*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6106–6110
Marseille, 11–16 May 2020


# Issues and Perspectives from 10,000 Annotated Financial Social Media Data

**Chung-Chi Chen**[⋆] **Hen-Hsen Huang**[†‡] **Hsin-Hsi Chen**[⋆‡]
[⋆] Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
[†]Department of Computer Science, National Chengchi University, Taiwan
[‡]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

**Abstract**

In this paper, we investigate the annotation of financial social media data from several angles. We present Fin-SoMe, a dataset with 10,000 labeled financial tweets annotated by experts from both the front desk and the middle desk in a bank's treasury. These annotated results reveal that (1) writer-labeled market sentiment may be a misleading label; (2) writer's sentiment and market sentiment of an investor may be different; (3) most financial tweets provide unfounded analysis results; and (4) almost no investors write down the gain/loss results for their positions, which would otherwise greatly facilitate detailed evaluation of their performance. Based on these results, we address various open problems and suggest possible directions for future work on financial social media data. We also provide an experiment on the key snippet extraction task to compare the performance of using a general sentiment dictionary and using the domain-specific dictionary. The results echo our findings from the experts' annotations.

**Keywords:** financial social media, sentiment analysis, market sentiment

## 1. Introduction

Much effort is being applied to natural language processing (NLP) technologies in various domains. In the NLP community, many researchers have begun to use machine learning on financial and economic data. Recently, workshops and shared tasks such as ECONLP'19, FinNLP'19, FNP'19, DSMM'18, FiQA'18, and SemEval'17 Task 5 have been organized. This reflects the increasing interest of NLP researchers in financial and economic domains. In this paper, we focus on financial social media data and offer an in-depth analysis of the data that are useful for future work.

Mining crowd opinions has been a crucial challenge given the prevalence of smartphones and social media. Indeed, decision-makers have begun to consult crowd opinions in their deliberations. Many topic-specific social media platforms have been developed. The social trading platform is a typical example, which provides a forum for investors to discuss their ideas, trading strategies, and analyses of financial instruments. On StockTwits, a leading social trading platform, investors use cashtags to indicate financial instruments. For example, Apple Inc. and Microsoft Corporation are tagged as $AAPL and $MSFT, respectively. Furthermore, investors can also annotate their tweets with a bullish or bearish label. Financial tweets with the writer-labeled annotations have been used as training and test data (Li and Shah, 2017). However, the reliability of writer-labeled data is sometimes poor (Huang et al., 2018). For instance, tweet (T1) is labeled as bullish but lacks enough content. (T2), assigned a bullish label, is a description of the trading volume of $DPW. Note that both soaring and collapsing stocks are characterized by huge trading volumes. When the content of such tweets is associated with bullish or bearish labels and is used as training data, the performance of the resulting model may be greatly diminished.

**(T1)** $AAPL today...

**(T2)** $DPW 50 MILLION VOLUME!!!

Specifically, although an investor's label provides information about the mentioned cashtag, it does not always represent the content of a tweet. Therefore, when analyzing financial social media data, one important pre-process is to prune those tweets that lack the investor's opinions from the training data . However, no prior work addresses this issue. In this work, we provide a large annotated dataset for this purpose.

Statements with founded predictions are more convincing than empty, unfounded forecasts. How to identify which tweets are reasonable and analyze these tweets in-depth is an important topic of the research on financial social media data. In addition, given the annotated results, we find that few investors discuss their gains and losses on social trading platforms. This indicates that investor's performance cannot be evaluated simply by extracting information from a single tweet.

In this paper, we address critical problems when using financial social media data, and discuss directions for future research. The contributions of our work are three-fold as follows.

1. We conduct a careful study of financial social media data with an adequate annotated dataset.
2. We publish Fin-SoMe, a dataset for academic use with 10,000 tweets labeled from four aspects.
3. We present several perspectives toward research on financial social media data.

## 2. Related Work

Sentiment analysis in financial social media data has been a research focus in the recent decade. Several works (Bollen et al., 2011; Si et al., 2013; Sul et al., 2017; Vanstone et al., 2018) show that financial social media data provides trading clues. However, to the best of our knowledge, little work takes a close look at the contents of individual investors' tweets.

Maks and Vossen (2013) show that evaluations may differ between the writer and readers of a given product review;

they recommend to use product reviews based on readers' ratings rather than the writer's rating. However, most works on financial social media data use writer-labeled information to construct lexicons (Li and Shah, 2017) or train sentiment classifiers (Deng et al., 2018). Evidence presented in Section 4.1 will show that the direct use of these writer-labeled financial tweets yields poor results because only 30% of labeled tweets have content related to the market sentiment label, 15% provide no related content, and 55% of labels are ambiguous. Our statistical results invalidate the assumption of the previous work that holds that writer labels are reliable. Future work should focus on how to preprocess the data collection in a better way, for instance, identifying whether a tweet provides a description of market sentiments. The Fin-SoMe dataset presented in this paper will be useful for such a purpose.

Xu and Cohen (2018), who provide StockNet, a price prediction dataset with tweets and historical prices. A total of 88 stocks are selected with a price period from 01/01/2014 to 01/01/2016. However, the price provided for $GMRE starts from 06/30/2016, indicating that the future price data is being used to select former targets, that is inadequate for backtesting. Furthermore, it seems strange that the average number of tweets for $AAPL would average 29.80 per day in the StockNet training set. Indeed, when using Tweepy to crawl Twitter tweets for 143 days from 31/05/2018, we find the average number of tweets that mention $AAPL is 8,411.88. Previous cases show that in-depth understanding of a specific domain and the data used is important and necessary before conducting experiments using the data.

To the best of our knowledge, no work has provided an expert-annotated dataset as large as that presented here. Most existing datasets use financial tweets directly. Fin-SoMe is the first sizable expert-annotated dataset, and can be used freely by the academic community. Our statistical results and the problems we identify provide insight into the construction of datasets with financial social media data and the future researches on financial social media.

# 3. Dataset

In this section, we describe the collected financial tweets and the annotation process in detail. The annotated results have been released under the CC BY-NC-SA 4.0 license with the related tweets' IDs.

## 3.1. Statistics of the Collected Dataset

From StockTwits, we collected 10,000 writer-labeled tweets that were labeled by the writers as either bullish or bearish. Each financial tweet contains at least one cashtag, as in (T1) and (T2). Table 1 shows basic dataset statistics. As with Li and Shah (2017), bullish tweets are more prevalent than those with a bearish label. Only 18% of users publish bearish statements. On average, there are 2.19 cashtags in a tweet.

## 3.2. Annotation

We hired both front-desk and middle-desk experts from a bank treasury. The front-desk expert (RN) is working in the treasury marketing unit, and the middle-desk expert (RA) is working in the risk management department. Annotators spent five months completing the labeling of 10,000 tweets. We investigate a tweet from four angles:

- The market sentiment (bullish/bearish) of the tweet;
- The presence or absence of the reason in the tweet supporting the investor's analysis;
- The writer's sentiment (positive/negative);
- The gain/loss of the writer's trade.

Because writers do not always show the market sentiment, the writer's sentiment, or gain/loss in their tweets, annotators assigned a "None" label to those tweets without the related description.

| Number of | Bullish | Bearish |
|---|---|---|
| Tweets | 8,567 | 1,433 |
| Unique users | 3,731 | 822 |
| Cashtags | 19,133 | 2,726 |

Table 1: Fin-SoMe statistics.

| RA \ RN | Bullish | Bearish | None | Sum |
|---|---|---|---|---|
| Bullish | 2,785 | 105 | 205 | 3,095 |
| Bearish | 183 | 258 | 74 | 515 |
| None | 4,360 | 563 | 1,467 | 6,390 |
| Sum | 7,328 | 926 | 1,746 | 10,000 |

Table 2: Annotation results for market sentiment.

# 4. Annotation Results and Findings

## 4.1. Market Sentiment and Writer Sentiment

Table 2 shows the annotation results of market sentiment. Only 2,785 and 258 tweets have a consistent agreement on bullish and bearish statements, respectively. This suggests that only 30% of financial tweets are written with explicit content providing clues for market sentiment. In most cases, the sentiment is ambiguous, i.e., the interpretation depends on the reader's risk attitude. In example (T3), RN labels it as bullish and RA as none. We have the following two observations from this example. (1) Whereas RN annotates tweets with implicit descriptions of market sentiment, RA assigns market sentiment labels only to tweets with explicit descriptions. (2) If the writer's own labeled market sentiment toward $MARA is bullish, why would he/she close the $MARA position?

> **(T3)** $MARA Sold for $10,656 profit. Wowza, what a day. I'll be back lol

| Writer \ Reader | Bullish | Bearish | Sum |
|---|---|---|---|
| Bullish | 88.83% | 1.54% | 90.37% |
| Bearish | 2.69% | 6.93% | 9.63% |
| Sum | 91.52% | 8.48% | 100.00% |

Table 3: Comparison of writer and reader labels.

| | Positive | Negative |
|---|---|---|
| Bullish | 89.66%/ 89.71% | 0.18%/ 0.57% |
| Bearish | 1.96%/ 3.26% | 8.20%/ 6.46% |

Table 4: RN/RA - Annotation results for market sentiment and writer sentiment.

| RA \ RN | Founded | Unfounded | Sum |
|---|---|---|---|
| Founded | 541 | 5,712 | 6,253 |
| Unfounded | 152 | 3,595 | 3,747 |
| Sum | 693 | 9,307 | 10,000 |

Table 5: Annotation results for founded/unfounded.

This may suggest that the writer has a neutral, even bearish, market sentiment toward $MARA's short-term price movement, and may have a bullish market sentiment toward long-term price movement.

In Table 3, we evaluate the consistency between the writers' labels and those fully agreements by annotators. A total of 97.06% $(88.83/(88.83+2.69))$ of the tweets that are considered bullish by the annotators are assigned with the same label by the writers. For bearish tweets, only 81.81% $(6.93/(1.54+6.93))$ are labeled bearish by both annotators and writers. Writer-labeled data may not only be ambiguous, but also misleading to investors. In (T4), the writer's label is bearish, but the annotators label it as bullish. It shows that we cannot use the data collected from social trading platforms directly. We also underscore the need to clean up the data collected.

> **(T4)** $CEI Only down 7% if this is not a Bullish sign i do not know what it.

Table 4 further shows that the market sentiment and the writer sentiment may be different. It indicates that disambiguating between the market sentiment and the general sentiment is needed when analyzing the financial social media data. In Section 5, we will show that the market sentiment and the general sentiment can result in very different performances in the experiments related to the financial social media data.

### 4.2. Founded/Unfounded and Gain/Loss

Table 5 shows the annotation results reflecting whether writers offer reasons for their analyses. Only 5% of tweets are fully agreed-upon by annotators. Since most writers do not provide evidence supporting their predictions, evaluating the reliability of a writer's prediction is difficult. One possible evaluation approach is to consider their historical performance. Therefore, we annotate the gain/loss information in financial tweets. (T5) provides an example with gain/loss information, i.e., this writer earned 18% on a $DG position.

> **(T5)** $DG In at 72 out at 85. Nice profit.

According to Table 6, it is rare that we are able to extract such a fine-grained performance evaluation in (T5). Up

| RA \ RN | Gain | Loss | None |
|---|---|---|---|
| Gain | 60 | 0 | 128 |
| Loss | 0 | 12 | 25 |
| None | 207 | 30 | 9,538 |

Table 6: Annotation results for writer gain/loss.

to 95% of financial tweets do not contain gain/loss information. It indicates that we cannot evaluate investor performance directly by extracting gain/loss information from the content, and must therefore infer performance from past tweets. For example, capturing the buying price and selling price of the same writer (Chen et al., 2019) can be used to evaluate the performance of this writer.

## 5. Comparison of Market Sentiment and General Sentiment

In order to show the difference between market sentiment and general sentiment, we experiment on the expert-annotated dataset (Cortis et al., 2017). We compare the performance of using general dictionary, SentiWordNet (Baccianella et al., 2010) with that of using market sentiment dictionary, NTUSD-Fin (Chen et al., 2018a). The experimental results echo the finding of our annotation results of writers' sentiment and the market sentiment.

### 5.1. Dataset

Semeval-2017 task 5 dataset (Cortis et al., 2017) is adopted in this paper. There are 2,030 tweets collected from Twitter and Stocktwits. This dataset was annotated by three independent experts. The details of the toolkit for annotation and the annotating process are described in Daudert et al. (2019). They selected piece(s) of tweets containing opinions for a certain cashtag as the key snippet. In our experiment, we test the performance of the models in extracting the key snippet as the experts with different sentiment dictionaries. (T6) is an example of the key snippets of the target cashtag, $VXX (iPath S&P 500 VIX Short-Term Futures ETN) and $SPX (S&P 500 Index).

> **(T6)** $VXX on the move up again should bring $SPX down into the close. we see.

**Annotated result of (T6)**:

- $SPX: should bring $SPX down into the close.
- $VXX: on the move up again

### 5.2. Approaches to Key Snippet Extraction

#### 5.2.1. Position-based Approach (PB)

Position-based approach is a strong baseline for this task. A tweet is separated into several segments based on punctuation marks, and the segment containing a target cashtag or a target company name will be regarded as the key snippet of the given target.

#### 5.2.2. Dependency-based Approach (DB)

The stanford parser (De Marneffe et al., 2006) is adopted for parsing the tweets. The dependency parse result for an

| | | P | R | F1 |
|---|---|---|---|---|
| PB | | 39.51 | 59.89 | 44.79 |
| DB | | 56.58 | 88.72 | 64.87 |
| GRU | | 40.06 | 80.66 | 50.00 |
| | GS | 68.91 | 68.91 | 65.78 |
| | MS | 76.82 | 76.36 | 73.32 |
| BiGRU | | 39.33 | 82.13 | 49.68 |
| | GS | 66.37 | 71.90 | 66.02 |
| | MS | 74.94 | 74.75 | 71.62 |

Table 7: Experimental results. (%)

| | Gain | Loss |
|---|---|---|
| Bullish | 75.00% | 6.94% |
| Bearish | 8.33% | 9.72% |

Table 8: Annotated gain/loss and writer-labeled market sentiment.

| | Gain | Loss |
|---|---|---|
| Positive | 83.08%/ 90.91% | 0.00%/ 0.00% |
| Negative | 1.54%/ 0.00% | 15.38%/ 9.09% |

Table 9: RN/RA-annotated gain/loss and writer sentiment.

n-word tweet is n triples in the form of $dep(word_i, word_j)$, where $word_i$ and $word_j$ has a dependency $dep$, $word_i$ is a parent of $word_j$, and $word_j$ is a child of $word_i$. The ancestors and the decedents of a target cashtag form the key snippet of the given target.

#### 5.2.3. Learning-based Approach

We adopt several neural network (NN) models to test the capability of the NN models for the proposed task. We use neural network models to determine which words in a tweet should be kept in the key snippet. Gated recurrent neural network (GRU) (Cho et al., 2014) and bidirectional GRU (Bi-GRU) are adopted in the experiments. The first layer of the GRU model and Bi-GRU model are gated recurrent unit layer and bi-directional gated recurrent unit layer, respectively. Both model follows by one densely-connected neural network layer, one dropout layer, one ReLU layer, and a sigmoid output layer. The output of this model determines whether each token in the tweet should remain.

The preprocessing procedure for the learning-based models is shown as follows. A financial tweet consists of words, cashtags, user id, numbers, URL, hashtags and emojis. Both hashtags and emojis are regarded as words in this paper. First, we remove user ids, URL and punctuation marks. Second, we replace numbers and cashtags by special symbols, "NUM" and "TICKER", respectively. In particular, the target cashtag and target company name are replaced by "TARGET". Finally, we transform the remaining tokens into lowercase.

### 5.3. Experimental Setup

We fold a 10-fold cross-validation for evaluating the learning-based approach. A total 10% of the instances in the training set is used as the validation set, and early stopping is triggered after five trial epochs. In this experiment, the cell size of GRU is 64, and the hidden dimension is 128, and the rate of dropout is 0.1. The input of the learning-based models includes (1) the word vector pretrained with the skip-gram scheme by the collected dataset with over 334K financial tweets, (2) word vector concatenating with the general sentiment score (GS) in the SentiWordNet (Baccianella et al., 2010), and (3) word vector concatenating with the market sentiment score (MS) in the NTUSD-Fin (Chen et al., 2018a).

### 5.4. Experimental Results

The results of different approaches are shown in Table 7. The F1-score of the baseline models PB and DB are 44.79% and 64.87%, respectively. We find that in the small dataset, adopting dictionary information is useful for the key snippet extraction task. Comparing the performances of using different dictionaries, both GRU and BiGRU perform better with the information of market sentiment dictionary than using general sentiment dictionary. It supports our findings from the annotations. The market sentiment and the general sentiment should be considered as different things when dealing with the financial textual data.

## 6. Discussion

Distinguishing writer sentiment and market sentiment is essential because they provide different clues. For example, we can obtain the forecast for certain financial instruments from investors via their market sentiments. However, writer sentiment does not provide this information. It can reflect investor's fear, which may be related to price volatility (Behrendt and Schmidt, 2018).

Considering writer sentiment and market sentiment with gain/loss information, we find clues that the gain/loss of investors is indeed related to writer sentiment but may be not directly related to their market sentiment. Table 8 shows that the market sentiment of investors may not be correlated to their gain/loss. However, as shown in Table 9, positive and negative writer sentiment are highly related to the gain/loss of the investors' position. This suggests features for future research that facilitates the evaluation of the performance of individual investors.

Detailed analysis of financial tweets has become a focus of researchers on financial social media data. Maia et al. (2018) provide a dataset for identifying the aspect of a financial tweet, and our previous work (Chen et al., 2018b) provide a taxonomy for the numerals in financial tweets, showing that numeral information is useful for price movement prediction. These studies afford a more thorough understanding of financial tweets. Beyond sentiments, more fine-grained comprehension toward context is imperative and promising.

## 7. Conclusion

Fin-SoMe provides evidence that the content of writer-labeled financial tweets may not include any cues for the writer's market sentiment. We present a detailed analysis of the textual information in financial tweets. Based on statistical results, we identify several challenges with financial social media data. We adopt the key snippet extraction task to show the difference between the market sentiment and

the general sentiment. Experimental results support our findings from annotated data—the market sentiment and the general sentiment should be distinguished when analyzing the financial textual data. We also suggest directions for future work. In order to broaden the usage of the annotations in Fin-SoMe, for the instances that did not get the consensus results from the annotations of RA and RN, the other expert selects the label from the annotations of RA and RN. We release the annotations in Fin-SoMe[1] under the CC BY-NC-SA 4.0 license for academic purposes.

## 8. Acknowledgement

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-106-2923-E-002-012-MY3, MOST 108-2218-E-009-051, MOST 108-2634-F-002-017, and MOST 109-2634-F-002-034, and by Academia Sinica, Taiwan, under grant AS-TP-107-M05.

## 9. References

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204.

Behrendt, S. and Schmidt, A. (2018). The twitter myth revisited: Intraday investor sentiment, twitter activity and individual-level stock return volatility. *Journal of Banking & Finance*, 96:355–367.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.

Chen, C.-C., Huang, H.-H., and Chen, H.-H. (2018a). Ntusd-fin: a market sentiment dictionary for financial social media data applications. In *Proceedings of the 1st Financial Narrative Processing Workshop (FNP 2018)*.

Chen, C.-C., Huang, H.-H., Shiue, Y.-T., and Chen, H.-H. (2018b). Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 136–143. IEEE.

Chen, C.-C., Huang, H.-H., and Chen, H.-H. (2019). Crowd view: Converting investorsâ opinions into indicators. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6500–6502. AAAI Press.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.

Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., and 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.

Daudert, T., Zarrouk, M., and Davis, B. (2019). CoSACT: A collaborative tool for fine-grained sentiment annotation and consolidation of text. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 34–39, Macao, China, 12 August.

De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Lrec*, volume 6, pages 449–454.

Deng, S., Huang, Z. J., Sinha, A. P., and Zhao, H. (2018). The interaction between microblog sentiment and stock return: An empirical examination. *MIS quarterly*, 42(3):895–918.

Huang, H.-H., Chen, C.-C., and Chen, H.-H. (2018). Disambiguating false-alarm hashtag usages in tweets for irony detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 771–777.

Li, Q. and 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.

Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., and Balahur, A. (2018). Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1941–1942. International World Wide Web Conferences Steering Committee.

Maks, I. and Vossen, P. (2013). Sentiment analysis of reviews: Should we analyze writer intentions or reader perceptions? In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 415–419.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., and Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–29.

Sul, H. K., Dennis, A. R., and Yuan, L. (2017). Trading on twitter: Using social media sentiment to predict stock returns. *Decision Sciences*, 48(3):454–488.

Vanstone, B. J., Gepp, A., and Harris, G. (2018). The effect of sentiment on stock price prediction. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 551–559. Springer.

Xu, Y. and Cohen, S. B. (2018). Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979.

[1] `http://nlg.csie.ntu.edu.tw/nlpresource/FinSoMe`