Mining User's Opinion Towards the Rising and Falling Trends of the Stock Market: A Hybrid Model

Haoda Qian[12] Liping Chen[1] Qiwen Zha[3,*]

[1]The State Key Laboratory of Management and Control for Complex Systems

Institute of Automation, Chinese Academy of Sciences Beijing 100190, China

[2]University of Chinese Academy of Sciences, Beijing, China

[3]China Academy of Industrial Internet

{qianhaoda2019, liping.chen}@ia.ac.cn, zhaqiwen@163.com

2021 IEEE International Conference on Intelligence and Security Informatics (ISI) | 978-1-6654-3838-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISI53945.2021.9624687

***Abstract*—Mining users' opinions towards the rising and falling trends of the stocks may help the management department estimate the risk and make timely decision. Existing methods ignore the effective fusion of domain information and pre-trained language models, hindering mining implicit semantic information. This paper proposes a hybrid method that adopts masked language modeling to obtain a domain-information-enhanced language model. Firstly, it generates an attention-mechanism-oriented masking based on words' importance, word-level polarity and terminology. Then, the masked words and their corresponding knowledge are predicted to acquire domain-aware language representation. Experimental results on two public financial sentiment analysis datasets show the efficacy of the proposed model.**

Keywords—stock sentiment analysis, pre-trained language models, domain information enhancement

I. INTRODUCTION

Mining users' opinions about the rising and falling trends of the stocks may facilitate management department understand market dynamics and estimate the risk [1]. Take a post "Not good for bear case for $XY" that expresses bullishness in the market as an example, it contains sentiment, terminology, and part-of-speech information. Among them, "good" and "bear" are sentiment words that indicate positive and negative polarity. "bear case" is a commonly used terminology in the market that expresses bearishness. The contained part-of-speech information lies in that "not" modifies "good ", "bear" modifies "case", and "not good" modifies "bear case". How to capture the above domain relevant information and unearth the latent semantic association among them is the key to identifying the market trends.

Recently, large-scale pre-trained language models such as BERT [2] have achieved good performance on a variety of NLP tasks. These models use LSTM or Transformer to acquire contextualized representation by performing pre-training tasks such as masked language modeling. FinBERT [3] first proposes a pre-trained language model for analyzing financial sentiment. It is trained on financial news corpus and employs layer-wise fine-tuning to prevent catastrophic forgetting. However, it neglects the integration of domain information that contains rich semantic information.

SentiLare [4] proposes a pre-training mechanism to fuse linguistic knowledge into pre-trained language models for predicting the overall rating of reviews. Inspired by the ideas that integrate linguistic features and pre-trained language models [4], we develop a hybrid method to mine the market trends by incorporating domain information with pre-trained language models. It performs domain information-aware masked language modeling, where significant tokens such as domain terminology words and sentiment words are assigned higher masking probability. Experimental results on two public financial sentiment analysis datasets show the efficacy of the proposed model.

II. PROPOSED MODEL

shows the framework of the proposed model. It contains three parts, e.g. Domain Information-Aware Masked Language Modeling, Domain Adaptation Pre-Training and User's Opinion Mining.

The model first obtains domain knowledge and exerts these information to calculate the word masking probability to identify important words, then performs domain adaptation pre-training to learn a domain-aware text representation, finally, the model is used to mine the users' opinion towards the stock market.

A. Domain Information-Aware Masked Language Modeling (DIA-MLM)

At present, most of the pre-trained models use random word masking strategy that treats all words equally. In order to incorporate the domain prior information into the mask task, it is necessary to give each word a domain-oriented mask probability. This paper proposes "domain information-aware masked language modeling" (DIA-MLM). To identify the correlation between word and domain sentiment information, each word is associated with a polarity label based on the financial sentiment lexicon. A domain terminology lexicon is constructed to identify financial terms in the text, which promotes language understanding in the financial domain. In addition, we add tf-idf to measure the importance of words. Finally, the weight score s_{x_i} for word x_i is determined by comprehensive information from these aspects, which is the product of these three types of weight. For polarity weight, a higher weight is given to words with positive / negative polarity since they are more likely to reflect the sentiment of the text. We also assign a higher weight for terms. The word weight of a sentence is normalized to produce a masking probability vector:

$$\alpha_i = s_{x_i|X} / \sum_{1}^{n} s_{x_j|X} \tag{1}$$

Intuitively, some words such as sentiment and transition words are more important than others for a specific task.

* Corresponding author

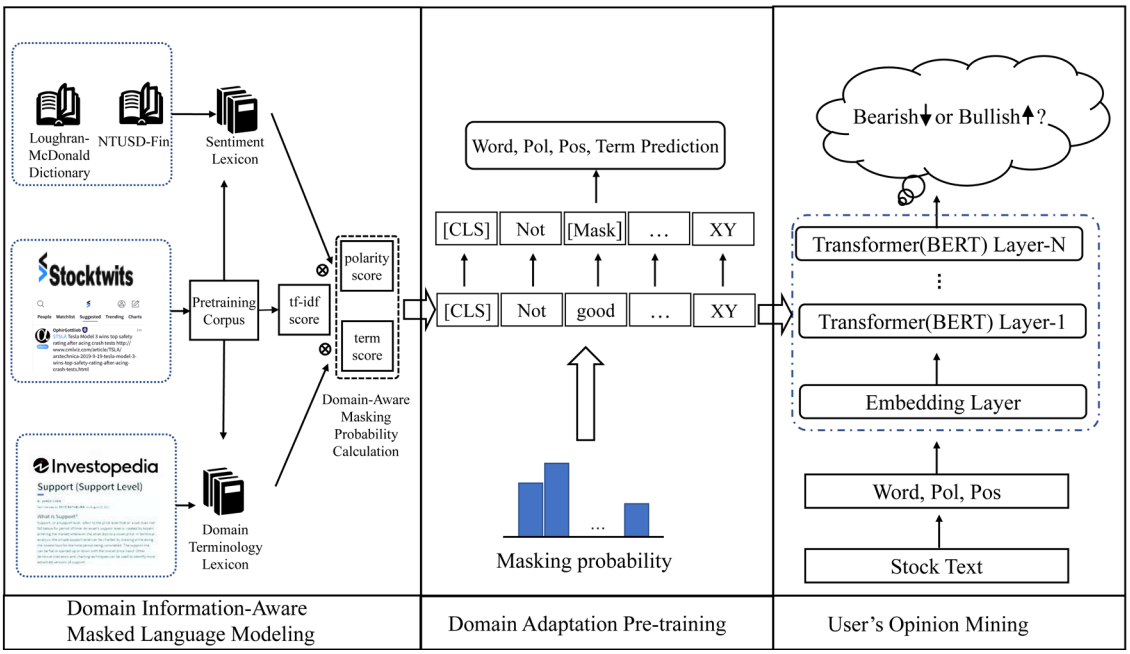

Figure 1 Overview of the proposed model

Domain information-aware masking method assigns different masking probabilities to words by selectively masking the important tokens for financial sentiment analysis. This mechanism can also relax the potential overfitting that learns general domain information and guides the model to learn more domain words.

B. Domain Adaptation Pre-Training (DAPT)

This pre-training step first utilizes DIA-MLM to mask words, then learns domain-aware and sentiment-aware language patterns through contextual language model by performing prediction for masked positions. In this phase, we assign polarity label and part-of-speech tag to each word. The input is the domain information-enhanced text sequence with some masked tokens $\hat{X}=\{x_t, pos_t, pol_t\}_1^n$, where x_i represents word i, pos_t and pol_t means pos tag and word-level polarity, respectively. The input embeddings of $\hat{X}$ contains the embedding used in RoBERTa along with the part-of-speech (POS) embedding and the polarity (POL) embedding. We obtain the representation vectors $\hat{H}$ with the input of $\hat{X}$:

$$\hat{H}=\left(\hat{h}_{cls},\hat{h}_1,\ldots,\hat{h}_n,\hat{h}_{sep}\right)=\text{Transformer}\left(\hat{X}\right) \tag{2}$$

where $\hat{h}_{cls}$ and $\hat{h}_{sep}$ are the hidden states of the special tokens [CLS] and [SEP] .The model predicts the word, pos tag, polarity label and whether the word is terminology at the masked positions. Thus, the overall training objective is minimized as follows:

$$\begin{aligned}\mathcal{L}_{DAPT}&=\sum_{t=1}^{n} m_t[\mathcal{L}_{MLM}(x_t)+\mathcal{L}_{MLM}(pos_t)+\\&\mathcal{L}_{MLM}(pol_t)+\mathcal{L}_{MLM}(term_t)]\\&=-\sum_{t=1}^{n} m_t[\log P(x_t\mid\hat{X})+\log P(pos_t\mid\hat{X})+\\&\log P(pol_t\mid\hat{X})+\log P(term_t\mid\hat{X})]\end{aligned} \tag{3}$$

where m_t equals to 1 if x_t is masked. The model predicts $P(x_t\mid\hat{X})$, $P\left(pos_t\mid\hat{X}\right)$, $P\left(pol_t\mid\hat{X}\right)$ and $P(term_t\mid\hat{X})$ based on the hidden states h_t .

This pre-training task exerts word information and several kinds of prior knowledge synergistically. The multi-layer transformers are used to generate knowledge-enhanced contextualized language representation.

III. EXPERIMENTS

A. Dataset

For pre-training, we construct a stock forum corpus from Stocktwits crawled from 2021/01/01 to 2021/05/01. The pre-training corpus contains 609K posts and 16M words.

To show the efficacy of the proposed model, we conduct experiments on Financial PhraseBank [5] and Fin-SoMe [6]. Financial PhraseBank is a news-based dataset for financial sentiment classification. The dataset contains 4,840 sentences selected from financial news and the sentiment label is either positive, neutral or negative. Fin-SoMe is a social-media-based market sentiment analysis dataset. It consists of 10,000 financial posts from stock forum annotated with labels including bullish, bearish and unsure.

B. Compared Models and Experimental Settings

We adopt BERT-based models (BERT [2], FinBERT [3]) and RoBERTa-based models (RoBERTa [7], BERTweet [8]) for comparison.

For financial sentiment lexicons, we use Loughran-McDonald Sentiment Lexicon [9] for news corpus and NTUSD-Fin [10] for social media corpus. The polarity words and term words are assigned with a higher masking probability. We consider seven pos tags including verb(v), adjective(a), adverb(r), noun(n), symbol(s), cardinal number(c) and others(o). The max text length for pre-training and fine-tuning are set to 128 and 64 respectively. The learning rate for pre-training and fine-tuning is set to 2e-5. The model is trained 1 epoch in the pre-training phrase and 5 epochs in the fine-tuning phrase. We employ the weight of RoBERTa for initialization. During the experiment, we stochastically select 80% of the data for training and the remaining 20% for testing and repeat the experiments for 5 times. Both the financial sentiment analysis datasets suffer

from class imbalance so *macro-f1* and *accuracy* are used as evaluation metrics.

C. Results and Analysis

We report the metric in TABLE 1. FinBERT achieves 85.8 and 57.16 in F1 on Financial PhraseBank and Fin-SoMe. FinBERT achieves lower F1 than BERT on Fin-SoMe, which indicates slightly different linguistic features between news and social media texts. Since RoBERTa is trained on larger corpus and uses robust optimization algorithm, it outperforms BERT-based models on both datasets, compared with FinBERT, it improves F1 from 85.8 to 86.8, 57.16 to 60.98, respectively. By training the model on general social media corpus, BERTweet performs better than RoBERTa on Fin-SoMe, with F1 of 85.57 and 61.25 on the two datasets, indicating that pre-training on domain corpus helps enhance downstream task performance. Different from the above methods, DIA-LM not only considers sentiment polarity and pos tag, but also integrates domain knowledge like financial sentiment lexicon and domain terminology, it further improves the performance of market sentiment analysis, with F1 of 87.64 and 61.72. The result demonstrates the advantages of the fusion of domain knowledge and pre-trained language model.

TABLE 1 Performance of compared methods

	Dataset	Financial PhraseBank		Fin-SoMe	
	Method	F1	Acc	F1	Acc
Baseline	Bert-based				
	BERT	85.2	86.3	59	75.47
	FinBERT	85.8	86.8	57.16	75.63
	Roberta-based				
	RoBERTa	86.8	87.58	60.98	75.77
	BERTweet	85.57	87.02	61.25	75.92
Ours	DIA-LM	87.64	88.16	61.72	76.5

D. Case Study

Figure 2 presents an application scenario of the proposed method. Take the stock $XY for an example, the mined polarity and terminology provide clues information for understanding why the stock trend is regarded as bearish or bullish. By making good use of the financial domain knowledge, the polarity words "overvalued", "risk", "broken" and "lower", the financial terminology like "capital structure" and "bear case" are identified. In detail, "capital structure" means the particular combination of debt and equity of a company, and “bear case” indicates a pessimistic outlook for the stock market. The domain information related to financial market fluctuations is leveraged in the pre-training and fine-tuning steps, the stock trends implicit in the user's opinion can thus be mined. The results reflect users’ confidence in the stock market, and the regulatory authorities may track market changes promptly to make effective decisions.

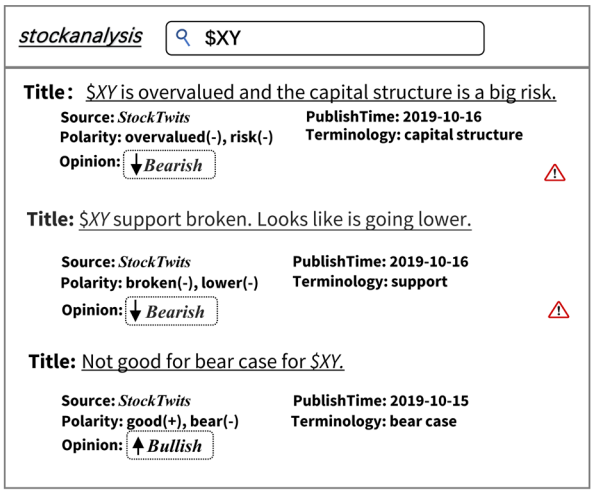

Figure 2 Case study on Fin-SoMe

IV. CONCLUSION

This paper proposes a hybrid method for analyzing users’ opinions towards the market trends, which adopts domain information-aware masked language modeling to pay more attention to domain-specific and task-specific words. The model can fuse other information such as multi-channel market related information from news media and social media, the writing style of the post, etc. in the future.

V. REFERENCES

[1] T. Loughran and B. Mcdonald, "Textual Analysis in Accounting and Finance: A Survey," *Behavioral & Experimental Finance eJournal,* 2016.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *NAACL-HLT*, 2019.

[3] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," *ArXiv,* vol. abs/1908.10063, 2019.

[4] P. Ke, H. Ji, S. Liu, X. Zhu, and M. Huang, "SentiLARE: Sentiment-Aware Language Representation Learning with Linguistic Knowledge," in *EMNLP*, 2020.

[5] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," *Journal of the Association for Information Science and Technology,* vol. 65, 2014.

[6] C.-C. Chen, H.-H. Huang, and H.-H. Chen, "Issues and Perspectives from 10, 000 Annotated Financial Social Media Data," in *LREC*, 2020.

[7] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *ArXiv,* vol. abs/1907.11692, 2019.

[8] D. Q. Nguyen, T. Vu, and A. Nguyen, "BERTweet: A pre-trained language model for English Tweets," in *EMNLP*, 2020.

[9] T. Loughran and B. McDonald, "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *Journal of Finance,* vol. 66, pp. 35-65, 2011.

[10] C.-C. Chen, H.-H. Huang, and H.-H. Chen, "NTUSD-Fin: A Market Sentiment Dictionary for Financial Social Media Data Applications," in 1st *Financial Narrative Processing Workshop*, 2018.