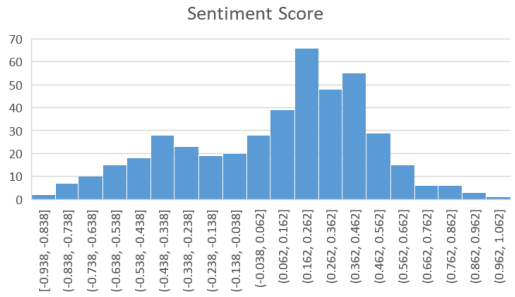

Figure 5: Distribution of sentiment scores – news headline.

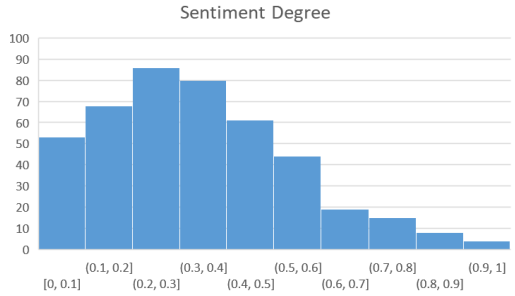

Figure 6: Distribution of degrees of sentiment – news headline.

The same as the experiments in Section 5.2, we use the average MSE and R2 of 100 times bootstrapping to evaluate the performance of each model. The experimental results are shown in Table 10. The CNN-Bi model gets the best in this experiment, and gets more than 7.5% improvement of MSE than the other models. Comparing the experimental result of news headline with that of social media data, the MSE of headlines is lower than that of social media data when using the CNN-Bi model, even we do not use the headlines as training data directly.

Table 10: Results of the fine-grained sentiment prediction on news headline. (%)

	CNN-Bi	CNN	CRNN	Bi-LSTM
MSE	26.45	36.89	39.57	34.71
R2	79.28	-149.99	-168.20	-135.22

7 RELATED WORK

Barnes et al. [1] compared the performance of several models in different sentiment analysis dataset, and recommended long short-term memory (LSTM) and Bi-LSTM models for fine-grained sentiment analysis. To compare the performance of LSTM model in this dataset with the other models, we also use the 675 provided instances to experiment with the LSTM model. The setting of the LSTM model is similar to the Bi-LSTM model, only change the bidirectional LSTM layer into LSTM layer. LSTM model performs worst in predicting bullish/bearish, and only gets about accuracy of 50% with different preprocessing procedures. The LSTM model gets 2.55% MSE in sentiment degree experiment, which is worse than the other models. MSE of the LSTM model predicting fine-grained sentiment score directly is 46.52%, which is better than the CNN and CRNN model, but is worse than the Bi-LSTM model and the CNN-Bi model. In sum, the experimental results show that the LSTM model is not suitable for the two-step model, but can perform better than the CNN and CRNN models when predicting fine-grained sentiment scores of financial social media data.

Few researches about financial social media data discuss the details of preprocessing process. Li and Shah [4] provided the sentiment-oriented word vector with their special preprocessing process. However, they did not analyze the impact of their preprocessing process. In this paper, we show the difference of experimental results between both coarse- and fine-preprocessing in our setting. Furthermore, the opinion-preprocessing, which remove many common terms (expected with neutral sentiment) in financial social media data, is also compared. The improvement of the experimental results shows the promising of the proposed preprocessing process.

Additional information can be considered in the future. Firstly, we do not use any lexicons in our experiments. The sentiment dictionaries for financial textual data are expected to be useful for sentiment analysis. Chen et al. [2] provided NTUSD-Fin, a dictionary for market sentiment analysis in financial social media. They showed the difference between general sentiment and market sentiment, and compared their dictionary with Loughran and McDonald [5]. The experimental results show the usefulness of their dictionary. Secondly, numerals always contain important information in financial textual data. Murakami et al. [6] generated market comments with the time-series price data. The experimental results show that numerals contain crucial information in their task. Therefore, numeral information can be taken into account in sentiment analysis task in the future.

8 CONCLUSION

We propose a two-step model for fine-grained sentiment analysis, and construct some rules for aspect classification. According to the experimental results, separating the fine-grained sentiment prediction into two steps, i.e., bullish/bearish and sentiment degree, can improve the performance. The effect of the preprocessing procedure is also shown. Besides, there may exist some equivocal instances of the aspect classification task in the provided dataset, and some cases are shown in Section 5.3. Furthermore, we point out some challenges as the future work of analyzing financial social media data, including (1) disambiguation, (2) outside reference, and (3) informal abbreviation.

ACKNOWLEDGMENTS

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-107-2634-F-002-011-, MOST-106-3114-E-009-008-, MOST-106-2923-E-002-012-MY3, and MOST-105-2221-E-002-154-MY3.

REFERENCES

[1] Barnes, J., Klinger, R., & Walde, S. S. I. 2017. Assessing State-of-the-Art Sentiment Models on State-of-the-Art Sentiment Datasets. *In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity*, pages 2–12

[2] Chen, C. C., Huang, H. H., & Chen, H. H. 2018. NTUSD-Fin: A Market Sentiment Dictionary for Financial Social Media Data Applications. In *Proceedings of the 1st Financial Narrative Processing Workshop* (FNP 2018).

[3] Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., & Davis, B. 2017. Semeval-2017 task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 519-535

[4] Li, Q., & Shah, S. 2017 Learning Stock Market Sentiment Lexicon and Sentiment-Oriented Word Vector from StockTwits. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pages 301–310

[5] Loughran, T., & McDonald, B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 2011, 66.1: 35-65.

[6] Murakami, S., Watanabe, A., Miyazawa, A., Goshima, K., Yanase, T., Takamura, H., & Miyao, Y. 2017. Learning to Generate Market Comments from Stock Prices. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. pages 1374-1384

[7] Oakes, M., Gaaizauskas, R., Fowkes, H., Jonsson, A., Wan, V., & Beaulieu, M. 2001. A Method based on the Chi-square Test for Document Classification. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 440-441