NTUSD-Fin: A Market Sentiment Dictionary for Financial Social Media Data Applications

Chung-Chi Chen, Hen-Hsen Huang, Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
{cjchen, hhhuang}@nlg.csie.ntu.edu.tw; hhchen@ntu.edu.tw

Abstract

Extracting information in textual data for further applications is one of the popular topics in financial domain in last decade. Although there exist some dictionaries for news and financial reports, few dictionaries focus on financial social media data. This paper constructs a market sentiment dictionary based on more than 330K labeled posts crawled from financial social media. There are 8,331 words, 112 hashtags and 115 emojis in our dictionary. The statistic results shows the difference between the sentiment and the market sentiment of the investors. Furthermore, the comparison of (1) general sentiment analysis and market sentiment analysis, and (2) the market sentiment of social media data and formal reports are discussed with the constructed dictionary. We find that some neutral words in general sentiment dictionary should be considered as the bullish/bearish words. The experimental results of our dictionary and that of the dictionary for financial formal documents show the usefulness of our dictionary in financial social media application.

Keywords: Financial dictionary, social media, sentiment analysis

1. Introduction

Textual data has been regarded as an important source when analyzing economic and financial phenomena. There are three major kinds of text resources, including official documents, news, and social media. They stand for different viewpoints of the same event. Official documents, published by government or company, provide the insiders' opinions. News are expected to propose objective opinons of the incident. Social media data, which are the most popular text data recently, contain plenty of crowd views. The informal vocabulary and syntax make social media data quite different from the other documents and make the analysis tasks more challenging.

Sentiment analysis is one of the hot topics in financial domain in last decade. Many empirical results show that textual sentiment is highly correlated to different aspects of financial phenomena. Bollen et al. (2011) show the correlation between tweet mood and Dow Jones Index. Sul et al. (2014) find that the emotion of tweets for certain company significantly affects its stock price. Furthermore, El-Haj et al. (2016) extend the sentiment analysis for sentence in PEAs into internal or external attribute, and compare the performance of human and machine.

Sentiment dictionary plays a crucial role in both dictionary-based and machine learning approaches. There are dictionaries for official documents (Loughran and McDonald, 2011) and news (Huang et al., 2013), but few dictionaries are available for social media data. Chen et al. (2014) present a dictionary for social media data applications. However, it is constructed based on the word list of Loughran and McDonald (2011), which is collected for formal documents like 10-K (annual report of company). Furthermore, the posts in the Seeking Alpha platform used by Chen et al. (2014) are dissimilar to nowaday Twitter-like social media posts. Li and Shah (2017) construct a domain specific sentiment lexicon with StockTwits data, and propose a state-of-the-art model for sentiment analysis. In this paper, we will construct a market sentiment dictionary based on over 330K labeled posts craweled from financial social media. Besides words, hashtags and emojis are also included.

In some topics, writers' sentiments are strongly related to their opinions. For example, in the review of a hotel, positive/negative sentiment of a customer is associated with good /bad opinion of the hotel. In constrast, the market sentiment (bullish/bearish) of an investor may not be derived from the positive/negative sentiments of the investor directly. More details will be discussed in Sections 4 and 5.

The sturcture of this paper is organized as follows. Details of the dataset are specified in Section 2. The methods to construct a sentiment dictionary are introduced in Section 3. Overview of the dictionary is shown in Section 4. We discuss some findings in Section 5. Finally, Section 6 concludes the remarks.

2. Fianacial Social Media Data

2.1 Data Source

StockTwits is a Twitter-like social media for investors to share their information and opinions of the market or a certain company. Figure 1 shows the graphical user interface (GUI) provided by StockTwits. The same as Twitter, it limits the length of each post to 140 characters. Under this limitation, users have to focus on a few main points they want to share in the posts. Users usually use cashtag ($ before ticker) to mark the instruments they mention. For instance, $MSFT stands for the security of Microsoft Corporation. In particular, the bullish and bearish bottoms allow users to label their market sentiment in the post.

Figure 1: GUI of Stocktwits

2.2 Dataset

From StockTwits, we crawled 334,798 labeled posts from 13,059 users. (The detail about StockTwits API please refer to the related documents [1].) In total there are 75,376 unique words, 3, 041 unique hashtags, and 451 unique emojis in the collection. The distribution of these elements for both sentiments is shown in Table 1.

Since the bull market is much longer and more profitable than the bear market in history, it is reasonable that people tend to find the bullish targets. The other reason for the unbalance distribution may be that short stock is costly than long. Therefore, 95.35% of users had published the bullish posts, while only 44.67% of users had published the bearish posts.

	Bullish	Bearish
Post	289,416	45,382
User	12,452	5,834
Word	69,114	25,956
Hashtag	2,507	715
Emoji	427	174

Table 1 : Distribution of dataset.

2.3 Quality of the Dataset

The collected dataset in this paper is the large financial social media dataset labeled by the original writers. Compared with the datasets that are labeled by additional annotators, ours is advantageous in the consistency between the text meaning and the label since the writers would not misunderstand the meaning in the posts written by themselves. Therefore, it is reasonable to assume this dataset is a high quality dataset.

3. Methods

To mine the bullish/bearish sentiment tokens, we applied four methods including chi-squared test, collection frequency, pointwise mutual information, and a convolutional neural network classifier. Before that, we perform the data preprocessing as follows. First, stopwords, punctuations, digits, URLs, user ids, and tickers are removed from the posts. Second, all characters are converted to lowercase. Third, we remove the posts less than 2 words. For example, users may just post one cashtag and give a sentiment label. Finally, the tokens appearing less than *n* times are not taken into consideration, where *n* is set to 100 for words and 10 for hashtags and emojis. We do not perform word stemmings, beacuse we would like to maintain the original results and keep the largest flexibility for the uses of the proposed sentiment dictionary.

3.1 Chi-Squared Test

Chi-squared test is used to determine if there exist the difference between expected and observed frequency. It is adopted to decide whether the token should be remained in our dictionary with the confidence level set to 95%.

3.2 Collection Frequency-Inverse Document Frequency

Collection frequency(CF) is calculated as

$$cf_s(t, D_s) = log(1 + f_{t,D_S}) \quad (1)$$

,where t is one of the tokens in the list of words, hashtags or emojis, and s stands for a sentiment (i.e., bearish or bullish). D_s is a set of posts labeled as s, and f_{t,D_s} is the frequency of the token t appearing in D_s. Inverse document frequency (IDF) is the most common weighting scheme used to extract the keywords of documents.

$$idf_s(t, D_s) = log \frac{N_s}{|\{d \in D_s : t \in d\}|} \quad (2)$$

where N_s is the number of posts in D_s. Collection Frequency- Inverse Document Frequency (CFIDF) can be computed as follows.

$$cfidf(t, s) = tf_s(t, D_s) \times idf_s(t, D_s) \quad (3)$$

We can obtain the degree of importance of a token according to its CFIDF score in sentiment s.

3.3 Pointwise Mutual Information

Pointwise mutual information (PMI) is used to measure the dependence of the events. With PMI, we can observe how much the token t is correlated to the sentiment s.

$$pmi(t, s) = log \frac{p(t,s)}{p(t)p(s)} \quad (4)$$

In order to maintain the information of frequency, we use the probability of the tokens in the dataset to weight pmi as (5), where f_t is the frequency of the token t and T is the total number of tokens in the dataset.

$$wpmi(t, s) = \frac{f_t}{T} \times pmi(t, s) \quad (5)$$

3.4 Convolutional neural networks

Convolutional neural networks (CNN) is adopted to train the word embedding of each token in our dataset. The input is the text in a StockTwits post. The output of CNN model is the classification result, i.e., bullish or bearish, of a input post. The sturcture of the model is shown in Figure 2. The loss function is binary cross entropy, and we use Adam algorithm (Kingma and Ba, 2015) to optimize the parameters.

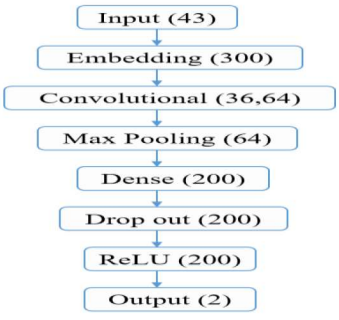

Figure 2: Structure and output size of CNN model

The word embedding scheme with neural network is widely used in natural language processing. We use the CNN model to train a classifier for market sentiment with our dataset, and use the output vector of embedding layer as the representation of each token. We calculate the cosine

[1] https://api.stocktwits.com/developers/docs

Bullish						Bearish					
Word		Hashtag		Emoji		Word		Hashtag		Emoji	
buy	14,489	stocks	202	😂	927	short	3,653	stocks	68	😂	184
today	13,191	sharkalerts	110	👍	518	going	2,094	noshamenate	40	💩	55
like	11,624	bitcoin	104	🚀	517	stock	2,075	sanofiwasright	34	🐻	35
go	10,959	cesium	102	😎	332	like	1,989	mnkdsecinvestigation	29	😎	26
get	10,829	trading	92	🙂	290	sell	1,897	forex	27	😹	24
going	10,203	bullish	71	😁	282	get	1,678	trading	26	🙂	21
back	9,768	stock	71	👀	242	lol	1,663	elliottwave	16	🙂	21
day	9,400	brachytherapy	63	🤔	233	today	1,626	earnings	15	😁	20
stock	8,590	tradesmart	60	💰	210	back	1,595	scambags	15	🤔	20
next	8,451	btfd	59	🤪	186	buy	1,592	markets	14	😝	17

Table 2 : Top 10 highly frequent tokens for sentiments *Bullish* and *Bearish*.

Bullish						Bearish					
Word		Hashtag		Emoji		Word		Hashtag		Emoji	
bully	86.74	stocks	80.19	🚀	82.75	bye	64.25	sanofiwasright	66.01	😂	60.78
dilly	86.74	sharkalerts	77.04	👍	82.23	ah	61.87	stocks	57.22	💩	56.64
bye	85.32	bitcoin	76.77	👀	82.14	junk	61.29	noshamenate	56.25	🐻	53.38
blah	83.87	cesium	76.54	😎	81.95	million	60.89	mnkdsecinvestigation	52.77	😎	51.40
energy	83.78	trading	75.83	😂	81.71	death	60.60	forex	51.43	😹	50.47
vs	83.77	pebbleproject	74.24	🙂	81.36	bubble	60.53	trading	51.13	🙂	49.33
sma	83.36	stock	73.85	😁	81.29	cents	60.39	elliottwave	46.82	🙂	49.33
billion	83.32	bullish	73.85	🤔	80.76	dip	60.33	scambags	46.19	🤔	48.90
candle	83.18	brachytherapy	72.85	💰	80.72	debt	60.33	earnings	46.19	😁	48.90
phase	82.99	tradesmart	72.43	💥	80.65	weekly	60.33	pennymikey	45.50	😝	47.40

Table 3 : Top 10 tokens ranking with CFIDF for sentiments *Bullish* and *Bearish*

similarity of each token with the "bullish" and "bearish", and subtact the cosine similarity with "bearish" from the cosine similarity with "bullish" to measure the tendency of each token. The token with positive score is considered to be the token with bullish tendency, and the token with negative score is considered to be the token with bearish tendency.

4. Analysis of Dataset

4.1 Analysis in Single Sentiment

The top 10 highly frequent tokens of both sentiments are shown in Table 2, respectively. First, we can find that "buy" is the most frequent word used in bullish posts. It is reasonable because the bullish label means a writer expects the price of the metioned instrument will rise, and may write down her/his action or ask the others to buy the mentioned instrument. "buy" is also in the top-10 tokens in bearlish posts. The reason is that it is unlimited for the rising price, but the minimun of the falling price is 0. In some bearish posts like (P1) "buy" is used to mention the buy back price of the instrument. Compared to bullish posts, the writers of bearish posts tend to use "short" but not "sell" to narrate the action they take.

(P1) *$SKX if you have profits, sell now and buy back below 30. You will thank me later*

Second, the uses of hashtags in financial social media can be formulated in the following four ways: (1) To tag the instruments they focus on (stock, bitcoin, forex), (2) To use an unique tag to store their posts (e.g., mnkdsecinvestigation), (3) To label their sentiments (e.g., bullish, sharkalerts, btfd), and (4) To tag the method they used (e.g., elliottwave). Compared to common social media, hashtags are not frequently used in financial social media. Only 1.37% of bullish posts and 1.75% of bearish posts contain at least one hashtag.

Third, emoji is more popular than hashtag in financial social media. Total 2.81% of bullish posts and 1.86% of bearish posts contain at least one emoji. In other domains, e.g., hotel review, the smile emoji stands for positive comment. In contrast, the Face With Tears of Joy (😂) emoji gets the first place in both bullish and bearish posts. It is a special phenomenon in sentiment analysis. Since the sentiment of an investor may depend on her/his return from trading, the investor who longs the intrument will feel happy if the price rise, but the investor who shorts the insturment will feel sad in this situation. Therefore, the market sentiment of investors should be discriminated from the positive/negative sentiment of investors in financial social media data. We will show some evidence for this issue in Section 5. Some emojis may not imply sentimentss, for example, Thinking Face (🤔) appears in both top-10 lists.

Furthermore, we use CFIDF to calculate the weights of tokens. The results are shown in Table 3. In bullish post, some positive words such as "bully" and "dilly", get the

first place, and some negative words like "junk", "death" and "dip" are shown in the top-10 list of bearish.

4.2 Analysis between Two Sentiments

To analyze the correlation of tokens and both sentiments, we adopt PMI to sort out the critical tokens. The top-10 results are shown in Table 4. In bullish posts, "dominant", "bully", "blast" and "undervalued" get the front place. If the instrument is described with "undervalued", it is similar to the writer expects that this instrument is bullish. The same case is also shown in bearish posts. The word "overvalued" gets the third place in the list. In bearish posts, "junk", "garbage", "trash" and "turd" appear in the top-10 list, and these words are considered as negative words, when they are used to described things. Furthermore, "puts", standing for one kind of options that buyers have the right to sell the underlying asset at certain price, is in the bearish list. The emoji results could also show some clues for the writer's sentiments. The Ox (🐂) and Airplane Departure (🛬) emojis get the front place in bullish posts, and the Down-Pointing Triangle (🔻), Thumbs Down Sign (👎), Pile of Poo (💩), and Bear Face (🐻) emojis are in the top-10 list of bearish posts.

Bullish						Bearish					
Word		Hashtag		Emoji		Word		Hashtag		Emoji	
dominant	1.22	buy	1.27		1.14	bagholders	3.46	mnkdsecinvestigation	3.53	🔻	4.42
bully	1.21	early	1.27		1.14	junk	3.45	pennymikey	3.53	👎	3.68
updates	1.21	gainers	1.27	🐂	1.14	overvalued	3.42	noshamenate	3.53	💩	3.53
runner	1.21	mattel	1.27		1.14	pumpers	3.37	sanofiwasright	3.53	🐻	2.59
binance	1.21	oprah	1.27	$	1.14	scam	3.34	scambags	3.53		2.01
blast	1.21	shortsqueeze	1.27		1.14	garbage	3.32	short	3.39		1.82
floater	1.21	analyis	1.27	🛬	1.14	pig	3.25	scam	3.29		1.82
undervalued	1.21	biotech	1.27		1.14	trash	3.25	forex	3.04		1.67
accumulating	1.20	blocks	1.27		1.14	turd	3.19	market	2.75		1.57
blackberry	1.20	boolish	1.27		1.14	puts	3.11	elliottwave	2.40		1.52

Table 4 : Top 10 tokens ranking with PMI for sentiments *Bullish* and *Bearish*

Bullish						Bearish					
Word		Hashtag		Emoji		Word		Hashtag		Emoji	
bully	1.22	sharkalerts	1.34		1.15	bagholders	3.42	short	3.01	👎	3.56
updates	1.22	club	1.34		1.15	junk	3.41	scam	2.91	💩	3.40
runner	1.22	bullboard	1.33		1.13	overvalued	3.38	forex	2.66	🐻	2.47
binance	1.22	dontbeasheep	1.33		1.13	pumpers	3.33	markets	2.37		1.89
blast	1.22	crypto	1.32		1.12	scam	3.30	elliottwave	2.02		1.70
floater	1.22	pennystocks	1.29		1.12	garbage	3.28	futures	1.48		1.69
undervalued	1.21	blockchain	1.28		1.12	pig	3.21	earnings	1.41		1.55
accumulating	1.21	moviepass	1.28		1.12	trash	3.21	stocks	1.16		1.45
blackberry	1.21	stockmarket	1.28		1.12	turd	3.15	trading	0.97		1.40
partnerships	1.21	timestamp	1.27		1.11	puts	3.07	study	0.86		1.39

Table 5 : Top 10 tokens ranking with PMI appearing in both sentiments *Bullish* and *Bearish*

Bullish						Bearish					
Word		Hashtag		Emoji		Word		Hashtag		Emoji	
buy	0.01	stocks	0.04		0.10	short	0.01	stocks	0.04		0.20
today	0.01	sharkalerts	0.02		0.06	stock	0.01	trading	0.02	💩	0.04
like	0.01	bitcoin	0.02		0.06	going	0.01	forex	0.01	🐻	0.03
go	0.01	trading	0.02		0.04	like	0.01	elliottwave	0.01		0.03
get	0.01	stock	0.01		0.03	sell	0.01	earnings	0.01		0.02
going	0.01	btfd	0.01		0.03	lol	0.01	markets	0.01		0.02
back	0.01	stockmarket	0.01		0.03	get	0.01	futures	0.01		0.02
day	0.01	club	0.01		0.03	back	0.01	study	0.01		0.02
shares	0.01	study	0.01		0.03	money	0.01	scam	0.01		0.02
stock	0.01	optionpros	0.01		0.02	today	0.01	short	0.01		0.02

Table 6 : Top 10 tokens ranking with WPMI appearing in both sentiments *Bullish* and *Bearish*

Bullish						Bearish					
Word		Hashtag		Emoji		Word		Hashtag		Emoji	
streamline	1.55	brent	1.25		1.16	fuh	-1.55	getyourshinebox	-1.26		-1.10
dinghy	1.55	crossover	1.23		1.11	pumptards	-1.35	timberrrrr	-1.23		-1.07
bitc	1.27	qnx	1.18		1.11	foolishness	-1.28	overpriced	-1.22		-0.99
awakes	1.26	buffett	1.17		1.11	bleeds	-1.28	pennymikey	-1.22		-0.94
rap	1.25	overwatch	1.15		1.11	grasshoppa	-1.26	scam	-1.18		-0.93
brent	1.25	powerhour	1.15		1.11	leeches	-1.24	overbought	-1.18		-0.92
crossover	1.23	paytheask	1.14		1.11	downgraded	-1.24	overvalued	-1.16		-0.89
attend	1.23	letsgo	1.14		1.11	timber	-1.24	bankruptcy	-1.16		-0.87
shaken	1.23	epyc	1.13		1.09	barev	-1.24	bitcoinfork	-1.15		-0.86
varta	1.22	wallstreetgames	1.13		1.09	myant	-1.23	forex	-1.15		-0.78

Table 7 : Top 10 tokens ranking with word embedding and cosine similarity for sentiments *Bullish* and *Bearish*

	NTUSD-Fin						SentiWordNet	
Word	Market sentiment	Chi squared	Bull freq.	Bull cfidf	Bear freq.	Bear cfidf	Sentiment	Word ID
buy	0.59	14711.71	14489	61.54	1592	52.32	0.00	buy#1
sell	-0.98	3581.53	5800	71.02	1897	51.60	0.00	sell#4
call	0.44	2211.63	2259	78.16	277	59.76	0.00	call#13
put	-0.49	973.82	1326	80.52	310	59.59	0.00	put#1
overvalued	-3.42	1625.99	54	71.49	172	59.75	0.25	overvalue#1
undervalued	1.21	1095.06	844	81.71	9	40.81	-0.38	undervalue#1

Table 8 : Comparision of NTUSD-Fin with SentiWordNet

To do the in-depth analysis, the tokens that appear in the posts of both sentiments are shown in Table 5. The results of hashtag and emoji of bullish posts are different from the results in Table 4. The hashtag "sharkalerts" gets the first place in the hashtag result. Moreover, none of emoji is the same as the emojis in Table 4. Rocket (🚀), Steam Locomotive (🚂) and Flexed Biceps (💪) are highly related to bullish. Moreover, we add the frequency information with WPMI, and the results are shown in Table 6. With frequency information, WPMI tends to pick out the general tokens. Most of tokens in Table 6 are the same as those scored with frequency in Table 2. Therefore, comparing with frequency and WPMI, the results scored by PMI contain most of specific tokens for market sentiment.

The top 10 tokens ranking by cosine similarity based on on word embedding for both "bullish" and "bearish" are listed in Table 7. Some tokens are different from those proposed by the other methods. The word, "downgraded", is picked out by this scoring method in bearish words, and the "overpriced", "overbought" and "overvalued" hashtags show high tendency toward bearish.

4.3 Dictionary Format

Because different information may have dissimilar usage, our dictionary provides various scoring methods including frequency, CFIDF, chi-squared value, market sentiment score and word vector for the tokens. Only the tokens appeared at least ten times and shown significantly difference between expected and observed frequency with chi-squared test are remained in our dictionary. The predetermined significance level is 0.05. The market sentiment score is calculated by substracting the bearish PMI from the bullish PMI. There are 8,331 words, 112 hashtags and 115 emojis in the constructed dictionary, NTUSD-Fin. Some examples are shown in Table 8. The distribution of these elements for both sentiments is shown in Table 9.

	Bullish	Bearish
Word	6,670	1,661
Hashtag	97	15
Emoji	103	12

Table 9 : Distribution of NTUSD-Fin.

5. Discussion

First, we discuss the gengeral sentiment and the market sentiment in financial social media. As we mentioned in Section 4, for the sentiment analysis in hotel reviews, the positive/negative sentiment of a writer is associated with a good/bad opinion of the hotel. This case is the same in movie reviews and product reviews. However, the sentiment of the investors may depend on the positions they hold. Therefore, the positive sentiment of the investor does not imply the bullish market sentiment for the mentioned target of this investor. It is worth distinguishing the market sentiments of the investors from the sentiments of the investors. To in-depth analysis, we compare the market sentiment scores with the sentiment scores in SentiWordNet 3.0 (Baccianella et al., 2010), a dictionary annotated with sentiments of all synsets in WordNet (Fellbaum, 2005) in Table 8. The results show the

differences between general sentiments and market sentiments. For instance, "buy" and "sell" are neutral words in SentiWordNet, but they get the positive and the negative market s scores in our dictionary, respectively. (P2) shows a bullish instance containing "buy", and (P3) shows a bearish instance containing "sell". With dictionary-based approach with SentiWordNet, the sentiment scores for both posts are zero. In contrast, the market sentiments provided by our dictionary are 0.64 and -0.93, suggesting correct information for market sentiments of investors. The words "call" and "put" are examples to show a similar phenomenon.

(P2) *$CHGG buy while you can...*
(P3) *$CADC sell*

In addition, the scores of "overvalued" and "undervalued" in our dictionary are opposite to the scores in SentiWordNet. This result shows the different between market sentiment and common sentiment. Because the price of overvalued instruments is expected to fall down, and the price of the undervalued instrument is expected to rise up, it is reasonable to be considered as bearish word and bullish word in financial data. However, "overvalue" and "undervalue" are the same as overestimate and underestimate in SentiWordNet, and get the positive and negative sentiment scores. The evidence illustrates the difference between general sentiment and market sentiment. (P4) and (P5) are the posts that will be misled by using SentiWordNet.

(P4) *$WTW very overvalued like $HLF*
(P5) *$MTBC so undervalued at these prices*

Second, we compare the words in our dictionary with those in the dictionary for formal documents in finance (Loughran and McDonald, 2011). There are 354 positive words and 2,355 negative words in their dictionary, and only 152 positive words and 329 negative words appear in our dictionary. This circumstance shows that the words used by investors in social media are different from those used in formal documents such as 10-K annual reports. Besides, "easy" is a positive word in the their dictionary, but gets a negative market sentiment score in our dictionary. It implies that not only the words are different between social media data and annual reports, but also the tendency of market sentiment of the same words may also different.

	Micro	Macro
Loughran and McDonald	21.67	23.02
NTUSD-Fin	61.23	40.22

Table 10 : Dictionary-based experimental results. (%).

Furthermore, we use dictionary-based model to test the performance of the dictionary of Loughran and McDonald (2011) and our dictionary. We use the number of positive words minus the number of negative words in the dictionary to score the sentiment of each tweet. The tweets get positive (negative) score will be considered as bullish (bearish) instances, and the tweets with zero score will be considered as neutral instances. SemEval-2017 Task 5 dataset (Cortis et al., 2017), which was collected from Twitter and StockTwits, is adopted for this experiment. In order to confirm all instances are discussing financial instruments, each instance contains at least one cashtag. There are total 2,030 instances in this dataset, including 1,318 bullish instances, 676 negative instances, and 36 neutral instances. Table 10 shows the micro- and macro-averaged F-score of both dictionaries. Our dictionary outperforms Loughran and McDonald's dictionary, which is constructed for formal documents. The experimental results shows the usefulness of the media-oriented dictionary.

In summary, the usage of the NTUSD-Fin dictionary is different from that of general sentiment dictionaries from several aspects. The purpose of this dictionary is to capture the market sentiment of the investors for the mentioned instruments in the social media platform, but not to predict the sentiments of the investors.

6. Conclusion and Future Work

In this paper, we distinguish the market sentiment of investors from the sentiments of investors. Not only the emoji shows the evidence of this phenomenon, the comparison with general sentiment dictionary shows too. The constructed market sentiment dictionary is based on a large-scale labeled data from financial social media. Words, hashtags and emojis are included in the dictionary. The POS tagging and the meaning of the words will be added in the future version. Our dictionary[2] is publicly available for research purpose.

7. Acknowledgements

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- and MOST 105-2221-E-002-154-MY3.

8. References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 2200–2204.

Bollen, J., Mao, H., & Zeng, X.J. (2011). Twitter Mood Predicts the Stock Market, *Journal of Computational Science*, 2(1): 1–8.

Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2014) Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *Review of Financial Studies* 27: 1367–1403.

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 (SemEval-2017)*, pages 519-535.

El-Haj, M., Rayson, P. E., Young, S. E., Walker, M., Moore, A., Athanasakou, V., & Schleicher, T. (2016). Learning tone and attribution for financial text mining. In *Proceedings of the tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1820-1825.

Fellbaum. C. (2005) Wordnet and wordnets. In K. Brown, editor, *Encyclopedia of Language and Linguistics*, pages

[2] http://nlg.csie.ntu.edu.tw/nlpresource/NTUSD-Fin/

665–670, Oxford, Elsevier.

Huang, A., Zang, A., & Zheng, R. (2013). Large sample evidence on the informativeness of text in analyst reports. *The Accounting Review*, 89(6): 2151-2180.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv, arXiv:1412.6980.

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 (CoNLL 2017)*, pages 301–310.

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.

Sul, H. K., Dennis, A. R., & Yuan, L. I. (2014). Trading on Twitter: The Financial Information Content of Emotion in Social Media. In *Proceedings of 47th Hawaii International Conference on System Sciences (HICSS)*, pages 806–815.