HEC MONTRÉAL
École affiliée à l'Université de Montréal

Three Essays on Advanced Natural Language Processing for Finance

par
Pan Liu

Thèse présentée en vue de l'obtention du grade Ph. D. en administration
(option Sciences des données)

Octobre 2023

HEC MONTRÉAL

École affiliée à l'Université de Montréal

Cette thèse intitulée :

Three Essays on Advanced Natural Language Processing for Finance

Présentée par :

Pan Liu

a été évaluée par un jury composé des personnes suivantes :

David Ardia
HEC Montréal
Président rapporteur

Gilles Caporossi
HEC Montréal
Directeur de recherche

Denis Larocque
HEC Montréal
Membre du jury

Liang (Alan) Zhang
Florida International University
Examinateur externe

Geneviève Gauthier
HEC Montréal
Représentant(e) du (de la) directeur(trice) de HEC Montréal

Résumé

L'extraction systématique d'informations exploitables à partir de données textuelles non structurées a été un défi persistant dans le domaine de la finance. En réponse, cette thèse se concentre sur le développement et l'application de méthodes avancées de traitement automatique du langage naturel (TALN) pour relever ce défi. Notre recherche présente trois essais qui proposent différents modèles de TALN pour extraire divers types d'informations des textes financiers, et démontrent leurs applications dans différents marchés financiers.

Le premier essai utilise la TALN pour étudier les liens entre les médias sociaux et le marché de la crypto-monnaie. Nous créons un modèle d'analyse du sentiment orienté vers la finance basé sur le transformateur, la dernière architecture de modèle d'apprentissage en profondeur TALN. Le modèle est pré-entraîné sur une grande quantité de textes bruts d'actualités commerciales afin qu'il apprenne la connaissance du langage financier et son vocabulaire. Il est ensuite affiné sur un ensemble de données étiquetées pour prédire le sentiment financier, sur lequel il obtient de hautes performances. Nous appliquons le modèle aux messages sur les médias sociaux concernant le bitcoin provenant de Twitter, Reddit, et Stocktwits. Nous combinons les résultats du sentiment avec les données du marché du bitcoin pour examiner les connexions entre le sentiment, l'attention, le désaccord sur les médias sociaux, et les activités de trading et les rendements du bitcoin. Nous montrons que 1) un sentiment plus élevé sur les réseaux sociaux entraîne une augmentation des rendements et du volume des trading de bitcoins, 2) une attention plus élevée et un désaccord sur les médias sociaux augmentent la volatilité des prix du bitcoin, 3) un changement positif du sentiment entraîne une diminution de la volatilité, et 4) l'ampleur de l'impact des différents médias sociaux varie.

Le deuxième essai développe un modèle TALN qui approfondit la mesure du sentiment orienté vers la finance. Contrairement au modèle du premier essai, ce modèle mesure le sentiment ciblé sur des entités spécifiques dans le texte, plutôt que le sentiment global d'une phrase. Nous proposons d'abord une architecture de modèle basée sur des prompts qui atteint une performance de pointe sur plusieurs jeux de données de référence pour

l'analyse générale du sentiment ciblé. Par la suite, avec un jeu de données de sentiment financier ciblé sur médias sociaux que nous créons, nous affinons ce modèle afin qu'il puisse mesurer le sentiment financier ciblé avec une grande précision. Nous l'appliquons ensuite à 23 millions de publications sur les médias sociaux liées à la finance provenant de différentes plateformes pour mesurer le sentiment financier envers 24 actions mème (actions qui gagnent une attention frénétique de la part des investisseurs particuliers sur les médias sociaux qui s'accompagne souvent d'un mouvement de prix spectaculaire) et 30 constituants du Dow Jones. Nos résultats montrent que le sentiment mesuré par notre modèle est positivement corrélé avec le rendement des prix et négativement corrélé avec la volatilité des prix, et que cette corrélation est plus forte pour les actions mème que pour les actions du Dow Jones. Nous démontrons en outre que la mesure du sentiment de notre modèle surpasse économiquement d'autres mesures du sentiment financier représentatives existantes.

Le troisième essai introduit un modèle pour mesurer la subjectivité du texte financier, qui pourrait fournir des informations supplémentaires au-delà du sentiment. L'analyse de la subjectivité, qui implique la différenciation entre les passages subjectives et objectives, est un sujet important en TALN, mais reste sous-exploré dans le domaine de la finance. Pour combler cette lacune, nous formons d'abord des annotateurs ayant une solide formation en commerce à étiqueter les subjectivités d'un ensemble de textes financiers. Ensuite, nous concevons un modèle de transformateur basé sur des prompts qui intègre spécifiquement le vocabulaire financier et est pré-entraîné sur des tâches de sémantique financière pour améliorer sa compréhension du langage de cette domaine. Enfin, après l'affinage sur notre jeu de données étiquetées, le modèle atteint une précision de test et un score F1 élevés.

Mots clés: Traitement automatique du langage naturel, grand modèle de langage, analyse du sentiment financier, analyse de la subjectivité, marché de la crypto-monnaie, marché boursier, médias sociaux, rapport d'analyste.

Méthodes de recherche: Intelligence artificielle et heuristique, analyse de contenu, recherche quantitative.

Abstract

Extracting actionable information systematically from unstructured textual data has been a persistent challenge in the domain of finance. In response, this dissertation focuses on developing and applying advanced natural language processing (NLP) methods to address this challenge. Our research presents three essays that propose different NLP models for extracting various types of information from financial texts, and demonstrate their applications in different financial markets.

The first essay uses NLP to study the connections between social media and the cryptocurrency market. We create a financial-oriented sentiment analysis model based on transformer, the latest NLP deep learning model architecture. The model is pretrained on a large amount of raw business news texts so that it learns financial language knowledge and vocabulary, and then finetuned on a labeled dataset to predict financial sentiment, on which it achieves high performance. We apply the model to social media posts concerning bitcoin from Twitter, Reddit, and Stocktwits. We combine the sentiment results with bitcoin market data to examine connections among social media sentiment, attention, disagreement, and bitcoin trading activities and returns. We show that 1) higher social media sentiment leads to higher bitcoin returns and trading volume, 2) higher social media attention and disagreement increase bitcoin price volatility, 3) positive changes in sentiment lead to a decrease in volatility, and 4) the magnitude of the impact of different social media varies.

The second essay delves deeper into the measurement of fine-grained financial-oriented sentiment. We develop a novel NLP model that can measure the sentiment targeted toward specific entities in the text, rather than the overall sentiment of a sentence as measured by the model in the first essay. First, we propose a prompt-based model architecture that achieves state-of-the-art performance on multiple benchmark datasets for general targeted sentiment analysis. Subsequently, with a high-quality human-annotated social media targeted financial sentiment dataset that we create, we finetune this model so that it can be specialized in measuring financial sentiment. We then apply the finetuned model to 23 million financial-oriented social media posts from different platforms to measure financial

sentiment toward 24 meme stocks (stocks that gain frenetic attention from retail investors on social media which is often accompanied by dramatic price movement) and 30 Dow Jones constituent stocks. Our results show that the sentiment measured by our model is positively correlated with price return and negatively correlated with price volatility, and that this correlation is stronger for meme stocks than for Dow Jones stocks. We further demonstrate that our model's sentiment measurement economically outperforms other representative existing financial sentiment measurements.

The third essay introduces a model to measure the subjectivity of financial text, which could provide additional information beyond sentiment. Subjectivity analysis, which involves the differentiation of subjective and objective statements, is an important topic in NLP, yet remains under-explored in the realm of finance. There's a lack of both models and labeled data dedicated to subjectivity in financial texts. To address this gap, we first train annotators with solid business education backgrounds to label the subjectivities of a set of financial texts. Next, we design a prompt-based transformer model that specifically incorporates financial vocabulary and is pre-trained on financial semantics tasks to enhance its domain language understanding. Finally, through finetuning on our labeled dataset, the model achieves high test accuracy and F1 score.

Keywords: Natural language processing, large language model, financial sentiment analysis, subjectivity analysis, cryptocurrency market, stock market, social media, analyst report.

Research methods: Artificial intelligence and heuristics, content analysis, quantitative research.

Table of contents

List of Tables

Chap 1.

Chap 2.

Chap 3.

List of Figures

List of abbreviations

BERT	Bidirectional Encoder Representations from Transformers
BOW	Bag Of Words
CNN	Convolutional Neural Network
DJ30	Dow Jones 30
EOS	End-Of-Sentence
FPB	Financial Phrase Bank
GPT	Generative Pretrained Transformer
LM Dictionary	Loughran & Mcdonald Financial Dictionary
LSTM	Long Short-Term Memory
MLM	Masked Language Model
MNLI	Multi-Genre Natural Language Inference
NLI	Natural Language Inference
NLP	Natural Language Processing
OOV	Out-Of-Vocabulary
PLM	Pretrained Language Model
QA	Question Answering
RNN	Recurrent Neural Network
RoBERTa	Robustly Optimized Bert Approach
SD	Standard Deviation
SOTA	State-Of-The-Art
SVM	Support Vector Machine
TSA	Targeted Sentiment Analysis

VAR	Vector Auto-Regression

To my beloved wife and parents.

Acknowledgements

The pursuit of PhD is a long journey, during which I often felt stressful, perplexed, and sometimes frustrated. But I hardly ever felt helpless or lonely, because along the way, I have been fortunate enough to have been accompanied by my supervisor, colleagues, family, and friends who have supported me so much.

I am profoundly grateful to my supervisor Prof. Gilles Caporossi for his unreserved support and guidance. He would always be there for discussion and help whenever I needed the most. His relentless passion and curiosity as a researcher have inspired me to keep exploring new ideas. And as a supervisor, he has not only given me the utmost support but also granted me the greatest freedom to pursue different ideas and paths. Merci Gilles, I will forever be indebted to your mentorship.

I am also deeply thankful to Prof. Denis Larocque and Prof. Hongping Tan. As members of my supervisory committee, they both have given invaluable comments and advises during my PhD journey. Beyond that, I also thank Denis for having interviewed and welcomed me into the HEC Data Science PhD program. And I was lucky to have crossed paths with Hongping, who introduced me to the research areas of NLP for finance and accounting, and guided me to navigate this unfamiliar field.

I also would like to thank my other coauthors, especially Xiaozhou Zhou, who has been a good research collaborator as well as a great friend. Moreover, I thank my fellow PhD colleagues with whom we shared similar experiences together and helped each other during the journey.

I would like to take this opportunity to thank the PhD program of HEC Montreal, Fin-ML & NSERC, Mitacs, TMX, Foundation J.A. DeSève, and the Financial Market Surveillance Intelligence Centre at UQAM, for their generous financial support.

Finally, my special thanks go to my beloved wife Yunmi and my parents. Their unconditional love and encouragement are the ultimate source of my motivation and courage, supporting every bit of progress I have ever made during my academic journey.

Chapter 1
Social Media Sentiment and Bitcoin Price Dynamics

Abstract[1]

Using NLP (natural language processing) and data from Twitter, Reddit, and Stocktwits, this study examines connections among social media sentiment, attention, disagreement, and bitcoin trading activity and returns. We show that 1) higher social media sentiment leads to higher bitcoin returns and trading volume, 2) higher social media attention and disagreement increase bitcoin price volatility, 3) positive changes in sentiment lead to a decrease in volatility, and 4) the magnitude of the impact from different social media varies.

[1] This essay is coauthored with Haibo Jiang (jiang.haibo@uqam.ca), Alexandre F. Roch (roch.alexandre_f@uqam.ca), and Xiaozhou Zhou (zhou.xiaozhou@uqam.ca).

It was presented at Financial Management Association (FMA) Annual Conference 2022, and has been accepted for presentation at FMA European Conference 2023, and French Finance Association Conference (AFFI) 2023.

1.1 Introduction

Despite the ongoing controversy, cryptocurrencies (digital currencies) have undoubtedly been considered by many as a new class of financial instruments, which can be used as an alternative currency, asset, and hedging instrument. Cryptocurrencies are based on blockchain technology, a cryptographic and decentralized technology that ensures the digitalization of trust and does not rely on any central authority such as governments or banks. As of October 2021, the total market capitalization of cryptocurrencies reached over $2,500 billion[2] with bitcoin being the most important and by far the largest cryptocurrency in terms of market capitalization (more than $1,000 billion, as of October 2021). Accompanying this sharp increase in cryptocurrency market size has been a remarkable engagement of retail investors and a lack of regulations. Cryptocurrency markets are less regulated than traditional financial markets because, as new global investable instruments traded 24 hours a day over the internet, having a globally legal and synchronized regulation system from all countries is quite difficult. In addition, traditional media is not always interested in timely reporting events involving cryptocurrency, which makes social media a primary source of information. The growing importance of social media in cryptocurrency trading, along with retail investors' considerable amount of time spent on general social media websites (e.g., Twitter) or financial-oriented social media websites (e.g., Reddit ad Stock-Twits), stamps cryptocurrencies as 'meme' type security. Meme security refers to the security whose price dynamics are mainly caused by sentiment on social media posts. For example, shares of GameStop stock skyrocketed more than 400% in one week in January 2021 and gained more than 1,600% for the whole month of January 2021. One of the underlying forces of this dramatic movement was amateur traders on "WallStreetBets", a popular online Reddit forum with more than 10 million active users, to bid up the stock price.

All these social media phenomena lead one to ponder their role in cryptocurrency trading activities. Today, the question is no longer whether social media affects cryptocurrency valuation, but how it affects it. With a large amount of data from Twitter, Reddit, and Stocktwits, this study first attempts to accurately measure the sentiment embedded in

[2] https://www.statista.com/statistics/730876/cryptocurrency-maket-value/

social media by using natural language processing (NLP) models and then investigates the extent to which this sentiment can be used to predict bitcoin trading activities and price dynamics. In addition, we also assess the impact of social media attention and disagreement on the bitcoin market.

Traditional media such as newspapers, online news media, and blogs have generally been one-way channels to communicate news and opinions to the general public. In these traditional channels, words and 'tone' are carefully chosen by journalists or newspaper editors in an effort to present unbiased information. Recent studies analyze firm's information release and document that both the 'tone' and the choice of words in firms' disclosure documents contain important information and are associated with company performance (Larcker and Zakolyukina (2012), Hobson et al. (2012), Allee and DeAngelis (2015)). However, with the rise of mobile technologies and online communities, social media has been considered a dynamic two-way channel of information updates (Cade (2018)). The popularity of social media provides both opportunities and challenges for the information environment. On one hand, social media provides an alternative that enables investors to communicate directly their analyses or views without editorial constraints. Due to its network effect, social media can diffuse information more rapidly among targeted groups of audiences. Consequently, the proliferation of social media helps to reduce information asymmetry among users and to mitigate adverse market reactions to negative news (Chen et al. (2014), Bartov et al. (2017), Tang (2018), Blankespoor et al. (2014), Lee et al. (2015)). On the other hand, social media platforms feature social transmission biases (Hirshleifer (2020)) and echo chamber effects. Pedersen (2021) shows the belief that spillovers from social network interactions can lead to excess volumes and volatility. Other empirical research in this area shows that social media widens the reach of false information and exacerbates investors' bias (Demarzo et al. (2003)).

Financial sentiment, broadly defined as the expressed view of a favorable or unfavorable prospect on the basis of an investor's beliefs, has been long posited as a determinant of asset price variation (Keynes (1936)). However, the question of how to accurately measure the sentiment embedded in social media is underexploited. Earlier studies on

textual analysis focused solely on the choice and the tone of words by counting the positive or negative words predefined by general-purpose dictionaries (Schrand and Walther (2000), McVay (2006), Larcker and Zakolyukina (2012), Allee and DeAngelis (2015)). Nevertheless, the positive and negative words defined by general-purpose dictionaries may not be suitable in the financial context. Loughran and McDonald (2011) show that almost three-fourths of the words identified as negative by the widely used Harvard dictionary are words typically not considered negative in a financial context. They further developed an alternative finance-specific sentiment lexicon which since then has been widely used in finance research (Engelberg et al. (2012), Garcia (2013), Chen et al. (2014), among many others[3]) for the analysis of formal financial statements (e.g., annual reports). In the context of social media, even sentiment derived by a finance-specific lexicon might be biased for several reasons. First, posts on social media often use non-standard informal English language (Liu et al. (2012)). Second, social media posts are often written in a social setting, and captures communications among a group of people with common interests (Park et al. (2015)). Third, languages used in social media posts present individuals' own views about the world (Back et al. (2010)). Dictionary-based measures may not be able to correctly identify the financial sentiment contained in these statements. In this paper, we contribute to the literature by filling this gap. Specifically, our study adopts a cutting-edge NLP model that is trained to measure the textual sentiment specifically in the context of finance. The model is far superior to dictionary-based methods in text understanding because it can capture the context and the order of words by treating a text as a sequence of words.

Our paper contributes to the literature in three important empirical dimensions. First of all, to the best of our knowledge, we are among the first to use an NLP model to measure financially-oriented sentiment embedded in social media. Several novel results emerge from our results. Second, we show that social media sentiment exhibits a Granger causality to future bitcoin returns and trading volumes, but not to future volatility. Social media sentiment instead has a contemporaneous (same-day) effect on volatility. More specifically, positive changes in sentiment lead on average to a decrease in volatility

[3] See Loughran and McDonald (2016) for surveys.

during the day. We also find that the impact of sentiment on bitcoin trading is different among the three social media platforms that we consider. Third, in addition to sentiment, our study also shows that a rise in attention and disagreement increases uncertainty by raising volatility and skewness, to a lesser extent.

The rest of the paper proceeds as follows. In Section 2, we describe the data that we use in our empirical tests. In Section 3, we discuss the construction of social media sentiment measures. In Section 4, we present and discuss the results of empirical tests. We offer robustness checks in Section 5. Section 6 concludes.

1.2 Data

In this study we use six different datasets: 1) social media textual data from Twitter, Reddit, and StockTwits, 2) articles released in traditional media outlets (e.g., Wall Street Journal), 3) daily Bitcoin price from Coinmarketcap.com, 4) intraday bitcoin transaction data from Kaiko, 5) Google search data series, and 6) daily financial index data, obtained from Yahoo Finance.

To obtain the sentiment embedded in social media, we first scrape bitcoin-related social textual data via the Application Programming Interface (API) provided by Twitter, Reddit, and StockTwits for the period between January 2017 and December 2020. Reddit and StockTwits typically feature discussions from more financially-savvy users and offer an advantage in extracting the sentiment of cryptocurrency traders, which may ultimately have an impact on bitcoin's short-term returns.[4] On the other hand, Twitter offers a relatively "noisy" sentiment because postings on Twitter also contain general news. For Twitter and Reddit, the posting messages are scraped with the keyword "Bitcoin". In StockTwits, one can filter a cryptocurrency with a hashtag that ends with ".X" (e.g., $BTC.X for Bitcoin). We use this convention to download all postings related to Bitcoin in StockTwits during our sample period. The scraped postings of these social media include posted messages, dates, and timestamps. For our sample period, our final bitcoin-related textual dataset includes 28.7 million messages from Twitter, 6.57 million from

[4] See Betzer and Harries (2021), Hu et al. (2021), Diangson (2021), Agrawal et al. (2018), and Awais and Yang (2021).

Reddits, and 1.14 million from StockTwits. With this dataset, we further measure the sentiment of every post/tweet using an NLP learning model, which is documented in Section 3. The sentiment of a post/tweet is a continuous numeric value between -1 (negative) and 1 (positive). The daily sentiment is then defined as the average sentiment and disagreement as the standard deviation of the sentiment of every message posted during a given day. Attention is proxied by the number of postings during the day as in Da et al. (2011).

In order to test whether social media sentiment has an additional impact on bitcoin beyond traditional media sentiment, we also compute sentiments of bitcoin embedded in the Wall Street Journal (WSJ), Dow Jones Newswires (DJN), and Reuters. The articles are collected via the Factiva database. Specifically, to avoid articles about bitcoin-related companies instead of bitcoin, we take bitcoin as our keyword and choose the cryptocurrency market as the main subject of our search criteria. Finally, we obtain a total of 1,450 articles published in the three traditional media. The numbers of articles for DJN, Reuters, and WSJ are 614, 577, and 259, respectively. On average, there are at least two articles per day that can be used to compute the sentiment of traditional media during our sample period. As for sentiment computation, we apply the same NLP algorithm that was used for social media postings.

Our daily bitcoin price and volume data are from Coinmarketcap.com, which is a leading source of cryptocurrency data. It collects and aggregates information from over 200 major exchanges and provides daily data on open, close, high, low prices, and volume. For each cryptocurrency, Coinmarketcap.com calculates its volume-weighted price of all prices reported at each exchange. To conduct our analysis of realized volatility, we also use intraday-level data. The intraday transaction data used in this paper are from the leading cryptocurrency market data provider Kaiko. Its raw cryptocurrency data covers 20,000+ pairs across worldwide exchanges. Our dataset is at the tick-by-tick level, including unique trade id, exchange code, currency pairs, prices, volumes, trade directions, and timestamps, for all exchanges where Bitcoin is traded.

Google search data series for the word ”Bitcoin” are downloaded from Google. We further reconstruct Google trends daily data as in Liu and Tsyvinski (2020). The market daily index series are from Yahoo finance. The indexes used to capture financial market dynamics include SP&500, MSCI Global, Gold Index, USD Index, VIX, and U.S. Treasury bond yield.

1.3 Social Media Sentiment

1.3.1 Background

With the growing availability of digital textual data and computing technology, the measurement of financial sentiment embedded in texts has received increased research interest. Some pioneering papers (Tetlock, 2007; Kothari, Li and Short, 2009) popularized the simple dictionary-based approach (i.e., counting within a text the presence of positive and negative words predefined by a sentiment dictionary). Later, Loughran and McDonald (2011) pointed out that the general-purpose dictionaries commonly used by previous researchers often misclassify words in the financial context. They curated a finance-specific sentiment lexicon, which has since been broadly adopted by other researchers in the domain of finance. Besides the dictionary-based approach, classical machine learning methods such as naïve Bayes have also been widely explored (Antweiler and Frank, 2004; Das and Chen, 2007; Huang, Zang and Zheng, 2014). Those methods are usually supervised, training statistical models that learn from examples of texts with sentiment labeled by human experts. As shown in the aforementioned papers, finance researchers traditionally rely on the bag-of-words methods, which treat a text as a collection of independent words that can be represented by a vector of word counts. Due to their simplicity, the bag-of-words methods lack the ability to capture the context and the order of words which are crucial for interpreting the semantics of a text.

Recently, the development of natural language processing techniques has introduced more sophisticated models that are capable of recognizing the sequential nature of text and preserving the dependencies between words. The past decade has witnessed rapid advances in the field of NLP (Mikolov, Chen, Corrado and Dean, 2013; Bahdanau, Cho and Bengio, 2014; Cho, van Merrienboer, Gulcehre, Bahdanau, Bougares, Schwenk and

Bengio, 2014; Pennington, Socher and Manning, 2014), with the help of deep learning, a subset of machine learning that features deep neural network models capable of tackling complex unstructured data such as texts and images. Specifically, revolutionary breakthroughs have been achieved recently by the novel transformer-based language models (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser and Polosukhin, 2017) such as GPT (Generative Pretrained Transformer) (Radford, Narasimhan, Salimans and Sutskever, 2018; Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry and Askell, 2020) and BERT (Bidirectional Encoder Representations from Transformers) (Devlin, Chang, Lee and Toutanova, 2019), which successfully bring the model performance on many NLP tasks to the human level. Those transformer models are gigantic in size, often have hundreds of millions of parameters. A such model is first pretrained on large unlabeled text datasets like the whole Wikipedia corpus, so that it can encode abundant linguistic knowledge. Then it only requires a small amount of labeled data to be finetuned on specific tasks such as sentiment analysis, due to its ability to transfer the knowledge it has learned from the unlabeled corpus to the downstream tasks. Incentivized by the breakthroughs in the NLP, some researchers begin to explore those cutting-edge models for their application in finance. Araci (2019) and A. H. Huang et al., 2022 both show that by pretraining the BERT model on finance specific corpora (the two papers both name their finetuned models as FinBERT) and then fine-tuning it on sentiment analysis, the model can achieve state-of-the-art performance for various financial sentiment analysis datasets.

In this paper, we develop our own transformer model named “FinRoBERTa” to compute the financial sentiment measurement on the social media postings associated with bitcoin. The backbone architecture of the FinRoBERTa is the cutting-edge RoBERTa (Robustly optimized BERT approach) model (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer and Stoyanov, 2019), an improved derivative of the BERT model. We pretrain the model on finance domain corpora and fine-tune it on financial sentiment data. Our model achieves state-of-the-art performance for financial sentiment analysis on the test set.

1.3.2 FinRoBERTa Financial Sentiment Model

Model Pretraining

The purpose of pretraining a language model is to leverage large text corpora as self-labeled data to teach the model language knowledge. The RoBERTa model that we adopt in this paper uses the Masked Language Model (MLM) technique in which we randomly mask some of the words from the input text, and let the model try to predict the masked word based on its context (Devlin et al., 2019).

The original base version of the RoBERTa model has 12 layers of the transformer neural network modules, comprising 110 million parameters. It's pretrained on 160GB of general English-language corpora including books, news, online texts, etc., and takes days with over a thousand Nvidia V100-32GB GPUs, a massive amount of computing resources (Liu et al., 2019). Through pretraining, the model learns rich general language knowledge. However, the word distribution of financial corpora can be quite distinct from that of the general corpora because the financial domain uses a lot of its own techincal jargon. Researchers in different specialized domains have reported that a transformer model pretrained on domain corpora can outperform the generic model on domain-specific tasks (Huang, Altosaar and Ranganath, 2019; Lee, Yoon, Kim, Kim, Kim, So and Kang, 2019). In this regard, we pretrain the FinRoBERTa model from scratch on a 2.6GB English corpora of 2.5 million financial news collected from Factiva database. To reduce the computational needs, we adopt a smaller version of the RoBERTa model architecture with 6 transformer layers totaling 57M parameters, and decrease the vocabulary size from the original 50K to 30K.

The model was implemented using the Huggingface Transformers python library (Wolf, Chaumond, Debut, Sanh, Delangue, Moi, Cistac, Funtowicz, Davison and Shleifer). We pretrain the model for 4 epochs (cycles over the whole dataset) on a server equipped with 4 Nvidia V100-16GB GPUs.[5] The pretraining took around 80 hours.

[5] The computing resource is supported by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca).

After pretrained on the financial corpora, the FinRoBERTa model can demonstrate a grasp of financial language knowledge when fulfilling the MLM task, as shown in the example in Table 1.

[Insert Table 1 here]

It is also worth noticing that the vocabulary the model learns from the financial corpora features many common financial technical terms that are not captured by the generic RoBERTa vocabulary, such as “IPO”, “EPS”, “ROE”, “EBITDA”, “GAAP”, “CFA”, “WSJ”, etc. This added awareness of finance terminology would contribute to its superior performance on finance specific tasks.

Model Fine-Tuning

Fine-tuning is the process of further training the pretrained model on the labeled data of a target task such as sentiment analysis, so that the model learns to solve the specific task. Since the pretrained language model has already encoded abundant language knowledge, it can solve the end task much better given a limited amount of labeled training data, compared to traditional machine learning models that are trained merely on those labeled data.

The fine-tuning is performed with the financial phrase bank (FPB) (Malo, Sinha, Korhonen, Wallenius and Takala, 2014). It contains around 5,000 sentences randomly selected from financial news. Each sentence is manually labeled with the financial sentiment as either negative, neutral, or positive, independently by 5 to 8 annotators with adequate background in finance and business. The data is divided into 4 subsets that each contain sentences meeting a certain agreement level, i.e., the percentage of annotators agreeing on the same label for a sentence, namely 100%, >75%, >66% and >50%. Crucially, the financial sentiment here is different from ordinary sentiment: it’s defined as the potential impact of a new information on future financial events from an investor’s point of view.

We follow the same fine-tuning method as in the RoBERTa and BERT papers, by appending a 2-layer neural network for classification to the Fin-RoBERTa model, that

takes the sentence embedding from the transformer layers as input and outputs the predicted probabilities for the 3 sentiment classes: Negative, Neutral, and Positive. The training criteria is the crossentropy loss which measures the divergence of the predicted class probabilities from the true class. The loss of each class is adjusted by a weight of $1/\sqrt{\%\ of\ the\ class\ in\ the\ data}$ to alleviate the class imbalance issue.

Considering that a sentence without a clear majority agreement by people should be very vague in its sentiment, it's hard to justify using it as golden standard for the model. So, we chose to fine-tune our model only on the FPB subdatasets with 100%, >75%, and >66% agreement separately (FPB-100, FPB-75, FPB-66). We split each data into 3 sets: 60% for training, 20% for validation, and 20% for test.

The test results on the different datasets are shown in Table 2. The performance is in par with the state-of-the-art performance by similar finance specific transformer models of even larger size (Araci, 2019; H. Huang et al., 2022), and it exceeds the traditional dictionary approach (Loughran and McDonald, 2011) and classical machine learning approach (Malo et al., 2014) by a large margin.

[Insert Table 2 here]

1.3.3 Sentiment Measurement on Bitcoin Social Media Postings

We apply the fine-tuned FinRoBERTa model to measure the financial sentiment of the social media postings described in the last section. In view of the trade-off between label quality and data quantity, we choose to use the model fine-tuned on the FPB-75 dataset. We first clean the texts to get rid of the noises such as web address. Then we input each text to the fine-tuned FinRoBERTa model to predict its financial sentiment. The output sentiment score is between -1 and $+1$, calculated as the predicted probability of being positive minus that of being negative, so that a more negative (positive) score means more negative (positive) sentiment. A score of 0 is interpreted as neutral since it has an equal probability of being positive as being negative. The measured financial sentiment distributions of postings from different social media platforms are shown in Table 4.

To further validate our measurement, we manually compare the financial sentiment measured by our FinRoBERTa model against the general sentiment measured by a conventional NLP model and the financial sentiment measured with the classical Loughran-McDonald dictionary. For the general sentiment, we apply TextBlob[6], a popular NLP library that uses an expert-crafted English sentiment lexicon and linguistic rules to measure sentiment in a text. By manually examining the sample results, we confirm that our financial sentiment captures well a text's financial implication from an investor's perspective, whereas the general sentiment and the dictionary-based financial sentiment often fail badly. Table 3 shows 10 representative examples that compare the three sentiment measures.

[Insert Table 3 here]

1.4 Empirical Results

1.4.1 Summary Statistics

Table 4 presents the summary statistics for the variables used in our study. Panel A shows the key statistics of bitcoin daily returns, volumes (in million of bitcoins, denoted by Vol), sentiment (denoted by Sent), number of postings (denoted by Nb), and disagreement (denoted by Dis). For sentiment-related variables, we present the statistics for the three social media sources (i.e., Twitter, Reddit, and StockTwits). Bitcoin has a daily average return of 0.07% with a skewness of -1.48 and kurtosis of 21.08, suggesting an overall increase during the analyzed period but accompanied by more negative observations and a relatively large number of extreme values. Figure 1 plots daily realized variance and kurtosis from January 2018 to January 2021. The realized variance and kurtosis were at their peak in March 2020. The corresponding trading volume during this period has a mean 18.62 million with a standard deviation of 13.80 m which implies that 68% of observations fall into the large interval between 4.82 and 32.42 billion dollars. The average sentiment for each of the three social media is all slightly positive (0.04 for Twitter, 0.02 for Reddit, and 0.03 for StockTwits) during the sample period, in line with an overall positive sentiment during the sample period. The number of postings, a proxy

[6] https://textblob.readthedocs.io/en/dev/

for attention, varies with social media. Twitter, the largest social media in the world, has a mean of 19.45 thousand bitcoin-related postings per day. Reddit (Finance subreddit) and StockTwits are more financially oriented social media and contain, on average, 4.1 and 1.0 thousand postings per day, respectively. Another sentiment-related measure, disagreement, is around 0.27 and remains stable among three social media.

[Insert Table 4 here]

Table 5 presents the correlation matrix for bitcoin returns, social media sentiments (raw and orthogonal), and traditional media sentiments. The orthogonal sentiments are the residuals from the regression of sentiments on lagged bitcoin returns. On average, the correlations between return and the three social media are around 0.45, which is quite similar for orthogonal sentiment but much higher than the correlation between return and traditional media (0.13).

[Insert Table 5 here]

1.4.2 The Impact of Social Media on Bitcoin Volume and Return

We first look at how social media affects bitcoin daily trading dynamics such as returns and volume. We estimate the following vector autoregression (VAR) model with daily bitcoin returns and various social media measures (sentiment, attention and disagreement).

$$x_t = c_x + \sum_{\tau=1}^{2} \alpha_{x,\tau}\, x_{t-\tau} + \sum_{\tau=1}^{2} \alpha_{y,\tau}\, y_{t-\tau} + \alpha_z Z_{t-1} + \varepsilon_{x,t}, \qquad (1)$$

$$y_t = c_y + \sum_{\tau=1}^{2} \beta_{x,\tau}\, x_{t-\tau} + \sum_{\tau=1}^{2} \beta_{y,\tau}\, y_{t-\tau} + \beta_z Z_{t-1} + \varepsilon_{y,t}, \qquad (2)$$

where x_t and y_t are variables of interest on day t, which include bitcoin daily returns (Ret_t), bitcoin daily trading volume (Vol_t), social media sentiment (i.e., Twitter $Sent_t^T$, Reddit $Sent_t^R$, or Stocktwits $Sent_t^S$), social media attention (Nb_t), and disagreement (Dis_t). Z_{t-1} represents control variables (e.g., traditional media sentiment ($Sent_{t-1}^{trad}$)).

The results are reported in Table 6. Coefficients of the sentiment of day $t-1$ on returns on day t are statistically significant and positive for Twitter and StockTwits, suggesting that a higher social media sentiment on a given day can lead to a positive bitcoin return the next day. The opposite causality, i.e., a positive bitcoin return on day $t-1$ also leads to a higher social media sentiment on day *t* given that the coefficients of bitcoin returns are also statistically significant and positive. However, this Granger causal relationship of bitcoin returns on sentiment is smaller than that of social media sentiment on bitcoin returns. It's also worth noticing that higher sentiment at day $t-2$ inversely predicts a lower return on day t, which may suggest a reversal effect caused by a correction of the initial market overreaction.

[Insert Table 6 here]

We further apply the VAR model to daily bitcoin trading volume and the three social media sentiment measures. The results (also in Table 6) show that a positive sentiment of Twitter and Reddit on bitcoin can lead to a significant increase in bitcoin trading volume. However, the opposite is not true in a statistically significant way. Combined with the results of bitcoin returns and sentiment, we conclude that a positive sentiment results in a stronger buy intention and then leads to higher returns. The results further indicate that attention, measured by the number of postings, has a time-varying impact on trading volume. Specifically, more attention in social media can lead to an increase in trading volume next day. However, this increase will be offset by a decrease in trading volume in two days.

Finally, social media disagreement of Twitter and Reddit has a significant net negative impact on bitcoin trading volume, while social media disagreement of StockTwits has a significant positive impact on bitcoin trading volume. The intuition is that when there is more difference in opinion, investors tend to trade less the cryptocurrency. On the other hand, a higher trading volume can lead to different levels of disagreement for Twitter and Reddit. Recall that the profiles of Twitter and Reddit users are more general than those of StockTwits. Our results suggest that the opinion divergence is more persistent in Twitter and Reddit than that in StockTwits which contains more financially oriented users.

It is worth noting that the sentiment of traditional media does not affect the bitcoin return and trading volume in general. When the sentiment of traditional media is used as a control variable in various VAR models as reported in Table 6, its coefficients are not significant except for one case. In the VAR model of bitcoin return and the sentiment of traditional media, the lagged traditional media sentiment has insignificant coefficient, while the past bitcoin returns have positive and significant impact on the traditional media sentiment.

As alternatives to the raw measures, we also calculate the abnormal sentiment, attention, and disagreement, defined as a measure at day t minus the average of that measure during the past 5 days, e.g., $AbnSent_t = Sent_t - \sum_{n=1}^{5}(Sent_{t-n}/5)$. Table 7 shows the results with those abnormal measures by applying the same VAR model as above. Similarly, we find postive correlation between abnormal sentiment(attention) at day t-1 and return(volume) at day t.

[Insert Table 7 here]

1.4.3 Determinants of Bitcoin Price Volatility and Higher Moments

We now turn our attention to the impact of social media on bitcoin price volatility, skewness and kurtosis. We consider the following OLS model:

$$M_oM_t = \beta_0 + \beta_1 \times MoM_{t-1} + \beta_2 \times SocMe_{t-1} + \beta_3 \times Ret_{t-1} + \beta_z Z_t + \epsilon_t, \qquad (3)$$

$$RV_t = \sum_{i=1}^{n} r_{i,t}^2, \qquad (4)$$

$$Skew_t = \frac{\sum_{i=1}^{n} r_{i,t}^3}{(n-1) \times \sigma_t^3}, \qquad (5)$$

$$Kurt_t = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{1}^{N} r_{i,t}^4}{\sigma_t^4}, \qquad (6)$$

where MoM_t stands for bitcoin daily realised variance (RV_t), skewness ($skew_t$), or kurtosis ($kurt_t$) on day t, all three defined as the median over the nine most active bitcoin

exchanges.[7] $SocMe_t$ corresponds to the social media related variables (i.e., sentiment, attention, and disagreement) and Ret_{t-1} is the lagged bitcoin daily return. $r_{i,t}$ is the i-th 5-min bitcoin return on day t and $n = 288$ is the number of 5-min interval during the day. Z_t represents control variables (e.g., traditional media sentiment ($Sent_t^{trad}$))

Table 8 shows that, without including the lagged bitcoin return as one of control variables, social media sentiment has a significant negative impact on realized volatility. However, when controlling with lagged bitcoin returns, the social media sentiment no longer has a significant impact on bitcoin future price volatility. In these three cases, lagged bitcoin return has a significant negative impact, providing evidence on the phenomenon known as the leverage effect in asset pricing literature (Bollerslev et al. (2006), Carr and Wu (2017), and among others).

[Insert Table 8 here]

Table 9 provides mixed evidence, after controlling for lagged returns, that social media sentiment has little or no impact on daily return skewness but a significant impact on daily kurtosis. Given that sentiment has a significant positive impact on bitcoin returns, the result implies that positive sentiment is likely to cause more extreme bitcoin returns, but not a mild asymmetry of returns.

[Insert Table 9 here]

By putting all three social media variables together, Table 10 confirms that sentiment does not affect future volatility, but attention and disagreement do. Regarding the intraday bitcoin return skewness, only the coefficients of sentiment and attention from StockTwits are positively significant. Further, disagreement of Twitter and Reddit have a positive significant impact on bitcoin return skewness, suggesting that when there is a divergence in social media sentiments, it is more likely to observe more positive intraday bitcoin returns. Finally, the bitcoin intraday returns' kurtosis is related to sentiment and disagreement, but not the attention. More specifically, a more positive (negative) sentiment from Twitter and Stocktwits results in an increase in the probability of

[7] These nine exchanges are Bibox, BeQuant, BitForex, Bit-Z, Binance, EXX, Huobi, OkEX, and ZB.

extremely positive (negative) returns. Also, our results suggest that disagreement from Twitter and Reddit is also an important factor to drive more extreme observations.

[Insert Table 10 here]

Table 11 shows that it is not the lagged social media sentiment, but contemporaneous social media sentiment, that affects bitcoin price volatility. Our results show that the coefficient of sentiment variation ($\Delta S_t = S_t - S_{t-1}$) at day t has a significant negative impact on volatility at the same day, suggesting that a rise in positive sentiment can reduce bitcoin price volatility during the same day, even when the lagged returns variable is included as a control. The results in Table 11 also confirm the positive relation between sentiment variation and bitcoin intraday return skewness and the positive relation between lagged social media sentiment variation and bitcoin intraday return kurtosis.

[Insert Table 11 here]

1.5 Robustness Check

1.5.1 VAR Model with Additional Control Variables

In Table 12, we revisit the previous VAR models with common financial indices as controlled variables:

$$x_t = c_x + \sum_{\tau=1}^{2} \alpha_{x,\tau}\, x_{t-\tau} + \sum_{\tau=1}^{2} \alpha_{y,\tau}\, y_{t-\tau} + \sum_{\tau=1}^{2} \alpha_{z,\tau}\, z_{t-\tau} + \varepsilon_{x,t}, \qquad (7)$$

$$y_t = c_y + \sum_{\tau=1}^{2} \beta_{x,\tau}\, x_{t-\tau} + \sum_{\tau=1}^{2} \beta_{y,\tau}\, y_{t-\tau} + \sum_{\tau=1}^{2} \beta_{z,\tau}\, z_{t-\tau} + \varepsilon_{y,t}, \qquad (8)$$

where z_t is control variable for VAR model and x_t and y_t have the same definitions as in equations (1) and (2).

We used the following controlled variables in the above regressions: lagged traditional media sentiment, MSCI World Index, US dollar index (DXY), gold prices, Invesco DB Commodity Index, Dow Jones Commodity Index (DJCI), crude oil prices, SPDR S&P

500 ETF, VIX volatility index and Yield of U.S. 10-year treasury note (TNXT). Table 12 confirms the results of the relationship between return, volume, and social media related variables. Specifically, the results show that 1) a higher social media sentiment can lead to a positive bitcoin return next day, 2) a positive sentiment on bitcoin can lead to an increase in bitcoin trading volume, however, the opposite is not always true, 3) more attention in social media can lead to an increase in trading volume next day, however, this increase will be offset by a decrease in trading volume in two days, 4) social media disagreement has a significant net negative impact on bitcoin trading volume.

[Insert Table 12 here]

1.5.2 Principal Components of Social Media-Related Variables

$$MoM_t = \beta_0 + \beta_1 \times MoM_{t-1} + \beta_2 \times PC_{t-1}^{SocMe} + \beta_3 \times Ctrol_{t-1} + \epsilon_t, \qquad (9)$$

where PC_t^{SocMe} relates to the principal components of the corresponding social media related variables. MoM_t, $SocMe_t$, and $Ctrol_t$ have the same definitions as in equation (3).

Using principal components to capture information embedded in three social media-related variables, Table 13 indicates that lagged social media sentiment do not have impact on bitcoin price volatility, however, social media attention does have.

[Insert Table 13 here]

1.5.3 VAR Model with Google Trend as a Control Variable

Liu and Tsyvinski (2020) show that the investor attention significantly predicts one-week to six week ahead cumulative coin market returns. They use a weekly measure of Google search for "Bitcoin" as a proxy for investor attention. Following Liu and Tsyvinski (2020), we construct the deviation of Google searches for the word "Bitcoin" in a given day compared with the average of those in the preceding thirty days. We further standardize the daily deviation measure to have a mean of zero and a standard deviation of one.

Table 14 reports the results of VAR models including the lagged Google trend measures as the control variable. We confirm that the investor attention, measured in terms of Google trend measure, has a positive and significant impact on the next day bitcoin returns when it is included in the VAR model of bitcoin returns and social media sentiment. Nevertheless, compared to Table 6, results of social media sentiment remain robust, which means social media sentiment captures different and much richer information than Google searches for the word "Bitcoin".

[Insert Table 14 here]

1.6 Conclusion

Using a state-of-the-art NLP sentiment model and social media posts/tweets related to bitcoin from Twitter, Reddit, and Stocktwits, we investigate the relations and causality effects of social media sentiment, attention and disagreement, on bitcoin trading activity, returns, volatility and higher moments. First, we provide evidence of a reciprocal causality effect between higher social media sentiment and positive bitcoin returns, leading to a complex interplay between these two quantities. Furthermore, we show that positive bitcoin sentiment and increased attention (proxied by the number of posts/tweets) lead to an increase in trading volume in subsequent days.

The relation between volatility and sentiment is more subtle. We do not find any evidence that sentiment directly affects volatility, although it affects daily returns kurtosis. On the other hand, we provide evidence that positive changes in social media sentiment lead to a decrease in daily realized volatility, and an increase in daily returns skewness. We further show that higher social media attention and disagreement increase bitcoin price volatility. Overall, these findings are consistent among the three social media sources we used, although the magnitude of the impact from different social media varies.

References

Agrawal, S., Azar, P. D., Lo, A. W., & Singh, T. (2018). Momentum, mean-reversion, and social media: Evidence from stocktwits and twitter. *The Journal of Portfolio Management*, *44*(7), 85-95.

Allee, K. D., & DeAngelis, M. D. (2015). The structure of voluntary disclosure narratives: Evidence from tone dispersion. *Journal of Accounting Research*, *53*(2), 241-274.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of finance*, *59*(3), 1259-1294.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Awais, M., & Yang, J. (2021). Does Divergence of Opinions make better minds? Evidence from Social Media. Working Paper.

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological science*, *21*(3), 372-374.

Bahdanau, D., Cho, K., Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 15)*.

Bartov, E., Faurel, L., & Mohanram, P. S. (2018). Can Twitter help predict firm-level earnings and stock returns?. *The Accounting Review*, *93*(3), 25-57.

Betzer, A., Harries, J.P. (2021). How online comments affect stock trading - the case of gamestop. *Financial Markets and Portfolio Management*, forthcoming.

Blankespoor, E., Miller, G. S., & White, H. D. (2014). The role of dissemination in market liquidity: Evidence from firms' use of Twitter™. *The accounting review*, *89*(1), 79-112.

Bollerslev, T., Litvinova, J., & Tauchen, G. (2006). Leverage and volatility feedback effects in high-frequency data. *Journal of Financial Econometrics*, *4*(3), 353-384.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Cade, N. L. (2018). Corporate social media: How two-way disclosure channels influence investors. *Accounting, Organizations and Society*, *68*, 63-79.

Carr, P., & Wu, L. (2017). Leverage effect, volatility feedback, and self-exciting market disruptions. *Journal of Financial and Quantitative Analysis*, *52*(5), 2119-2156.

Chen, H., De, P., Hu, Y., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, *27*(5), 1367-1403.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 14)*, Association for Computational Linguistics. pp. 1724–1734.

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *The journal of finance*, *66*(5), 1461-1499.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management science*, *53*(9), 1375-1388.

DeMarzo, P. M., Vayanos, D., & Zwiebel, J. (2003). Persuasion bias, social influence, and unidimensional opinions. *The Quarterly journal of economics*, *118*(3), 909-968.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding, in *NAACL 19*, Association for Computational Linguistics. pp. 4171–4186.

Diangson, B., Jung, N. (2021). Bet if on reddit: The effects of reddit chatter on highly shorted stocks. Working paper.

Engelberg, J. E., Reed, A. V., & Ringgenberg, M. C. (2012). How are shorts informed?: Short sellers, news, and information processing. *Journal of Financial Economics*, *105*(2), 260-278.

Garcia, D. (2013). Sentiment during recessions. *The journal of finance*, *68*(3), 1267-1300.

Hirshleifer, D. (2020). Presidential address: Social transmission bias in economics and finance. *The Journal of Finance*, *75*(4), 1779-1831.

Hobson, J. L., Mayew, W. J., & Venkatachalam, M. (2012). Analyzing speech to detect financial misreporting. *Journal of Accounting Research*, *50*(2), 349-392.

Hu, D., Jones, C.M., Zhang, V., Zhang, X. (2021). The rise of reddit: How social media affects retail investors and short-sellers' roles in price discovery. Working paper .

Huang, A. H., Zang, A. Y., & Zheng, R. (2014). Evidence on the information content of text in analyst reports. *The Accounting Review*, *89*(6), 2151-2180.

Huang, K., Altosaar, J., Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342 .

Keynes, J.M. (1936). The general theory of employment, interest and money. London: Macmillan .

Kothari, S. P., Li, X., & Short, J. E. (2009). The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review*, *84*(5), 1639-1670.

Larcker, D. F., & Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, *50*(2), 495-540.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Lee, L. F., Hutton, A. P., & Shu, S. (2015). The role of social media in the capital market: Evidence from consumer product recalls. *Journal of Accounting Research*, *53*(2), 367-404.

Liu, F., Weng, F., & Jiang, X. (2012, July). A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1035-1044).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, Y., & Tsyvinski, A. (2021). Risks and returns of cryptocurrency. *The Review of Financial Studies*, *34*(6), 2689-2727.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*(4), 1187-1230.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance*, *66*(1), 35-65.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, *65*(4), 782-796.

McVay, S. E. (2006). Earnings management using classification shifting: An examination of core earnings and special items. *The accounting review*, *81*(3), 501-531.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, *108*(6), 934.

Pedersen, L.H. (2021). Game on: Social networks and markets. Working Paper.

Pennington, J., Socher, R., Manning, C.D. (2014). Glove: Global vectors for word representation, in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP 14)*, pp. 1532–1543.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Schrand, C. M., & Walther, B. R. (2000). Strategic benchmarks in earnings announcements: the selective disclosure of prior-period earnings components. *The Accounting Review*, *75*(2), 151-177.

Tang, V. W. (2018). Wisdom of crowds: Cross-sectional variation in the informativeness of third-party-generated product information on Twitter. *Journal of Accounting Research*, *56*(3), 989-1034.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, *62*(3), 1139-1168.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).

Huang, A. H., Wang, H., & Yang, Y. (2022). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research*.

List of Tables

Masked Input Text	"Dow *[mask]* 900 points for worst day of year amid fears of new Covid variant."				
Original word masked	"fell"				
Top 5 predicted words	"fell"	"falls"	"loses"	"rose"	"shed"
Predicted probability	0.30	0.05	0.04	0.04	0.03

Table 1: Example of the pretrained FinRoBERTa model performing the MLM task. The model is asked to predict the masked word given the rest of the input text.

Dataset \ Model	FinRoBERTa	RoBERTa Base
FPB-100	**0.9602**	0.7604
FPB-75	**0.9267**	0.7313
FPB-66	**0.8848**	0.6745

Table 2: Test accuracy of the FinRoBERTa model and the generic RoBERTa Base model on FPB dataset with different agreement levels. The lower the agreement level of the data is, the harder it is for any model to achieve high accuracy, because the sentiment in sentences with a lower agreement level are less clear even to an expert.

Text	TextBlob	LM Dict	FinRoBERTa
“If I use bitcoin as a store of value, transaction volume is not a very interesting metric for me. That being said, bitcoin transaction volume has been increasing when measured in terms of the goods and services that can be purchased with it.”	-0.192	0.000 (Pos: 1, Neg: 1)	0.943
“Why a Top Analyst Thinks Bitcoin Price Could Fall By 20% Before Bottoming”	0.500	0.000 (Pos: 0, Neg: 0)	-0.890
“that’s why I’m here asking what the best route is to buy a bitcoin”	1.000	0.3536 (Pos: 1, Neg: 0)	-0.004
“Good entry point or y’all waiting? No Moon boys please. I’m expecting Bitcoin to correct down to 10k so OMG should drop down to $10ish as well. Thoughts?”	0.130	0.000 (Pos: 1, Neg: 1)	-0.957
”Bitcoin gets 15% down in just a day. The cryptocurrency value shows its lowest level in months. Digital currency prices fell considerably for the second consecutive day due to the impact produced by Goldman Sachs and its decision to stop its plans to launch a persistent cryptocurrency desk. Ethical hacking specialists report that the price of a unit of Bitcoin, the most widely known digital currency in the world, fell by more than $1.1k USD in a period of 24 hours, representing a decrease...”	0.061	-0.198 (Pos: 0, Neg: 2)	-0.980
“And yet Bitcoin is slowly clawing back market dominance”	-0.150	0.000 (Pos: 1, Neg: 1)	0.976
”Yes, BitMEX Liquidations Caused Bitcoin Price to Crash; Here’s How”	0.000	-0.354 (Pos: 0, Neg: 1)	-0.768
”So what’s wrong with Bitcoin Cash, in terms of its technological changes? So far, all I’ve heard is ”it’s too simple” even though it, thus far, has greatly improved the usability of Bitcoin as a currency.”	0.100	0.218 (Pos: 2, Neg: 1)	0.988

”Gold have like 5k years, also have industrial usage. Bitcoin is just money, nothing else. If can’t be the best on that, is done.”	1.000	0.267 (Pos: 1, Neg: 0)	-0.002
”They’ll hit the entry points. Bitcoin is going to outpace badly in the war though. Just by being an always available alternative to an ever growing list of inflationary and manipulated currencies built under a system that heavily favors the interests of the banks and lawmakers.”	-0.166	0.000 (Pos: 2, Neg: 2)	0.866

Table 3. Examples of bitcoin-related social media posts measured with 1) General sentiment by TextBlob, 2) Financial sentiment by Loughran-McDonald dictionary (LM Dict), and 3) Financial sentiment by our FinRoBERTa model. All scores range from -1 (most negative) to 1 (most positive). When using the dictionary to measure the sentiment of a text, we first delete the stopwords, i.e. extremely common words which have little value for determining the sentiment, such as “the”, “he”, “in”, “that”, etc. Then, we compute the total number of words left in the text, and count the number of positive and negative words in it according to the dictionary. Last, we calculate $\mathrm{p} = (\mathrm{num\_positive\_words} - \mathrm{num\_negative\_words})/\mathrm{total\_num\_words}$, and the $\mathrm{sentiment\ score} = \mathrm{sqr(p)}$ if $\mathrm{p} > 0$ and $-\mathrm{sqrt}(-\mathrm{p})$ if $\mathrm{p} <= 0$. In the “LM Dict” column, the $(\mathrm{num\_positive\_words}, \mathrm{num\_negative\_words})$ are also shown below the sentiment score.

	Bitcoin		Sentiment (Sent)			Number (Nb)			Disagreement (Dis)		
	Ret	Vol	Twitter	Reddit	StockTwits	Twitter	Reddit	StockTwits	Twitter	Reddit	StockTwits
Min	-0.46	2.92	-0.09	-0.06	-0.04	2.43	0	0.20	0.21	0	0.23
Max	0.17	74.16	0.16	0.23	0.10	69.73	16.08	8.68	0.39	0.46	0.39
Mean	0.0007	18.62	0.04	0.02	0.03	19.45	4.10	0.99	0.28	0.26	0.29
Std	0.04	13.80	0.02	0.02	0.02	9.06	1.67	0.86	0.03	0.03	0.02
Skewness	-1.48	0.95	-0.43	1.01	-0.09	1.81	2.45	3.74	0.47	-0.39	0.26
Kurtosis	21.08	3.39	5.52	15.10	2.74	7.21	12.16	23.94	3.66	14.24	3.35

Table 4: Descriptive statistics. Volume is in millions, number of posts is in thousands.

	Returns	Twitter	Reddit	StockTwits	Traditional Media	Twitter Orth	Reddit Orth
Twitter	0.4415						
Reddit	0.3247	0.7663					
StockTwits	0.4178	0.6395	0.5130				
Traditional Media	0.1339	0.3311	0.2754	0.1756			
Twitter Orth	0.5497	0.8617	0.6704	0.5720	0.3015		
Reddit Orth	0.3751	0.6222	0.9284	0.4369	0.2404	0.7221	
StockTwits Orth	0.4555	0.5148	0.4237	0.9573	0.1410	0.5975	0.4564

Table 5: Correlation matrix for bitcoin returns, sentiments and orthogonal sentiments. The orthogonal sentiments are the residuals from the regression of sentiments on lagged bitcoin returns.

	Vars X_t, Y_t	Y_{t-1}		Y_{t-2}		X_{t-1}		X_{t-2}		$Sent_{t-1}^{trad}$	
Twitter	Ret, Sent	0.2513***	(2.6693)	-0.1317**	(-1.7356)	0.1864***	(8.6359)	0.0180	(0.8355)	-0.0048	(-1.2749)
	Ret, Nb	0.0034	(0.1462)	-0.0078	(-0.3342)	-0.1382***	(-2.4015)	0.0886*	(1.5410)	-0.0029	(-0.7881)
	Ret, Dis	0.0609	(0.7765)	0.0869	(1.1086)	-0.0226*	(-1.3152)	-0.0218	(-1.2752)	-0.0023	(-0.6414)
	Ret, Vol	0.0003	(0.0135)	0.0157	(0.7010)	-0.0250	(-0.4152)	0.0495	(0.8343)	-0.0035	(-0.9609)
	Vol, Nb	0.1235***	(3.1948)	-0.1535***	(-3.9793)	-0.0344	(-0.9513)	0.0284	(0.7814)	0.0053	(0.9250)
	Vol, Sent	0.1884*	(1.5578)	0.0278	(0.2392)	0.0144	(1.2822)	-0.0071	(-0.6351)	0.0031	(0.5061)
	Vol, Dis	0.0464	(0.3478)	-0.2222**	(-1.6805)	0.0263***	(2.4449)	-0.0174*	(-1.6232)	0.0066	(1.1453)
Reddit	Ret, Sent	0.0838	(0.9083)	-0.0712	(-0.8553)	0.1380***	(8.2014)	0.0226*	(1.3159)	-0.0032	(-0.8511)
	Ret, Nb	-0.0425*	(-1.4811)	0.0136	(0.4795)	-0.0945**	(-2.0484)	0.0092	(0.1992)	-0.0031	(-0.8630)
	Ret, Dis	0.0479	(0.7494)	-0.0144	(-0.2254)	-0.0320*	(-1.5240)	-0.0239	(-1.1445)	-0.0027	(-0.7372)
	Ret, Vol	0.0003	(0.0135)	0.0157	(0.7010)	-0.0250	(-0.4152)	0.0495	(0.8343)	-0.0035	(-0.9609)
	Vol, Nb	0.0321	(0.6826)	-0.0829**	(-1.7720)	0.0131	(0.4623)	-0.0274	(-0.9668)	0.0064	(1.1199)
	Vol, Sent	0.2032*	(1.5066)	0.0736	(0.5632)	-0.0014	(-0.1386)	0.0042	(0.4268)	0.0038	(0.6329)
	Vol, Dis	0.1037	(0.9982)	-0.1464*	(-1.4207)	0.0171*	(1.3448)	-0.0175*	(-1.3704)	0.0072	(1.2587)
StockTwits	Ret, Sent	0.2302***	(2.6161)	-0.1082*	(-1.3144)	0.0813***	(4.1014)	0.0363**	(1.8552)	-0.0034	(-0.9199)
	Ret, Nb	-0.0533***	(-2.5297)	0.0552***	(2.5976)	-0.0953*	(-1.4889)	0.0458	(0.7182)	-0.0024	(-0.6662)
	Ret, Dis	-0.0102	(-0.1359)	-0.0781	(-1.0334)	0.0212	(1.1751)	-0.0056	(-0.3123)	-0.0035	(-0.9349)
	Ret, Vol	0.0003	(0.0135)	0.0157	(0.7010)	-0.0250	(-0.4152)	0.0495	(0.8343)	-0.0035	(-0.9609)
	Vol, Nb	0.1305***	(3.3520)	-0.2123***	(-5.5364)	-0.1477***	(-3.3026)	0.1424***	(3.1738)	0.0053	(0.9398)
	Vol, Sent	0.1323	(1.0662)	0.0306	(0.2487)	0.0005	(0.0437)	0.0002	(0.0204)	0.0056	(0.9548)
	Vol, Dis	-0.0492	(-0.4033)	0.1776*	(1.4640)	0.0030	(0.2751)	-0.0082	(-0.7522)	0.0082*	(1.4048)
Traditional Media	Ret, Sent	0.0001	(0.0278)	0.0030	(0.9588)	1.3223***	(4.4484)	0.9862***	(3.2927)		
	Vol, Sent	0.0051	(0.9727)	0.0159***	(3.0529)	0.1405	(0.8220)	0.1053	(0.6153)		

Table 6. VAR models with two lags results on the impact of social media on bitcoin volume and return. The two variables (e.g., "Ret, Sent") in the first column for each row represent X_t and Y_t of a VAR model defined in equations (1 & 2), correspondingly. Coefficients

Y_{t-1}, Y_{t-2} are the loadings in the equation (1) for X_t, and coefficients X_{t-1}, X_{t-2} are the loadings in the equation (2) for Y_t.

	Vars X_t, Y_t	Y_{t-1}		Y_{t-2}		X_{t-1}		X_{t-2}		$Sent_{t-1}^{trad}$	
Twitter	Ret, AbnSent	0.102*	(1.38)	-0.076	(-1.20)	0.153***	(9.05)	-0.040**	(-2.25)	-0.000	(-0.01)
	Ret, AbnNb	-0.007	(-0.40)	0.006	(0.37)	-0.132***	(-2.49)	0.071*	(1.34)	0.001	(0.18)
	Ret, AbnDis	0.068	(1.22)	0.016	(0.28)	-0.038**	(-2.31)	-0.000	(-0.02)	0.000	(0.12)
	Vol, AbnNb	0.107***	(3.40)	-0.152***	(-4.85)	0.001	(0.03)	-0.011	(-0.34)	0.005	(1.05)
	Vol, AbnSent	0.129	(1.24)	-0.031	(-0.30)	0.011	(1.27)	-0.013*	(-1.53)	0.006	(1.09)
	Vol, AbnDis	-0.062	(-0.61)	-0.238***	(-2.36)	0.039***	(3.90)	-0.025***	(-2.51)	0.007*	(1.29)
Reddit	Ret, AbSent	0.061	(0.83)	0.012	(0.18)	0.119***	(8.31)	-0.002	(-0.13)	0.000	(0.08)
	Ret, AbnNb	-0.041**	(-1.87)	0.016	(0.74)	-0.113***	(-2.67)	0.032	(0.75)	0.000	(0.13)
	Ret, AbnDis	0.036	(0.74)	0.005	(0.10)	-0.043**	(-2.30)	-0.022	(-1.18)	0.001	(0.16)
	Vol, AbnNb	0.075**	(1.95)	-0.148***	(-3.85)	0.028	(1.11)	-0.049**	(-1.98)	0.005	(1.01)
	Vol, AbSent	0.194**	(1.70)	-0.080	(-0.71)	0.007	(0.88)	-0.010	(-1.21)	0.006	(1.12)
	Vol, AbnDis	0.010	(0.12)	-0.191**	(-2.26)	0.037***	(3.43)	-0.041***	(-3.77)	0.007*	(1.35)
StockTwits	Ret, AbSent	0.129**	(1.96)	-0.051	(-0.79)	0.062***	(3.58)	0.003	(0.19)	0.000	(0.13)
	Ret, AbnNb	-0.026*	(-1.59)	0.039***	(2.35)	-0.094**	(-1.69)	0.098**	(1.76)	0.001	(0.20)
	Ret, AbnDis	0.015	(0.27)	-0.045	(-0.81)	0.009	(0.55)	0.008	(0.45)	0.000	(0.09)
	Vol, AbnNb	0.147***	(4.48)	-0.245***	(-7.60)	-0.175***	(-4.63)	0.175***	(4.64)	0.006	(1.10)
	Vol, AbnSent	0.079	(0.78)	0.028	(0.28)	0.009	(0.97)	-0.008	(-0.93)	0.007	(1.28)
	Vol, AbnDis	-0.172**	(-1.82)	0.156*	(1.64)	0.018**	(1.88)	-0.027***	(-2.81)	0.007*	(1.39)

Table 7. Results of the same VAR models as in table 6 applied to abnormal measures.

Const	RV_{t-1}	Ret_{t-1}	$Sent_{t-1}^{T}$	$Sent_{t-1}^{R}$	$Sent_{t-1}^{S}$	$Sent_{t-1}^{trad}$
0.0014***	0.4934***		-0.0113**			-0.0004
(5.19)	(18.20)		(-2.13)			(-1.11)
0.0012***	0.4944***			-0.0132**		-0.0004
(6.23)	(18.28)			(-2.07)		(-1.28)
0.0015***	0.4890***				-0.0180***	-0.0004
(6.74)	(18.32)				(-3.32)	(-1.35)
0.0006**	0.4863***	-0.0284***	0.0086			-0.0004
(2.12)	(18.54)	(-8.78)	(1.54)			(-1.28)
0.0009***	0.4787***	-0.0267***		0.0022		-0.0003
(4.93)	(18.24)	(-8.66)		(0.34)		(-0.93)
0.0010***	0.4767***	-0.0264***			0.0001	-0.0003
(4.28)	(18.36)	(-8.23)			(0.02)	(-0.87)

Table 8. OLS regressions of realised bitcoin intraday returns variance on sentiment. RV is the median of daily realised variance over 9 most active exchanges. Independent variables are the lagged RV, bitcoin returns, and sentiment over the three social media sources and the traditional media. Sent_{t-1}^{T}, Sent_{t-1}^{R}, Sent_{t-1}^{S}, and $\mathrm{Sent}_{t-1}^{trad}$ represent the sentiment from Twitter, Reddit, StockTwits, and the traditional media, correspondingly.

Skewness						
Const	MoM_{t-1}	Ret_{t-1}	$Sent^{T}_{t-1}$	$Sent^{R}_{t-1}$	$Sent^{S}_{t-1}$	$Sent^{trad}_{t-1}$
0.0000***	-0.2263***		-0.0008***			-0.0000
(4.51)	(-7.73)		(-4.69)			(-1.09)
0.0000***	-0.2254***			-0.0007***		-0.0000*
(2.99)	(-7.65)			(-3.37)		(-1.75)
0.0000**	-0.2172***				-0.0004**	-0.0000**
(2.26)	(-7.38)				(-2.19)	(-2.34)
0.0000	-0.1457***	-0.0011***	0.0000			-0.0000
(0.34)	(-4.98)	(-9.79)	(0.09)			(-1.28)
0.0000	-0.1456***	-0.0011***		0.0000		-0.0000
(0.54)	(-4.99)	(-10.36)		(0.12)		(-1.31)
-0.0000	-0.1375***	-0.0012***			0.0005**	-0.0000
(-1.59)	(-4.77)	(-11.02)			(2.51)	(-1.63)

Kurtosis						
Const	MoM_{t-1}	Ret_{t-1}	$Sent^{T}_{t-1}$	$Sent^{R}_{t-1}$	$Sent^{S}_{t-1}$	$Sent^{trad}_{t-1}$
0.0000***	0.2331***		-0.0001**			-0.0000
(2.79)	(7.89)		(-2.31)			(-1.33)
0.0000**	0.2372***			-0.0001		-0.0000*
(2.04)	(8.01)			(-1.31)		(-1.76)
0.0000***	0.2370***				-0.0001**	-0.0000*
(2.79)	(8.07)				(-2.29)	(-1.78)
-0.0000*	0.2472***	-0.0002***	0.0001***			-0.0000
(-1.72)	(8.87)	(-11.84)	(2.82)			(-1.58)
-0.0000	0.2448***	-0.0002***		0.0001**		-0.0000
(-0.44)	(8.78)	(-11.88)		(2.26)		(-1.29)
-0.0000	0.2415***	-0.0002***			0.0001**	-0.0000
(-1.11)	(8.73)	(-11.77)			(2.53)	(-1.08)

Table 9. OLS regressions of higher bitcoin intraday returns moments (skewness and kurtosis) on sentiment. In top panel, the dependent variable MoM is the daily realised skewness. In bottom panel, MoM is daily realised kurtosis. In both cases, a median is taken over the 9 most active bitcoin exchanges. Independent variables are the lagged values of skewness or kurtosis, bitcoin returns, and sentiment over the three social media sources and the traditional media. $Sent^{T}_{t-1}$, $Sent^{R}_{t-1}$, $Sent^{S}_{t-1}$, and $Sent^{trad}_{t-1}$ represent the sentiment from Twitter, Reddit, StockTwits, and the traditional media, correspondingly.

	Const	MoM_{t-1}	Ret_{t-1}	$Sent^{T}_{t-1}$	Nb^{T}_{t-1}	Dis^{T}_{t-1}
$MoM = RV_t$	-0.0037***	0.4063***	-0.0297***	0.0048	0.0751***	0.0110**
	(-2.71)	(13.53)	(-9.29)	(0.88)	(5.57)	(2.25)
$MoM = Skew_t$	-0.0001***	-0.1566***	-0.0011***	-0.0001	0.0007	0.0004***
	(-3.04)	(-5.34)	(-9.76)	(-0.50)	(1.63)	(2.85)
$MoM = Kurt_t$	-0.0000**	0.2368***	-0.0002***	0.0001**	0.0001	0.0001*
	(-2.31)	(8.38)	(-11.85)	(2.29)	(1.14)	(1.82)
	Const	MoM_{t-1}	Ret_{t-1}	$Sent^{R}_{t-1}$	Nb^{R}_{t-1}	Dis^{R}_{t-1}
$MoM = RV_t$	-0.0041***	0.3800***	-0.0271***	-0.0004	0.5220***	0.0120***
	(-3.51)	(13.03)	(-9.02)	(-0.06)	(6.96)	(2.65)
$MoM = Skew_t$	-0.0001***	-0.1556***	-0.0011***	-0.0002	0.0011	0.0005***
	(-3.04)	(-5.32)	(-10.15)	(-0.85)	(0.45)	(3.08)
$MoM = Kurt_t$	-0.0000**	0.2357***	-0.0002***	0.0001	0.0001	0.0001**
	(-2.12)	(8.35)	(-11.80)	(1.43)	(0.19)	(2.06)
	Const	MoM_{t-1}	Ret_{t-1}	$Sent^{S}_{t-1}$	Nb^{S}_{t-1}	Dis^{S}_{t-1}
$MoM = RV_t$	-0.0001	0.3838***	-0.0291***	0.0032	0.9857***	0.0005
	(-0.05)	(12.89)	(-9.02)	(0.55)	(6.08)	(0.09)
$MoM = Skew_t$	0.0001	-0.1351***	-0.0013***	0.0006***	0.0116**	-0.0003
	(1.00)	(-4.69)	(-11.46)	(3.07)	(2.47)	(-1.52)
$MoM = Kurt_t$	0.0000	0.2367***	-0.0002***	0.0001***	0.0014	-0.0001
	(1.18)	(8.51)	(-12.09)	(2.98)	(1.60)	(-1.55)

Table 10. OLS regressions of daily realised variance, skewness and kurtosis on sentiment, attention, and disagreement. Independent variables are the lagged values of MoM, lagged bitcoin returns, lagged sentiment, lagged number of tweets / post (in millions), and lagged disagreement, taking on social media post at a time.

	Const	MoM_{t-1}	Ret_{t-1}	$\Delta Sent^{T}_{t-1}$	$\Delta Sent^{T}_{t}$	$\Delta Sent^{trad}_{t-1}$	$\Delta Sent^{trad}_{t}$
$MoM = RV_t$	0.0009***	0.4959***	-0.0258***	0.0063	-0.0269***	-0.0006**	-0.0002
	(7.59)	(19.70)	(-7.86)	(0.99)	(-4.63)	(-2.52)	(-0.98)
$MoM = Skew_t$	0.0000	-0.1507***	-0.0013***	0.0006***	0.0008***	-0.0000	0.0000
	(0.96)	(-5.29)	(-11.65)	(2.95)	(4.28)	(-0.32)	(0.55)
$MoM = Kurt_t$	0.0000*	0.2404***	-0.0002***	0.0001*	-0.0000	-0.0000	-0.0000
	(1.83)	(8.69)	(-11.19)	(1.87)	(-0.51)	(-1.61)	(-0.12)
	Const	MoM_{t-1}	Ret_{t-1}	$\Delta Sent^{R}_{t-1}$	$\Delta Sent^{R}_{t}$	$\Delta Sent^{trad}_{t-1}$	$\Delta Sent^{trad}_{t}$
$MoM = RV_t$	0.0009***	0.4875***	-0.0266***	0.0107*	-0.0168***	-0.0007***	-0.0004*
	(7.69)	(19.37)	(-8.68)	(1.67)	(-2.70)	(-2.70)	(-1.65)
$MoM = Skew_t$	0.0000	-0.1442***	-0.0012***	0.0004*	0.0003	0.0000	0.0000
	(0.92)	(-5.02)	(-11.27)	(1.71)	(1.21)	(0.24)	(1.47)
$MoM = Kurt_t$	0.0000*	0.2426***	-0.0002***	0.0001**	-0.0000	-0.0000	-0.0000
	(1.82)	(8.79)	(-11.70)	(1.96)	(-1.17)	(-1.50)	(-0.00)
	Const	MoM_{t-1}	Ret_{t-1}	$\Delta Sent^{S}_{t-1}$	$\Delta Sent^{S}_{t}$	$\Delta Sent^{trad}_{t-1}$	$\Delta Sent^{trad}_{t}$
$MoM = RV_t$	0.0009***	0.4853***	-0.0281***	0.0056	-0.0177***	-0.0007***	-0.0004*
	(7.72)	(19.26)	(-8.98)	(0.98)	(-3.25)	(-2.62)	(-1.80)
$MoM = Skew_t$	0.0000	-0.1423***	-0.0013***	0.0010***	0.0003*	0.0000	0.0000
	(0.95)	(-5.01)	(-12.25)	(4.96)	(1.91)	(0.04)	(1.37)
$MoM = Kurt_t$	0.0000*	0.2363***	-0.0002***	0.0001***	-0.0000	-0.0000	-0.0000
	(1.85)	(8.59)	(-12.30)	(3.32)	(-0.21)	(-1.56)	(-0.27)

Table 11. OLS regressions of daily realised skewness and kurtosis on sentiment variation. Independent variables are the lagged values of skewness or kurtosis, lagged bitcoin returns, and variations in sentiment defined as $\Delta Sent_t = Sent_t - Sent_{t-1}$.

Twitter											
Vars X_t, Y_t	Y_{t-1}		Y_{t-2}		X_{t-1}		X_{t-2}		$Sent_{t-1}^{trad}$		Other Control Variables
Ret, Sent	0.267***	(2.85)	-0.111*	(-1.46)	0.185***	(8.59)	-0.003	(-0.14)	-0.005	(-1.21)	Y
Ret, Nb	0.006	(0.25)	-0.010	(-0.44)	-0.136***	(-2.36)	0.076	(1.24)	-0.002	(-0.63)	Y
Ret, Dis	0.054	(0.69)	0.081	(1.04)	-0.023*	(-1.35)	-0.020	(-1.10)	-0.002	(-0.49)	Y
Ret, Vol	0.001	(0.04)	0.015	(0.66)	-0.032	(-0.53)	0.035	(0.56)	-0.003	(-0.78)	Y
Vol, Nb	0.133***	(3.45)	-0.163***	(-4.24)	-0.036	(-1.00)	0.031	(0.84)	0.006	(1.01)	Y
Vol, Sent	0.175*	(1.44)	0.045	(0.38)	0.016*	(1.42)	-0.009	(-0.80)	0.004	(0.62)	Y
Vol, Dis	0.059	(0.44)	-0.265**	(-2.00)	0.025**	(2.33)	-0.016*	(-1.51)	0.007	(1.24)	Y
Reddit											
Vars X_t, Y_t	Y_{t-1}		Y_{t-2}		X_{t-1}		X_{t-2}		$Sent_{t-1}^{trad}$		Other Control Variables
Ret, Sent	0.101	(1.11)	-0.052	(-0.62)	0.137***	(8.17)	0.007	(0.39)	-0.003	(-0.76)	Y
Ret, Nb	-0.042*	(-1.48)	0.014	(0.48)	-0.086**	(-1.87)	-0.002	(-0.04)	-0.003	(-0.70)	Y
Ret, Dis	0.042	(0.65)	-0.006	(-0.10)	-0.032*	(-1.53)	-0.026	(-1.16)	-0.002	(-0.57)	Y
Ret, Vol	0.001	(0.04)	0.015	(0.66)	-0.032	(-0.53)	0.035	(0.56)	-0.003	(-0.78)	Y
Vol, Nb	0.033	(0.71)	-0.086**	(-1.84)	0.019	(0.66)	-0.032	(-1.12)	0.007	(1.21)	Y
Vol, Sent	0.203*	(1.50)	0.095	(0.72)	-0.000	(-0.03)	0.003	(0.31)	0.004	(0.71)	Y
Vol, Dis	0.112	(1.07)	-0.168*	(-1.63)	0.017*	(1.34)	-0.018*	(-1.37)	0.008*	(1.35)	Y
StockTwits											
Vars X_t, Y_t	Y_{t-1}		Y_{t-2}		X_{t-1}		X_{t-2}		$Sent_{t-1}^{trad}$		Other Control Variables
Ret, Sent	0.244***	(2.79)	-0.068	(-0.82)	0.083***	(4.23)	0.014	(0.70)	-0.003	(-0.79)	Y
Ret, Nb	-0.052***	(-2.48)	0.054***	(2.56)	-0.089*	(-1.38)	0.030	(0.43)	-0.002	(-0.51)	Y
Ret, Dis	-0.006	(-0.08)	-0.095	(-1.25)	0.018	(1.03)	0.012	(0.61)	-0.003	(-0.78)	Y
Ret, Vol	0.001	(0.04)	0.015	(0.66)	-0.032	(-0.53)	0.035	(0.56)	-0.003	(-0.78)	Y
Vol, Nb	0.137***	(3.53)	-0.220***	(-5.76)	-0.146***	(-3.24)	0.141***	(3.12)	0.006	(1.05)	Y
Vol, Sent	0.127	(1.02)	0.036	(0.28)	0.003	(0.27)	-0.002	(-0.21)	0.006	(1.05)	Y
Vol, Dis	-0.067	(-0.55)	0.163*	(1.33)	0.002	(0.21)	-0.008	(-0.71)	0.008*	(1.44)	Y

Table 12. VAR models with two lags with financial control variables. The two variables (e.g., "Ret, Sent") in the first column for each row represent X_t and Y_t of a VAR model defined in equations (7 & 8), correspondingly. Coefficients Y_{t-1}, Y_{t-2} are the loadings in the equation (7) for X_t, and coefficients X_{t-1}, X_{t-2} are the loadings in the equation (8) for Y_t.

	Const	MoM_{t-1}	Ret_{t-1}	$Sent_{t-1}^{C1}$	$Sent_{t-1}^{C2}$	$Sent_{t-1}^{C3}$
$MoM = RV_t$	0.0010***	0.4831***	-0.0279***	0.0026	-0.0047	-0.0176
	(7.67)	(18.29)	(-8.45)	(0.65)	(-0.63)	(-1.58)
$MoM = Skew_t$	0.0000	-0.1445***	-0.0012***	0.0001	0.0007***	0.0003
	(0.92)	(-4.95)	(-10.19)	(0.70)	(2.77)	(0.94)
$MoM = Kurt_t$	0.0000*	0.2481***	-0.0002***	0.0001***	0.0000	-0.0000
	(1.82)	(8.88)	(-11.81)	(2.75)	(0.48)	(-0.16)
	Const	MoM_{t-1}	Ret_{t-1}	Dis_{t-1}^{C1}	Dis_{t-1}^{C2}	Dis_{t-1}^{C3}
$MoM = RV_t$	0.0010***	0.4483***	-0.0283***	0.0099***	-0.0111**	-0.0002
	(8.16)	(16.44)	(-9.49)	(2.71)	(-2.13)	(-0.02)
$MoM = Skew_t$	0.0000	-0.1515***	-0.0011***	0.0004***	-0.0004**	0.0000
	(0.92)	(-5.29)	(-11.52)	(3.21)	(-2.26)	(0.02)
$MoM = Kurt_t$	0.0000*	0.2233***	-0.0002***	0.0001***	-0.0001**	-0.0000
	(1.85)	(8.01)	(-12.18)	(2.59)	(-1.97)	(-0.23)
	Const	MoM_{t-1}	Ret_{t-1}	Nb_{t-1}^{C1}	Nb_{t-1}^{C2}	Nb_{t-1}^{C3}
$MoM = RV_t$	0.0012***	0.3626***	-0.0278***	0.0863***	0.4900***	0.4307**
	(9.46)	(12.31)	(-9.63)	(6.42)	(4.80)	(2.31)
$MoM = Skew_t$	0.0000	-0.1478***	-0.0011***	0.0009**	-0.0023	0.0123**
	(0.92)	(-5.16)	(-11.33)	(2.08)	(-0.70)	(2.03)
$MoM = Kurt_t$	0.0000*	0.2291***	-0.0002***	0.0001	-0.0005	0.0019*
	(1.84)	(8.21)	(-12.03)	(1.10)	(-0.82)	(1.70)

Table 13. OLS regressions of daily realised variance, skewness and kurtosis on principle components of sentiments. Independent variables are the lagged values of MoM, lagged bitcoin returns, lagged first three principal components of sentiment (C1 to C3).

Twitter												
Vars X_t, Y_t	Y_{t-1}		Y_{t-2}		X_{t-1}		X_{t-2}		$Sent_{t-1}^{trad}$		$Goog_{t-1}$	
Ret, Sent	0.255***	(3.06)	-0.103*	(-1.58)	0.198***	(11.77)	-0.002	(-0.11)	-0.002	(-0.77)	0.005***	(2.41)
Ret, Nb	-0.007	(-0.42)	0.012	(0.70)	-0.132***	(-2.59)	0.051	(1.00)	0.000	(0.01)	0.005***	(2.65)
Ret, Dis	0.029	(0.50)	0.009	(0.15)	-0.038**	(-2.33)	-0.000	(-0.00)	-0.000	(-0.02)	0.004**	(2.00)
Ret, Vol	-0.013	(-0.73)	0.017	(0.97)	0.025	(0.49)	0.061	(1.18)	-0.000	(-0.11)	0.005***	(2.36)
Vol, Nb	0.113***	(3.62)	-0.164***	(-5.24)	0.042*	(1.34)	-0.022	(-0.69)	0.006	(1.08)	-0.010***	(-2.83)
Vol, Sent	0.198**	(1.80)	-0.005	(-0.04)	0.008	(0.92)	-0.005	(-0.58)	0.004	(0.65)	-0.007**	(-1.85)
Vol, Dis	-0.035	(-0.34)	-0.220**	(-2.19)	0.037***	(3.72)	-0.025***	(-2.45)	0.007*	(1.31)	-0.004	(-1.00)
Reddit												
Vars X_t, Y_t	Y_{t-1}		Y_{t-2}		X_{t-1}		X_{t-2}		$Sent_{t-1}^{trad}$		$Goog_{t-1}$	
Ret, Sent	0.121*	(1.58)	-0.011	(-0.16)	0.135***	(9.84)	0.016	(1.12)	-0.001	(-0.35)	0.005***	(2.35)
Ret, Nb	-0.038**	(-1.73)	0.025	(1.12)	-0.107***	(-2.63)	0.025	(0.62)	0.000	(0.01)	0.005***	(2.47)
Ret, Dis	0.002	(0.03)	-0.015	(-0.31)	-0.044***	(-2.40)	-0.021	(-1.14)	-0.000	(-0.04)	0.005***	(2.44)
Ret, Vol	-0.013	(-0.73)	0.017	(0.97)	0.025	(0.49)	0.061	(1.18)	-0.000	(-0.11)	0.005***	(2.36)
Vol, Nb	0.084**	(2.23)	-0.163***	(-4.24)	0.059***	(2.39)	-0.060***	(-2.44)	0.005	(1.04)	-0.011***	(-2.88)
Vol, Sent	0.283***	(2.39)	-0.047	(-0.41)	0.004	(0.46)	-0.003	(-0.37)	0.004	(0.73)	-0.007**	(-1.98)
Vol, Dis	0.037	(0.43)	-0.171**	(-2.01)	0.031***	(2.86)	-0.039***	(-3.57)	0.007*	(1.38)	-0.005*	(-1.30)
StockTwits												
Vars X_t, Y_t	Y_{t-1}		Y_{t-2}		X_{t-1}		X_{t-2}		$Sent_{t-1}^{trad}$		$Goog_{t-1}$	
Ret, Sent	0.200***	(2.95)	-0.080	(-1.25)	0.079***	(4.75)	0.007	(0.42)	-0.001	(-0.22)	0.005***	(2.65)
Ret, Nb	-0.037**	(-2.18)	0.036**	(2.17)	-0.098**	(-1.81)	0.081*	(1.49)	0.000	(0.03)	0.006***	(2.67)
Ret, Dis	0.011	(0.20)	-0.041	(-0.73)	0.008	(0.46)	0.008	(0.47)	-0.000	(-0.09)	0.005***	(2.52)
Ret, Vol	-0.013	(-0.73)	0.017	(0.97)	0.025	(0.49)	0.061	(1.18)	-0.000	(-0.11)	0.005***	(2.36)
Vol, Nb	0.152***	(4.56)	-0.241***	(-7.50)	-0.150***	(-4.06)	0.169***	(4.56)	0.006	(1.11)	-0.004	(-1.12)
Vol, Sent	0.067	(0.65)	0.011	(0.11)	0.005	(0.52)	-0.007	(-0.84)	0.007	(1.25)	-0.007**	(-1.85)
Vol, Dis	-0.168**	(-1.78)	0.157**	(1.66)	0.013*	(1.35)	-0.024***	(-2.57)	0.007*	(1.43)	-0.006**	(-1.80)

Table 14. VAR models with two lags controlling Google trend. The two variables (e.g., "Ret, Sent") in the first column for each row represent X_t and Y_t of a VAR model defined in equations (7 & 8), correspondingly. Coefficients Y_{t-1}, Y_{t-2} are the loadings in the equation (7) for X_t, and coefficients X_{t-1}, X_{t-2} are the loadings in the equation (8) for Y_t. $Goog_{t-1}$ is the lagged Google trend measure.

List of Figures

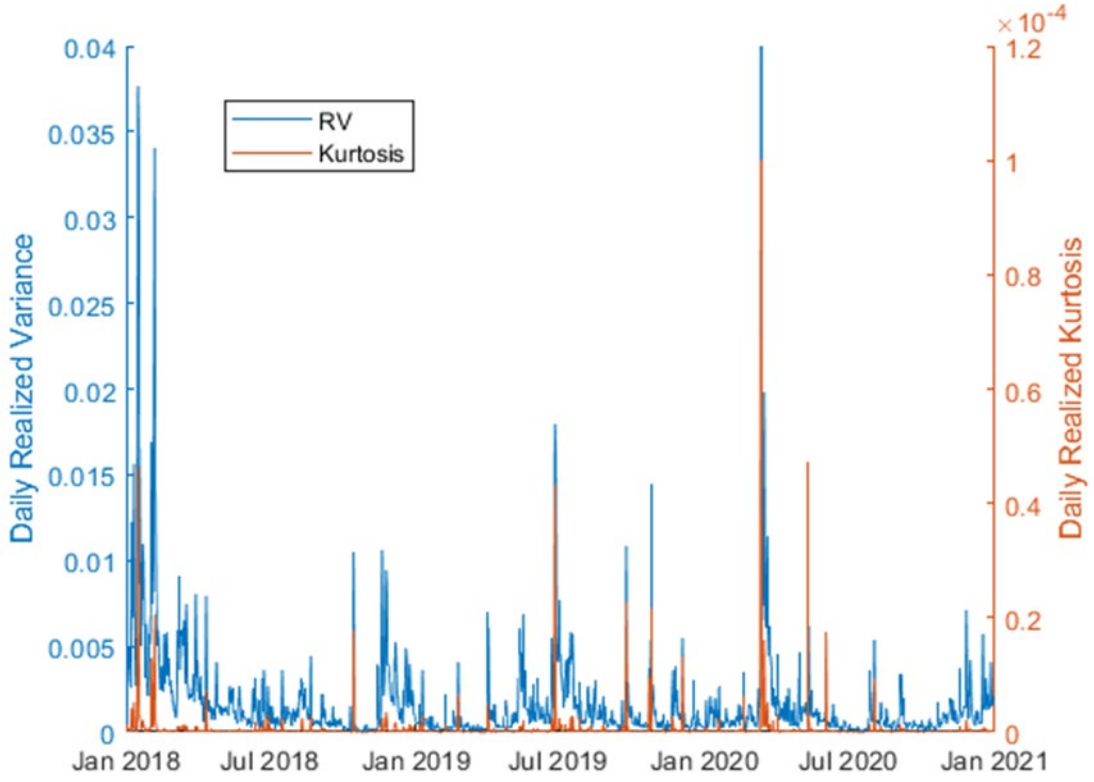

Figure 1. Daily realized variance and kurtosis, which are estimated from intraday 5-minute bitcoin returns.

Chapter 2
Targeted Financial-Oriented Social Media Sentiment Measurement: Natural Language Processing Approach

Abstract[8]

This study develops a natural language processing model that measures financial-oriented sentiment targeted toward specific firms in social media texts. First, we create a human-annotated social media targeted financial sentiment dataset. Then, we propose a prompt-based model architecture that achieves state-of-the-art performance on multiple benchmark datasets for general targeted sentiment analysis. Subsequently, we finetune this model using our annotated dataset, which allows it to measure targeted financial sentiment with high accuracy. We apply it to 23 million financial-oriented social media posts from different platforms to measure financial sentiment toward 24 meme stocks (stocks that gain frenetic attention from retail investors on social media which is often accompanied by dramatic price movement) and 30 Dow Jones constituent stocks. Our results show that the sentiment measured by our model is positively correlated with price return and negatively correlated with price volatility, and that this correlation is stronger for meme stocks than for Dow Jones stocks. We further demonstrate that our model's sentiment measurement economically outperforms other representative financial sentiment measurements by comparing the returns of the same trading strategy built upon them separately.

[8] This essay is coauthored with Gilles Caporossi (gilles.caporossi@hec.ca), Feng Zhan (feng.zhan@uwo.ca), and Xiaozhou Zhou (zhou.xiaozhou@uqam.ca). It has been accepted for presentation at the 2023 China Finance Review International & China International Risk Forum Joint Conference.

2.1 Introduction

A growing number of investors are turning to social media as a source of information with the advent of mobile technology and online communities. Furthermore, a significant number of business and political influencers, as well as traditional media agencies, have been utilizing social media such as Twitter, as a primary means of communicating with their audiences. As such, their posts contain specific information and could cause a significant reaction in the market. In addition, social media popularity has combined with the increased activity of retail investors in recent years, contributing to their growing influence in the marketplace, as exemplified by the "meme stocks" frenzy. The term 'meme' refers to a stock that is receiving intensive attention on social media such as Twitter, Reddit and StockTwits forums, and concurrently experiencing dramatic price movement. The growing influence of social media leads to increased research interests in its role in the financial market. Pedersen (2022) introduces a model to explain the mechanism of information propagation through social network and the consequent affects to the market. Hirshleifer (2020) highlights the transmission bias of economic and financial signals induced by social media. Financial sentiment, long recognized as an important market impacting factor, has been one of the focuses in studying social media. Sentiment from StockTwits is shown to forecast short-term stock index return (Renault, 2017), and is used to study the source of disagreement among investors which is the foundation of trading (Cookson & Niessner, 2019), and echo chamber effect which leads to investors' confirmation bias (Cookson, Engelberg, et al., 2022). More broadly, Cookson, Lu, et al. (2022) studies the social media sentiment and attention from Twitter, StockTwits, and Seeking Alpha, finding that sentiment-induced retail imbalances predict positive returns while attention-induced ones have the opposite market outcomes. Although the importance of financial sentiment in social media is widely recognized, its accurate measurement, however, is very challenging and less studied. In this paper, we propose targeted sentiment analysis (TSA) model using advanced natural language processing (NLP) methods, to address this problem.

Targeted sentiment, defined as the sentiment targeted towards a specific entity or aspect within a text, is a more refined measurement compared to the common overall sentiment

of a text. Often, a piece of text can contain multiple entities or aspects, which may bear completely different sentiments. For example, considering the sentence:

"Morgan Stanley says Disney could surpass Netflix in the streaming market."

The sentiment expressed would be totally different depending on which firm is the target of interest: neutral for *Morgan Stanley*, positive for *Disney*, and negative for *Netflix*. A common sentiment model can only predict one sentiment for the sentence as a whole regardless of which firm is being focused on, hence has no means to distinguish the individual sentiments towards the different firms separately. By contrast, only a TSA model designed to measure the fine-grained sentiment specific to a given target, could correctly predict the different sentiments given the different firms.

Targeted sentiment is especially important when studying financial oriented social media. For many traditional financial texts such as the earnings report or the financial analyst report, usually each text explicitly concerns only one particular firm, and this targeted firm is clearly documented in the database. Moreover, a such text is often in the form of a long document, and we know for sure the majority of its content is about the known target. Thus, for those type of texts, there's little problem attributing the overall sentiment obtained by pooling the measurements of all the parts of the document to the target firm. However, for the social media texts which are usually short posts by individual users, we don't have prior knowledge of which post is about what. In practice, to study a certain firm, we have to collect all financial oriented posts mentioning this firm. But unlike a long document with clear target, we have no guarantee weather the overall sentiment of a such post is really about this firm. That's why a TSA model is crucial for accurately measuring the sentiment conveyed by social media posts.

Measuring financial-oriented sentiment towards specific firms within social media texts is a challenging problem due to several reasons.

First, there's a scarcity of labeled data and advanced models for **financial-oriented sentiment** analysis. Although the measuring of textual sentiment has been extensively studied in the field of NLP, most of the data and research focus on general sentiment,

typically customers' likes or dislikes expressed in online reviews. However, the financial sentiment, defined as view of a favorable or unfavorable prospect from an investor's perspective, is very different and is much less investigated. There's a lack of large-scale labeled financial sentiment dataset to train and evaluate models, because of the difficulty of the labeling which requires expert knowledge. Unlike the general sentiment for which there are numerous public datasets of online reviews of size up to hundreds of millions, such as the Stanford Sentiment Treebank of movie reviews by Socher et al. (2013) and the Amazon Review Data by Ni et al. (2019), to our knowledge, only two datasets are publicly available for financial sentiment. One is the Financial Phrase Bank (FPB) dataset (Malo et al., 2014) that provides ~5000 financial news sentences each labeled with financial sentiment by multiple annotators. The other is the Financial Opinion Mining and Question Answering (FiQA) dataset (Maia et al., 2018) that contains 436 news and 675 social media posts from financial web pages, each labeled with fine-grained aspect-targeted financial sentiment. In terms of modeling, researchers in finance used to rely on word-counting based on tailored dictionaries of financial sentiment keywords, the most popular of which is proposed by Loughran & McDonald (2011). Besides dictionaries, classical statistical models such as naïve Bayes were also widely adopted (Antweiler & Frank, 2004; Das & Chen, 2007; Huang et al., 2014). Those models, treating a text as a simple bag of words with their order and context disregarded, are incapable of capturing complex semantics, thus leading to inaccurate measurements of the true sentiment in many cases. More recently, large pretrained language model (PLM) based on transformers, a novel NLP deep-learning architecture, has become dominant in the NLP field and achieved significant performance breakthroughs across various NLP tasks. Riding on this trend, some researchers in the domain of finance also begin to explore adapting the transformers models to financial sentiment analysis, and reported improved performance compared to traditional models (Araci, 2019; Jiang et al., 2022; A. H. Huang et al., 2022).

Second, measuring **targeted sentiment** requires more advanced models and is less studied compared to commonly measuring sentiment at the whole sentence or document level. Classical lexicon and machine learning based models, relying heavily on hand crafted rules and feature engineering, can perform well on sentence or document level

sentiment tasks, but have difficulty measuring the fine-grained targeted sentiment. Only with the recent development of deep neural network, can the complex dependency and interaction between the target and its context be effectively modeled. Apart from more complex model, labeled targeted sentiment data is also more burdensome to get. As we mentioned before, sentence or document level sentiment data can be easily obtained by mass amount from the online reviews. But fine-grained targeted sentiment datasets are much scarcer, only few public English datasets are available. The most used are the SemEval2014 laptops and restaurants review datasets (Pontiki et al., 2014) which contain online reviews on laptops and restaurants with each review manually labeled with sentiment towards different aspects, such as "service", "staff", "food", etc. for restaurants. Another one is the Twitter dataset (Dong et al., 2014) containing tweets manually labeled with sentiment towards different targets including celebrities, products, and companies. In the domain of finance, the only public targeted financial sentiment dataset available is the aforementioned FiQA dataset.

Third, the informal nature of social media texts makes them more challenging for NLP models (Farzindar & Inkpen, 2020). Non-standard or even incorrect grammatical structure and word spelling are very common in social media posts. Also, like many web content, social media are plagued with much more noise in terms of irrelevant content compared to formal media. Moreover, being short in length and often in a conversational nature, a social media post often provides very limited contextual information that is essential for language understanding. Considering all those issues, social media requires more advanced models that are more tolerant of informal language and noise, and have better ability comprehending texts in relation to their contexts.

In this paper, we address the above challenges by developing an NLP model that measures the financial sentiment targeted towards specific firms in social media texts. First, we create a targeted financial sentiment dataset of ~3000 social media posts, each annotated by multiple people with academic background in business to ensure the quality. This dataset adds to the rare public data resources regarding both financial sentiment and targeted sentiment. Then, we propose a novel NLP model architecture based on the prompt paradigm, which functions as reformulating the TSA task to imitate the natural

language inference (NLI) task on which the backbone transformers model was well pretrained with massive data. Our model proves to achieve state-of-the-art (SOTA) performance on multiple benchmark datasets for general TSA task. Subsequently, we finetune the model based on our targeted financial sentiment dataset, which enables it to measure targeted financial sentiment with high accuracy. Finally, we apply our finetuned model to over 23 million financial-oriented social media posts between 2020 and 2022 to measure financial sentiment towards 24 meme stocks and 30 Dow Jones constituent (DJ30) stocks. We show that the sentiment measured is positively correlated with price return and negatively correlated with price volatility. Moreover, we demonstrate that this correlation between social media sentiment and price is significantly stronger for meme stocks than for DJ30 stocks. We further construct a sentiment-based trading strategy using different financial sentiment measures. The return differences demonstrate that our model's sentiment measurement economically outperforms the other two representative financial sentiment measurements.

2.2 Background and Literature

2.2.1 Advanced Natural Language Processing

NLP is the subfield of artificial intelligence that aims at enabling computers to process and analyze human language, in order to perform relevant tasks such as machine translation, sentiment analysis, document summarization, question and answering, etc. Recent years have seen huge advancement of NLP due to several key factors including vast growth in computing power, increased availability of a large linguistics data, development of highly successful machine learning algorithms, and richer understanding of the language structure and its deployment in social contexts (Hirschberg & Manning, 2015). The rest of the chapter will cover some basic concepts and the development of modern NLP.

Representation Learning

For textual data to be processed by algorithms, first they need to be represented as numeric vectors. Classical NLP methods often treat a document as a **bag of words (BOW)**, i.e., a collection of independent words (or n-grams) that can be simply represented by a vector

of the counts of each word in the vocabulary. However, words represented in this way lose their semantic meanings, they become atomic units that have no inherent relationship to one another. For instance, the concept of synonym or antonym is completely absent. Furthermore, a BOW representation has the dimensionality equivalent to the size of the entire vocabulary, resulting in a large and usually sparse vector. This can lead to significant computational inefficiency.

A milestone for representation learning is achieved with the novel **word embeddings** methods: featurized word representations that preserve semantic information of words. Instead of merely being a numeric encoding without meaning, the new word representations can capture syntactic and semantic regularities that enable analogy reasoning. Mikolov, Yih, et al. (2013) found that using the word embedding vectors they generated, semantic relationships can be represented using simple arithmetic, e.g., "$King - Man + Woman = Queen$". The word embeddings are learnt from large unlabeled text corpuses, which are easy to obtain. The generated word embeddings can then be applied for downstream tasks, which can greatly boost their performances since more semantic information of words can be of great value to those tasks. The most influential word embedding algorithms include *word2vec* (Mikolov, Chen, et al., 2013) and *GloVe* (Pennington et al., 2014). Those advanced word embedding models have greatly boosted the performance of many NLP tasks, because of their ability to extract and preserve words' meaning by simply being pretrained on a large unlabeled corpus. This concept of gaining general knowledge from training on large unlabeled data, in order to later apply the knowledge learned to other downstream tasks, is the core idea of transfer learning, which we will introduce in the following.

Transfer Learning

A major assumption for statistical learning algorithms is that the training data and the future data to be applied on must be in the same feature space and of the same distribution. However, this can't always be satisfied for real-world applications: often we only have a small amount of labeled data for our task of interest (target task), but we may have enough data from another related task (source task) where the feature space or distribution is

different. In this case, if we can let the model "pre-train" on the source task data and then transfer the knowledge it leans to apply on the target task, it would improve the performance without expensive additional data labeling effort. This reasoning leads to the development of transfer learning.

Transfer learning was earlier popularized in the field of computer vision (CV), because during pretraining CV models can effectively gain automatic feature extraction capabilities such as detecting the edge of objects or identifying shapes. Those capabilities will benefit all sorts of downstream CV tasks. Similarly, since NLP tasks also share common knowledge about the language, transfer learning was naturally introduced into the NLP field, and has become a fundamental methodology today. According to a summary by Ruder et al. (2019), the most common process of transfer learning in NLP today consists of two phases: 1. a pretraining phase in which general representations are learned on a source task or domain; 2. an adaptation phases in which the learned knowledge is applied to a target task or domain.

As we mentioned, word embedding is a typical case of transfer learning. Word embedding algorithms can extract word vectors that preserve the semantic meanings of words from merely unlabeled corpus, which corresponds to the phase of pretraining. Then in the adaptation phase, those word vectors are applied to represent the input texts of target tasks, which greatly improves the performance compared to using representation that contain no prior semantic knowledge of the word such as BOW. Despite its huge success, word embedding is essentially a "shallow" form of transfer learning, since it can only learn and represent individual word-level knowledge. This leads to one critical flaw: the representation of a word is not context-specific, whereas a word can have very different meanings in different contexts. In order to capture and transfer deeper-level language knowledge such as contexts and interactions of words within an entire text, researchers developed the method of language model pretraining.

Pretrained Language Model (PLM)

Language model is a type of NLP model that learns to probabilistically predict the next word given any previous sequence of words. The goal is to be able to assign a probability

for any given sentence or paragraph, based on the probability distribution it learned from the training language corpus (Goldberg, 2017). The pretraining of a language model usually means training the model on large-scale unlabeled corpus, with the task of reconstructing the original text given a text that has been artificially corrupted with certain noise functions. Pretraining enables the model to gain useful insights on the language, thus learning the representation of texts that captures deeper-level language knowledge including contextualized semantics.

Given a capable PLM architecture and proper pretraining tasks design, the more diverse data the model is pretrained on, the more comprehensive language knowledge it can encode (Pérez-Mayos et al., 2021). Although labeled text data are rare, fortunately, unlabeled text data is abundant and cheap to obtain, thanks to rapid digitalization and information boom. Many public large-scale corpuses are collected from sources like Wikipedia, news, books, web crawl, etc. Nowadays, it is common for large PLMs to be pretrained on more than a hundred GB of raw text data, such as the RoBERTa model (Liu et al., 2019). PLMs empowered by deep neural network, especially large PLMs developed in recent years, have become the fundamental technologies of NLP (Li, 2022).

As a simple usage example, the PLM can generate new texts that mimic the style of those in the training corpus, e.g., a PLM trained on a corpus of Shakespeare's poems and plays can "write" Shakespeare style texts. Beyond this little funny application, language modeling is actually a critical NLP task that lays the foundation for many other higher-level tasks. As an example, when performing French-English translation, given the input sentence *"Tom est gros"*, the model may have to weigh between *"Tom is large"* and *"Tom is fat"* as for output. If the model has good language knowledge, then it should recognize "Tom" as most likely a person's name, thus the more plausible adjective that follows should be "fat" instead of "large". So, the model should evaluate that $Probability("Tom\ is\ fat") > Probability("Tom\ is\ large")$, which leads to the correct translation. From this example, we may have a glimpse of insight that evaluating the probability of a sentence implies the judgement on the semantic and contextual information, sometimes even world knowledge in that sentence. Recent large PLMs pretrained on immerse amount of textual data have demonstrated the ability to incorporate

such complex context semantics and knowledge, leading to a revolutionary development of modern NLP.

2.2.2 NLP Development

In its early days, NLP relied heavily on linguistics study, researchers often attempted to predefine dictionaries and rules to decipher human language for the computer. However, language has proved to be too complex for this approach: language can be ambiguous, fuzzy, context-dependent, and often requires reasoning based on common sense. Gradually, researchers turned to the machine learning approach of applying statistical models over a large amount of data so that algorithms can learn empirical language patterns and knowledge by themselves. Classical machine learning models such as naïve Bayes and support vector machines achieved notable successes in NLP. However, those models usually treat texts with BOW method, which limits the models' ability to capture the order and dependency of words which are crucial for language understanding.

Later, **recurrent neural network (RNN)** based models such as long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) and gated recurrent unit (GRU) (Cho et al., 2014), capable of handling texts as ordered sequences of words, dominated the NLP domain by greatly pushing the performance on many tasks. RNN is a special type of neural network whose units are connected recurrently along a sequence, allowing prior output based on previous values to affect subsequent input concerning the current value. This way RNN model can relate different parts of a sentence to understand the dependencies and contexts. For certain sequence to sequence tasks like machine translation, the input sequence and output sequence may have different lengths and the relationship between the two lengths is non-monotonic. This brings a problem for RNN which is good at mapping the input sequence to the output sequence only when the alignment between them is known a priori. To address this problem, the **encoder-decoder** architecture (Fig. 6) was proposed (Cho et al., 2014; Sutskever et al., 2014). The encoder is an RNN module that takes a variable-length sequence as input and output its encoding, i.e., a fixed-size vector that encapsulates all the information of input sequence. Then this encoding is passed to the decoder, another RNN module which is essentially a PLM, that predicts the output sequence with the highest possibility conditioning on the input.

However, there are bottlenecks with RNN encoder-decoder architecture. First, it needs to compress the information of a whole sentence into a fixed length vector, which leads to bad performance for long sentences due to loss of information. Second, the sequential computation of RNN precludes computing parallelization, which greatly limits the speed and scale of model training.

To address those issues, Bahdanau et al. (2014) first introduced a novel encoder-decoder model with the **attention mechanism**: when the model predicts a word at each time step, it searches the source sentence for the most relevant words, and use information of those words in addition to the previously predicted words to generate the current word prediction. In another word, like human reading, the model leans to pay attention only to those words in the source sentence that are relevant to the target word, instead of relying on encoding the whole source sentence. Also, without the recurrent modules, attention models are significantly more efficient with parallelization, which makes them way faster to train. The novel attention-based language models are called **transformers** model. Large pretrained transformers models have brought revolutionary improvements on almost all NLP tasks, greatly pushing the boundaries of NLP.

Two representatives and pioneers of transformers are the **GPT** (Generative Pretrained Transformer) (Radford et al. 2018) from OpenAI, and **BERT** (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019) from Google. The both adopts the pretrain-finetune paradigm to perform transfer learning. The initial version of GPT (GPT-1) has 117 million parameters, consisting of a 12-layer left-to-right transformer pretrained on a diverse corpus of unlabeled text, with a standard autoregressive PLM task, i.e., to predict the next word given the previous sequence of words. After the pretraining, the model is then fine-tuned on each specific down-stream task. For classification tasks, the finetuning involves simply concatenating a classification head, usually a shallow neural network, that takes the text embeddings generated by the PLM and make classification based on it. GPT with finetuning achieves SOTA performance on a wide range of NLP tasks. Its success demonstrates that large PLMs pretrained on massive raw text data are effective transfer learner when coupled with the fine-tuning approach. Due to the auto-regressive nature of GPT, it is especially suitable text generation related tasks.

The later BERT model pushed the SOTA even further. To overcome the unidirectional constraint of classical PLM, BERT invented a novel bidirectional pretraining objective called *"masked language model (MLM)"*: randomly mask some words within a text, and let the model predict those words based on the contextual texts from both sides. In addition to the MLM, BERT also uses a *"next sentence prediction (NSP)"* task that leans sentence-level representations. Thanks to those innovations and the huge size of model and data being used, BERT is proven to be impressively effective for transfer learning. With finetuning, BERT achieved amazing success by claiming the SOTA on almost all major NLP tasks, including both text and token classifications as well as text generation. It has been widely regarded as a new milestone for NLP, triggering a wave of transfer learning using pre-trained models. Also, there have been sizable research efforts related to its variants (e.g. by diminishing model size, like Albert (Lan et al. 2019); or improving model pretraining, like RoBERTa (Liu et al., 2019)), derivatives (e.g. extending BERT for specific tasks, such as BioBERT (Lee et al., 2019)), and interpretation (e.g. investigating its internal mechanism (Tenney et al. 2019)).

Training a large transformers model on massive data takes a prohibitively huge amount of computing resources. So, instead of training a transformers model from scratch, people usually use directly the public models already pretrained on huge raw texts data (and sometimes further trained on specific large labeled datasets of source tasks), as the backbone of their own model for downstream tasks.

2.3 Prompt-Based NLP Model for Targeted Sentiment Analysis

2.3.1 Related Works

Prompting Method

Using prompts to guide a PLM to perform different tasks is becoming a novel paradigm for effectively leveraging large PLMs. A prompt is a piece of text that we add to the PLM input so that the original task can be reformulated as a task that the PLM has already been pretrained on. As the example in Figure 1 shows, with the help of prompting, a sentiment analysis task can be restructured in a way similar to the MLM task, so that a model

pretrained with MLM task can perform sentiment analysis task directly without further task-specific training.

Prompting method was first popularized with the GPT-2 model (Radford et al., 2019) for zero-shot prediction, i.e. making prediction on a downstream task without training for this specific task. The enormous 1.5 billion parameters GPT-2 model was pretrained with 40GB of raw texts. Then the model parameters are frozen, and different prompts could be used to direct the model to perform different tasks including translation, reading comprehension, etc., without further tunning the model for those tasks. Yin et al. (2019) propose a prompting approach to reformulate text classification task as NLI task. They show that a BERT model further trained on NLI task can perform zero-shot classification. Beyond the zero-shot setting, Schick & Schütze (2021) introduce PET model that proposed further finetuning the model with prompt. Instead of freezing the PLM model, they further finetune the PLM's parameters with supervised training approach. While unlike traditional supervised learning using only input texts and labels, they add prompts specifically engineered for different tasks to guide the PLM to better leverage the patterns it leant during pretraining that are relevant to different tasks. Later, Gao et al. (2021) show that prompt-based finetuning of PLM on a small amount of labeled data can dramatically outperform standard finetuning. A formal definition of prompting method and a systematic survey are presented by Liu et al. (2021).

[Insert Figure 1 here]

Targeted Sentiment Analysis Models

The development of deep neural network enables researchers to build modern TSA models that are increasingly better at detecting this fine-grained sentiment. Tang et al. (2016) propose MemNet which adopts attention mechanism with external memory. It uses attention mechanism to explicitly model the target's relatedness to different parts of the texts semantically embedded in the memory. Wang et al. (2016) propose ATAE-LSTM which combines attention mechanism with LSTM. It concatenates target embedding with the representation of each word to let the aspect embedding play a role in computing attention weight. Chen et al. (2017) propose RAM model which uses bidirectional LSTM

to build memory from embeddings. The importance of different words in a sentence is weighted by their distance to the target word(s), closer words get higher weights. It also uses recurrent attention to focus on target-related information from memory.

More recently, transformers-based PLM have brought its success to TSA. Dai et al. (2021) show that fine-tuning a PLM on TSA task forces the PLM to implicitly learn more sentiment-word-oriented dependency trees compared to classical parser-provided dependency tree. Combining the induced tree with popular TSA models proves to elevate the performance to SOTA level. Tian et al. (2021) propose BERT-based TSA enhanced with word dependencies captured by an external key-value memory network (BERT-KVMN). They firstly extract the words associated to the target by parsing the dependency information of the sentence, then use KVMN to encode and weight such information to enhance TSA accordingly.

The novel prompting method is also being applied to TSA task. Seoh et al., 2021 build two different prompt-based models, one formulates the TSA task as a language modeling task, and use pretrained BERT or GPT-2 model as backbone model; the other formulates the TSA task as a NLI task, and used BERT further trained on NLI data as backbone model. Their approach proves to outperform standard supervised finetuned models for TSA.

2.3.2 Prompt-Based Targeted Sentiment Analysis Model

In this section we describe our method of prompt-based TSA using pretrained transformers model. We use BART-MNLI (Lewis et al., 2020), a powerful transformers model trained a large NLI task dataset, as the backbone of our model. In order to effectively leverage the language understanding capability of the backbone model, we design a prompt-based approach to reformulate the TSA task to imitate the NLI task. We further finetune our prompt-based model with labeled TSA data, to update the model's weights to be adapted to specific TSA tasks. Our model is different from the one in Seoh et al. (2021 in several important aspects. First, the model architectures for leveraging the inference prediction are different. We modify the BART-MNLI classification head to generate binary prediction during finetuning instead of keeping the three-class NLI

classification head. This enables our model to be more adapted to the actual TSA task with finetuning. Second, correspond to the model architecture, the prompting designs are different. We construct prompts for all three sentiment categories explicitly, instead of deducing the predictions from the inference results. Third, the backbone model we choose has better generalization ability and is more suitable for NLI task.

Backbone Model

Natural Language Inference: NLI is the problem of determining whether a text ("hypothesis") can logically be inferred from another text ("premise"). We call this inference relationship "entailment" if it's true, "contradiction" if it's false, or "neutral" if it's undetermined. For example, given the premise *"A child is playing football on the muddy playground in the rain."*, the relationship is entailment if the hypothesis is *"A person is playing sport outside amid bad weather."*, neutral if the hypothesis is *"A man loves football more than reading."*, and contradiction if the hypothesis is *"A man is afraid of getting wet in the rain."*. NLI task is a perfect testing ground for an NLP model's ability to capture the linguistic meanings of sentences. Hence a model trained to excel in NLI would be capable of extracting rich semantic representations of texts, which is a crucial basis for performing other downstream tasks.

There are two notable large public datasets for NLI, which have promoted a great amount of progress in NLP. The earliest one is the Stanford NLI (SNLI) corpus (Bowman et al., 2015), that contains 570K human-written hypothesis-premise pairs. To construct a set of pairs, a human annotator is given a true description of an image as the premise, and asked to come up with the three types of hypotheses: an alternate true description as entailment, a description that might be true as neutral, and a false description as contradiction. The way SNLI dataset is constructed constraint its text genre to be image captions which are descriptions of concrete visual scenes, thus lacking many important concepts such as time, mental states, etc. Modeled on the SNLI corpus, the later Multi-Genre NLI (MNLI) corpus (Williams et al., 2018) overcomes the earlier drawbacks by covering a wider range of genres of texts with different styles, formality, and topics. Its 433K human-annotated texts-pairs include both written texts like press releases, letters, fictions, travel guides,

etc., and spoken texts like face-to-face conversation, telephone transcripts, etc. The wide coverage of MNLI makes it a valuable source for training advanced NLP models that are good at domain adaptation and transfer-learning when solving different tasks in various domains.

BART-MNLI: BART is a transformers model developed by the Facebook AI team (Lewis et al., 2020). It adopts a standard sequence to sequence architecture, while creatively combines the bidirectional encoder of BERT and the autoregressive decoder of GPT. It is pretrained with a so-called text denoising task, which involves two steps: 1) corrupting the text with a noise function by masking arbitrary spans of words and randomly permuting sentences, and 2) letting the model learn to reconstruct the original text. Thanks to its special architecture and the pretraining task, which is proven to be very effective, BART achieves great performance on various common benchmarks for both text comprehension and text generation.

We adopt the BART-MNLI, i.e., BART model further trained on the MNLI dataset, as the backbone of our model, because of its proven capability of language understanding and domain generalization, as well as its easy public access[9]. During the training, BART model takes each input from MNLI in the form of a premise-hypothesis pair, adds a special token to separate the two, and appends another special token to mark the end of the sentence. The representation of the end of sentence token, EOS, is plugged into a classification head to make prediction.

Prompt Method and Model Design

For the model to effectively harness the inference capability of the backbone PLM, we follow two basic concepts when designing the prompt. First, the prompt must be constructed in a way that mimics the NLI in both form and logic. Second, the prompt needs to aim the model to perform inference on the specific aspect that we want to capture,

[9] We use the pretrained BART(large size)-MNLI model checkpoint available freely via the HuggingFace platform: https://huggingface.co/facebook/bart-large-mnli. Due to the sizes of the model and dataset, pretraining BART model on MNLI will take an immense computational resource, which is both impractical and unnecessary to do by our own.

i.e., sentiment in our case. Those lead to a basic cloze-style prompt design similar to the ones first proposed by Schick & Schütze (2021). When we construct a prompt in this way, we are actually imitating how human read a text and respond to the question of judging the sentiment. Furthermore, in order to direct the model to focus attention on the targeted entity, we also need to explicitly embed and indicate the target in the prompt.

The resulted prompt method is as follows: For the supervised training phase, given an input text with a specified target and a corresponding sentiment label, we construct three different prompts embedded with the target and separately with the three sentiments labels (see the illustrative example in Figure 2). We then assign a binary label to the inference relationship between the text and prompt pairs. Only the one embedded with the original sentiment label is true among the three prompts. Similar to the input of NLI data, we input the text-prompt pairs joined by special tokens to mark the boundaries, and take the embedding of the EOS token as the representation of the sentence pair. So, one row from the original training data will generate up to three labeled text-prompt pairs for the training. In the prediction phase, we want the model to predict the sentiment given an input text with a target. For that, we still construct the three different text-prompt pairs same as in the training, and let the model predict their individual probabilities of being true, then we take the sentiment label in the prompt with the highest probability as the sentiment prediction of the target.

[Insert Figure 2 here]

The model architecture is illustrated in Figure 3. The original BART-MNLI model consists of a two-layer three-class classification head appended to BART PLM. Except for changing the last layer of the classifier with a layer of binary output, we keep the all the rest pretrained parameters of the BART-MNLI model, for the sake of preserving as much the pretrained knowledge as possible.

[Insert Figure 3 here]

The above model input design can fit any TSA dataset that has a clear sentiment label for each observation, like the public SemEval2014 (Pontiki et al., 2014) and Twitter (Dong

et al., 2014) datasets. But in practice, when annotating a dataset with multiple people for each observation, different annotators will often have disagreement on the label, so that the original labeling is not clear-cut. Aggregating those disagreed labels into one hard label causes information loss. Now that we have the full original data of the labels made by multiple annotators on each text, we design a slightly different input to best leverage the extra information on our data. Instead of hard labeling an observation as clearly being of one sentiment label, we attribute to each label a probability score based on the percentage of annotators that vote for this label, as shown in the example in Figure 4. The soft label is then used to calculate the loss. This soft-label input method proves to generate better performance on our data.

[Insert Figure 4 here]

2.3.3 Experiment and Results

To test models' performance on general TSA, we use the three most widely adopted public benchmark datasets: the SemEval2014 laptops and restaurants review datasets (Pontiki et al., 2014), and the Twitter TSA dataset (Dong et al., 2014).

To test models' performance on financial TSA, we use the multiple-platform social media financial TSA data that we gathered, as well as the public FiQA dataset (Maia et al., 2018).[10]

Results on General Targeted Sentiment Analysis

We first train and test our model for general TSA. Table 1 shows the performance comparison between our prompt-based TSA model (Prompt-TSA) and other representative SOTA TSA models. Our model outperforms all the existing models on the three benchmarks. The results prove the strong ability of our model in performing the TSA task.

[10] Please see appendix for details on the model training implementation and hyperparameters.

[Insert Table 1 here]

Results on Financial Targeted Sentiment Analysis

We then train and test our model on two financial TSA datasets, including our own social media dataset, and the FiQA dataset. The FiQA dataset contains 436 news headlines and 675 social media posts from financial web pages, labeled with sentiment score targeted towards stocks (Maia et al., 2018). Our won social media dataset originally contains around 4K finance/investment-related social media posts from social media platforms, including StockTwits, Twitter, and Reddit. Each post is annotated by at least 5 people with adequate education background in business, as positive, negative, or neutral. Due to the ambiguous and noisy nature of the social media, the annotators often disagree on the labeling of the same sentence. To ensure the correctness of the labeling in our final sample, we only keep the posts for which over 80% of annotators agreeing on the same label. The size of our final sample is 1 355, which is comparable to that of the FiQA dataset.

Table 2 shows the performance of our Prompt-TSA model. We also tested two representative non-targeted financial sentiment models on those datasets. The first uses the Loughran & McDonald financial dictionary (LM Dictionary) (Loughran & McDonald, 2011); and the second uses FinRoBERTa (Jiang et al., 2022), financial PLM based on RoBERTa model pretrained on raw financial texts and finetuned on FPB dataset. Our model demonstrates superior performance with over 80% accuracy on both datasets, significantly outperforming the other two models.

Also, when comparing the cross test results, we notice that the model trained on our financial social media data shows better performance when applied to the FiQA dataset (acc = 68.80%) compared to the reverse (acc = 61.62%). This may indicate that our data has a higher training value that helps the model generalization.

[Insert Table 2 here]

2.4 Financial Implications of NLP-based Sentiment Analysis

2.4.1 Textual and Financial Data

We evaluate the proposed sentiment measure and its economic value empirically utilizing four different datasets for the time period between July 2020 and July 2022: 1) social media textual data from Twitter, Reddit and StockTwits for 24 meme stocks and 30 Dow Jones Industrial Average component stocks[11]; 2) traditional media articles (such as the Wall Street Journal) for the aforementioned meme and Dow Jones stocks; 3) intraday prices and volumes; and 4) daily Fama-French factors.

Using the application programming interfaces (APIs), we collect social media textual postings concerning the targeted companies from Twitter, Reddit, and StockTwits. For StockTwits and Twitter, by convention, a cashtag followed by a ticker is used as the keyword for identifying stock-related posts (e.g. $AAPL for investing-related discussion on Apple Inc.). For Reddit, however, the cashtag convention is not valid. Therefore, we first filter finance related posts by first restricting our download within selected investing-related subreddits[12], then we further filter the downloaded posts using NLP techniques to ensure the correct identification of the company names and tickers keywords[13]. In addition to the posted messages, all scraped social media postings include the date and timestamp information. Due to the unique design of Twitter and StockTwits, social postings from both platforms include the number of followers of the posters. Additionally, Twitter data includes the number of retweets for each post.

[11] See Appendix for the list of stocks we use in this paper. We select a meme stock if the stock appears in the monthly top 10 holdings of Roundhill MEME ETF at least twice during Dec 2021 (inception of the MEME ETF) to July 2022.

[12] The Reddit forum is composed of subreddits each devoted to a specific topic. We select 10 most influential finance/investing related subreddits based on the number of members and time of existence. Selected subreddits include: 'stocks', 'options', 'wallstreetbets', 'CanadianInvestor', 'SecurityAnalysis', 'InvestmentClub', 'RobinHood', 'investing', 'StockMarket', 'ValueInvesting'.

[13] When using company names and tickers as keyword to search and download posts on Reddit, the downloaded data could contain many irrelevant posts, because the simple keyword matching on Reddit API is case insensitive and superficial. For example, searching the meme stock tickers WISH will return many irrelevant posts containing the plain word “wish”. As another example, searching the company name Apple with return irrelevant posts mentioning the fruit “Apple”. To address this problem, first, for ticker based downloads, we filter them by requiring strict case-sensitive match of the ticker. Second, for name based downloads, we use named entity recognition to ensure that the keyword refers to a company instead of a generic meaning.

For our sample period, the meme stocks related (DJ30 stocks related) textual dataset includes 5.20 (2.34) millions Twitter posts, 11.68 (1.88) millions StockTwits posts, and 1.86 (0.44) millions Reddit posts, a daily average of 297 (107) tweets, 661 (86) stockstwits, and 106 (20) reddit posts for each meme stock (DJ30 stock), respectively. Using this dataset, we further measure the sentiment of every post using our proposed NLP learning model documented in Section 3. Generally, the sentiment of a post is expressed as a continuous numeric value between -1 (negative) and 1 (positive). We compute the daily sentiment based on the average sentiment[14] and disagreement based on the standard deviation of sentiment of all messages posted during a given period.

For the purpose of testing whether social media sentiment has additional effects on stock return and volatility beyond traditional media sentiment, we also used our proposed methodology to compute the financial sentiment embedded in the Wall Street Journal (WSJ) and use it as a control variable. The articles relating to the companies are gathered from the Factiva database. For all meme stocks (DJ30 stocks) during our sample period, we obtain 591 (2 366) articles in total, which equals to one (five) article(s) per day on average that can be used to compute the sentiment of traditional media.

In this study, we use the Trade and Quote (TaQ) dataset of the Wharton Research Data Services (WRDS) to determine the intraday price and volume. With intraday data, we calculate the daily return as the log difference between the close price of day t and that of $t-1$, and the realized volatility as the sum of squared 5-minute log return. The data of daily Fama-French 5 factors are from Kenneth French's web site[15].

[Insert Table 3 here]

From Table 3, we observe several interesting pieces of information. First, there seems to be no significant differences on both return and trading volumes between meme and DJ30 stocks. Overall, meme and DJ30 stocks combined has a return of 0.178 percent (with a standard deviation (SD) of 0.06) on daily basis. Meme stocks' daily return is positive at

[14] We also compute the followers weighted average sentiment for Twitter and StockTwits, and Retweets weighted average for Twitter.
[15] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

0.4 percent (SD = 0.08) and DJ30 stocks' daily return is negative at -0.009 percent (SD = 0.02). The average daily trading volume for both types of stocks combined is 15.94 (millions shares, with SD = 1.10), with meme stocks traded a bit higher at 16.05 (SD = 1.28) and DJ30 stocks at 15.88 (SD = 0.95). However, the differences between both daily return and volume are not statistically significant. Second, compared to both daily return and daily trading volume, there are great variations in terms of the number of messages/posts on meme and DJ30 stocks. On average, there are 217 twitter posts (SD = 1472) each day mentioning either meme or DJ30 stocks. For meme Stocks, there are 493 (SD = 2245) posts, while there are only 27 (SD = 161) posts covering DJ30 stocks. Similarly, there are 89 (SD = 542) Reddit posts covering both types of stocks, with 207 (SD = 822) posts for meme stock and 4 (SD = 64) posts for DJ30 stocks. Considering 'meme' stock refers to a stock with dramatic price movement that is mainly caused by sentiment on social media posts, the differences on the number of posts regarding meme stock and DJ30 stock are not surprising.

2.4.2 Performance of the Proposed Sentiment Measure

In order to assess the performance of sentiment measures, we examine whether they can distinguish meme stocks from DJ30 stocks in market impact tests. In theory, meme stocks' return and volatility should be more sensitive to social media sentiment than DJ30 stocks' return and volatility. Table 4 and Table 5 illustrate the multivariate regression results on the impact of the proposed sentiment measurements on both stock return and stock volatility. In Table 4, the dependent variable is stock return. For each sentiment measure, we perform one regression model with control of stock trading volume, sentiment measure from traditional media WSJ, and five Fama French Factors (FF1-FF5). Models 1 to 3 use three different sentiment indexes measured from Twitter, models 4 and 5 use two different sentiment indexes measured from Stocktwits. Model 6 uses one sentiment index measured from Reddit. And model 7 adopts one sentiment index aggregating different sources using principal component analysis.

[Insert Table 4 here]

[Insert Table 5 here]

The results in Table 4 indicate that, in general, social media postings have a positive impact on stock returns. All social media sentiment indexes are positive and statistically significant at 1 percent. In term of economic value, one standard deviation of changes in social media sentiment generates 0.0038 in stock return. For meme stock, this impact is more significant. The coefficients of all social media sentiment indexes are higher for meme stocks. Economically, one standard deviation changes in social media sentiment generates 0.0286 (0.0038+0.0248) in meme stock return. Our results are consistent with our theoretical predictions that sentiment has a significant impact on stock price movement and this impact is stronger for meme stock. Besides the significant impact of social media sentiment, Table 4 also reveals some other important factors. Our results show that trading volumes are positively and statistically significantly associated with stock return. Nevertheless, sentiment generated from traditional media source WSJ is negative and statistically significant associated with stock returns.

Table 5 repeats all models in Table 4 with stock volatility as the dependent variable. Unlike the consistent and significant positive relationship found between sentiment and stock returns, the relationships between social sentiment and stock volatility are mixed, depending on social media platform. In general, results in Table 5 show a negative and statistically significant relationship between social media sentiment and stock volatility. Nevertheless, when the sentiment is measured using PCA data, our results show a positive and statistically significant relationship between social media sentiment and stock volatility. For meme stock, except for one sentiment measure from twitter, all social media sentiment shows a significantly negative relationship with meme stock volatility.

2.4.3 Trading Strategy Based on Social Media Sentiment

We further test if our new Prompt-TSA sentiment measure economically outperforms the measures by the other two existing representative financial sentiment models: FinRoBERTa and LM dictionary. To do so, we compare the average daily return derived from these three measures for both meme stocks and DJ30 stocks. More specifically, we conduct a daily basis trading strategy: at day t, we first compute the daily social media sentiment for Twitter, StockTwits, and Reddit. We use each social media sentiment together with other control variables to predict the next day return, and then buy (sell) if

the predicted return is greater (less) than zero [16]. We close our position at the next-day's market closing and repeat the same procedure for day $t + 1$.

We compare the average daily return of meme stocks for strategies based on sentiments measured by Prompt-TSA, FinRoBERTa, and LM dictionary. The results presented in Table 6 show that the daily average return derived from our new proposed model is statistically significantly higher than those based on FinRoBERTa and LM dictionary measures for meme stocks. When applying the same strategy to DJ30 stocks, we find that none of the 3 sentiment measurements generate positive return on average, and the results do not indicate significant difference in performance. This suggests social media sentiment as an investment signal is more applicable to meme stocks than to blue-chip DJ30 stocks.

[Insert Table 6 here]

[Insert Table 7 here]

2.5 Conclusion

In this study, we develop a cutting-edge NLP model for financial-oriented social media that can measure financial sentiment targeted toward specific firms within a text. The model architecture itself is domain-agnostic and demonstrates state-of-the-art performance on multiple benchmark datasets for targeted sentiment analysis in general domains including online reviews and Twitter sentiment. Further, based on the high-quality human-annotated social media targeted financial sentiment dataset that we created, we are able to finetune our model so that it measures targeted financial sentiment with high accuracy, outperforming two other representative existing financial sentiment models (that are not as sophisticated and are not specifically designed for TSA) by a large margin. Then, we test the financial implication of our new sentiment measure using 25 million social media posts from Reddit, Twitter and StockTwits. Those posts are filtered

[16] Given that social media postings can arrive any time during the day and market operation time is between 9:30 and 16:00, we choose 16:00 as our cut-off time. Therefore, the daily return and sentiment for day t are the return and sentiment between 16:00h on day $t - 1$ and 16:00h on day t.

to be finance / investing relevant, and to concern the 24 meme stocks or Dow30 stocks. In general, the sentiment measured by our model shows predictive power for short-term future return and volatility of the targeted companies' stock prices. Higher social media sentiment forecasts higher return and lower volatility. Moreover, consistent with our hypothesis, this relationship is stronger for meme stocks than for Dow30 stocks. To further compare our model with the other two existing financial sentiment models, we construct a trading strategy using the different sentiment measures in parallel. The strategy based on our sentiment measure has a higher return compared to the others, indicating that our model outperforms the other two existing models economically.

There are several aspects that can be explored for future research. The first two aspects are about extending the model, provided that we can get proper different labeled targeted sentiment data. First, since our model architecture is agnostic to text genre or domain, we can naturally extend its application to other financial texts beyond social media such as business news; or even extend to other domain beyond finance, for example, measuring consumer sentiment towards companies for marketing research. Second, besides targeting companies, by adjusting the prompt, the model could also be modified to target different types of entities or even aspects/topics. The last aspect is about enhancing the model architecture. We could use more sophisticated technics such as automated prompting search (Shin et al., 2020) instead of empirical prompt construction.

Appendix

Implementation Details

We used Huggingface and PyTorch for our model construction and training.

Key hyperparameters for training our model include:

- Use AdamW optimizer.
- Learning rate =1e-5, with a warmup ratio = 0.1, and linear decay scheduler.
- Effective batch size = 64.
- Train for 3 epochs.

Computation platform: Compute Canada Narval HPC[17] server equipped with Nvidia A-100 GPUs.

All model performance results reported on the benchmark TSA datasets and on financial TSA datasets are based on 5-fold cross-validation tests.

[17] https://docs.alliancecan.ca/wiki/Narval/en

Stocks Lists

	DJ30 Comonente Stocks		Meme Stocks	
	Ticker	Company name	Ticker	Company name
1	AXP	American Express Co	DWAC	Digital World Acquisition Corp
2	AMGN	Amgen Inc	SOFI	SoFi Technologies Inc
3	AAPL	Apple Inc	DKNG	DraftKings Inc
4	BA	Boeing Co	BB	BlackBerry Ltd
5	CAT	Caterpillar Inc	HOOD	Robinhood Markets Inc
6	CSCO	Cisco Systems Inc	ROKU	Roku Inc
7	CVX	Chevron Corp	AFRM	Affirm Holdings Inc
8	GS	Goldman Sachs Group Inc	UPST	Upstart Holdings Inc
9	HD	Home Depot Inc	LCID	Lucid Group Inc
10	HON	Honeywell International Inc	TDOC	Teladoc Health Inc
11	IBM	International Business Machines Corp	AMD	Advanced Micro Devices, Inc.
12	INTC	Intel Corp	NET	Cloudflare Inc
13	JNJ	Johnson & Johnson	CLF	Cleveland-Cliffs Inc
14	KO	Coca-Cola Co	SQ	Block Inc
15	JPM	JPMorgan Chase & Co	RBLX	Roblox Corp
16	MCD	McDonald’s Corp	WISH	ContextLogic Inc
17	MMM	3M Co	BYND	Beyond Meat Inc
18	MRK	Merck & Co Inc	RIVN	Rivian Automotive Inc
19	MSFT	Microsoft Corp	AMC	AMC Entertainment Holdings Inc
20	NKE	Nike Inc	COIN	Coinbase Global Inc
21	PG	Procter & Gamble Co	MSTR	MicroStrategy Inc
22	TRV	Travelers Companies Inc	SNAP	Snap Inc
23	UNH	UnitedHealth Group Inc	PLTR	Palantir Technologies Inc
24	CRM	Salesforce Inc	GME	GameStop Corp.
25	VZ	Verizon Communications Inc		
26	V	Visa Inc		
27	WBA	Walgreens Boots Alliance Inc		
28	WMT	Walmart Inc		
29	DIS	Walt Disney Co		
30	DOW	Dow Inc		

References

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, *59*(3), 1259-1294.

Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063*.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015, September). A large annotated corpus for learning natural language inference.*Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal.

Chen, P., Sun, Z., Bing, L., & Yang, W. (2017). Recurrent attention network on memory for aspect sentiment analysis. Proceedings of the 2017 conference on empirical methods in natural language processing, Copenhagen, Denmark.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, oct). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar.

Cookson, J. A., Engelberg, J., & Mullins, W. (2022). Echo Chambers. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3603107

Cookson, J. A., Lu, R., Mullins, W., & Niessner, M. (2022). The Social Signal. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4241505

Cookson, J. A., & Niessner, M. (2019). Why Don't We Agree? Evidence from a Social Network of Investors. *The Journal of Finance*, *75*(1), 173-228.

Dai, J., Yan, H., Sun, T., Liu, P., & Qiu, X. (2021, Jun). Does syntax matter? A strong baseline for Aspect-based Sentiment Analysis with RoBERTa.*Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, *53*(9), 1375-1388.

Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., & Xu, K. (2014). Adaptive recursive neural network for target-dependent twitter sentiment classification. Proceedings of the 52nd annual meeting of the association for computational linguistics (ACL),

Farzindar, A. A., & Inkpen, D. (2020). *Natural Language Processing for Social Media* (3 ed.). Springer Cham.

Gao, T., Fisch, A., & Chen, D. (2021, August). Making Pre-trained Language Models Better Few-shot Learners.*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online.

Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1-309.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261-266.

Hirshleifer, D. (2020). Presidential Address: Social Transmission Bias in Economics and Finance. *The Journal of Finance*, *75*(4), 1779-1831.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, *9*(8), 1735–1780.

Huang, A. H., Zang, A. Y., & Zheng, R. (2014). Evidence on the Information Content of Text in Analyst Reports. *The Accounting Review*, *89*(6), 2151-2180.

Jiang, H., Liu, P., F. Roch, A., & Zhou, X. (2022). *Social Media and Bitcoin Price Dynamics*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020, July). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online.

Li, H. (2022). Language models: past, present, and future. *Communications of the ACM*, *65*(7), 56-63.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loughran, T. I. M., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35-65.

Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., & Balahur, A. (2018). *WWW'18 Open Challenge: Financial Opinion Mining and Question Answering* Companion Proceedings of the The Web Conference 2018, Lyon, France.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, *65*(4), 782-796.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies,

Ni, J., Li, J., & McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP),

Pedersen, L. H. (2022). Game on: Social networks and markets. *Journal of Financial Economics*, *146*(3), 1097-1119.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar.

Pérez-Mayos, L., Ballesteros, M., & Wanner, L. (2021, November). How much pretraining data do language models need to learn syntax?*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* The 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014, August). SemEval-2014 Task 4: Aspect Based Sentiment Analysis.*Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Renault, T. (2017). Intraday online investor sentiment and return patterns in the U.S. stock market. *Journal of Banking & Finance*, *84*, 25-40.

Rietzler, A., Stabinger, S., Opitz, P., & Engl, S. (2020). Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification. Proceedings of the 12th Language Resources and Evaluation Conference, Marseille.

Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). *Transfer Learning in Natural Language Processing* Proceedings of the 2019 Conference of the North, Minneapolis, Minnesota.

Schick, T., & Schütze, H. (2021, April). Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference.*Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* 16th Conference of the European Chapter of the Association for Computational Linguistics, Online.

Seoh, R., Birle, I., Tak, M., Chang, H.-S., Pinette, B., & Hough, A. (2021). Open Aspect Target Sentiment Classification with Natural Language Prompts. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana / Online.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020, November). AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts.*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the 2013 conference on empirical methods in natural language processing, Seattle, Washington.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, Cambridge, MA, USA.

Tang, D., Qin, B., & Liu, T. (2016, nov). Aspect Level Sentiment Classification with Deep Memory Network.*Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas.

Tian, Y., Chen, G., & Song, Y. (2021). Enhancing aspect-level sentiment analysis with word dependencies. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online.

Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. Proceedings of the 2016 conference on empirical methods in natural language processing, Austin, Texas.

Williams, A., Nangia, N., & Bowman, S. (2018, June). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference.*Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans.

Huang, A. H., Wang, H., & Yang, Y. (2022). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research*.

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong.

List of Tables

Model	Laptop		Restaurant		Twitter	
	Acc	F1	Acc	F1	Acc	F1
BERT-ADA	80.23	75.77	86.22	79.79	-	-
FT-RoBERTa GRAT	83.33	79.95	87.52	81.29	75.81	74.91
BERT-large KVMN	80.41	77.38	86.88	80.92	75.14	73.68
BERT PLM	81.10	76.83	87.5	80.78	-	-
GPT-2 PLM	80.73	77.13	86.99	80.02	-	-
BERT NLI	77.58	73.18	85.07	77.53	-	-
Prompt-TSA	**83.86**	**80.61**	**88.30**	**81.42**	**76.59**	**75.07**

Table 1. Models' performance on general TSA datasets. The best performance for each dataset is indicated in bold. The results of the other models are taken as reported from the corresponding papers: BERT-ADA (Rietzler et al., 2020), FT-RoBERTa GRAT (Dai et al., 2021), BERT-large KVMN (Tian et al., 2021), BERT PLM, GPT-2 PLM, and BERT NLI (Seoh et al., 2021).

Model	Social Media		FiQA	
	Acc	F1	Acc	F1
Prompt-TSA (SclMd)	**83.17**	79.92	68.80	63.18
Prompt-TSA (FiQA)	61.62	59.35	**81.6**	78.3
LM Dictionary	39.48	39.32	34.10	34.12
FinRoBERTa (FPB)	40.37	40.35	50.98	51.71

Table 2. Models' performance on financial TAS datasets. The name in the parenthesis indicates the dataset on which the model is finetuned on.

	Total Stocks						Meme Stocks						DJ30 Stocks					
	Nb_Obs	Mean	Std	25%	50%	75%	Nb_Obs	Mean	Std	25%	50%	75%	Nb_Obs	Mean	Std	25%	50%	75%
Return	15,059	0.00	0.06	-0.01	-0.00	0.01	6,268	0.00	0.09	-0.02	-0.00	0.03	8,791	-0.00	0.02	-0.01	-0.00	0.01
Volume	15,059	15.95	1.10	15.14	15.85	16.65	6,268	16.05	1.28	15.11	15.99	16.98	8,791	15.88	0.95	15.16	15.79	16.47
Sentiment_Twitter_1	15,059	0.16	0.28	-0.05	0.13	0.38	6,268	0.38	0.19	0.27	0.41	0.51	8,791	-0.00	0.21	-0.13	-0.00	0.12
Sentiment_Twitter_2	15,059	0.16	0.44	-0.12	0.18	0.47	6,268	0.39	0.31	0.21	0.42	0.61	8,791	-0.01	0.44	-0.27	-0.01	0.24
Sentiment_Twitter_3	15,059	0.15	0.37	-0.08	0.16	0.41	6,268	0.36	0.26	0.19	0.37	0.54	8,791	0.00	0.37	-0.21	-0.00	0.21
Nb_Message_Twitter	15,059	217.53	1,472.59	3.00	34.00	154.00	6,268	493.04	2,245.87	74.00	164.00	332.00	8,791	21.10	161.08	-9.00	7.00	27.00
Sentiment_Stocktwits_1	15,059	0.07	0.26	-0.09	0.07	0.24	6,268	0.19	0.19	0.06	0.18	0.31	8,791	-0.02	0.28	-0.18	-0.02	0.14
Sentiment_Stocktwits_2	15,059	0.18	0.47	-0.12	0.21	0.54	6,268	0.45	0.32	0.27	0.50	0.68	8,791	-0.02	0.47	-0.29	-0.02	0.25
Nb_Message_Stocktwits	15,059	498.06	3,063.64	2.00	31.00	189.00	6,268	1,159.58	4,661.88	82.00	216.00	599.50	8,791	26.39	221.75	-5.00	5.00	21.00
Sentiment_Reddit_1	15,059	0.04	0.50	-0.19	0.03	0.30	6,268	0.12	0.34	-0.07	0.10	0.30	8,791	-0.01	0.59	-0.33	-	0.30
Nb_Message_Reddit	15,059	88.62	542.34	-	5.00	26.00	6,268	207.14	822.73	9.00	26.00	96.00	8,791	4.11	64.18	-3.00	1.00	5.00
Price range	15,059	0.04	0.05	0.02	0.03	0.05	6,268	0.06	0.06	0.04	0.05	0.07	8,791	0.02	0.01	0.01	0.02	0.02
Disagreement_Twitter_1	15,059	0.74	0.11	0.67	0.74	0.81	6,268	0.77	0.10	0.70	0.77	0.84	8,791	0.71	0.11	0.65	0.72	0.79
Disagreement_Twitter_2	15,059	0.63	0.17	0.51	0.63	0.75	6,268	0.66	0.16	0.56	0.67	0.79	8,791	0.60	0.17	0.50	0.59	0.72
Disagreement_Twitter_3	15,059	0.67	0.17	0.56	0.69	0.80	6,268	0.72	0.15	0.63	0.74	0.83	8,791	0.63	0.18	0.52	0.66	0.77
Disagreement_Stocktwits_1	15,059	0.80	0.15	0.73	0.85	0.91	6,268	0.88	0.07	0.85	0.90	0.93	8,791	0.75	0.16	0.65	0.77	0.87
Disagreement_Stocktwits_2	15,059	0.59	0.20	0.48	0.59	0.74	6,268	0.64	0.18	0.52	0.65	0.79	8,791	0.56	0.20	0.45	0.54	0.70
Disagreement_Reddit_1	15,059	0.78	0.32	0.76	0.89	0.94	6,268	0.84	0.23	0.83	0.90	0.94	8,791	0.74	0.37	0.67	0.88	0.95
Price volatility	15,059	0.00	0.02	0.00	0.00	0.00	6,268	0.00	0.03	0.00	0.00	0.00	8,791	0.00	0.00	0.00	0.00	0.00

Table 3. Descriptive statistics. This table presents the descriptive statistics of all variables used in this paper for 30 DowJones and 24 meme stocks between July 2020 and March 2022. *Return* and *Volume* are stocks' daily return and trading volume, respectively. *Sentiment_Twitter_1*, *Sentiment_Twitter_2*, and *Sentiment_Twitter_3* are equally-weighted, followers-weighted, and retweets-number weighted Twitter sentiments. *Sentiment_Stocktwits_1* and *Sentiment_Stocktwits_2* are equally-weighted and followers-weighted Stocktwits sentiments. *Sentiment_Reddit_1* is for equally-weighted Reddit sentiment. *Nb_Message* is the corresponding number of messages for each source. *Price_Range* and *Price_volatility* are daily price range and 5minute realised volatility based on intraday transactions. 25%, 50%, and 75% relate to the first, the second, and the third quartile, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
S_Twitter_1	0.0498***						
	(13.2455)						
MmS_Twitter_1	0.1327***						
	(14.6748)						
S_Twitter_2		0.0112***					
		(12.0159)					
MmS_Twitter_2		0.0266***					
		(24.3864)					
S_Twitter_3			0.0113***				
			(6.5413)				
MmS_Twitter_3			0.0416***				
			(16.2421)				
S_Stocktwits_1				0.0229***			
				(6.4512)			
MmS_Stocktwits_1				0.1983***			
				(17.5822)			
S_Stocktwits_2					0.0058***		
					(6.3975)		
MmS_Stocktwits_2					0.0087***		
					(5.2196)		
S_Reddit_1						0.0029***	
						(3.8774)	
MmS_Reddit_1						0.0212***	
						(13.7659)	
S_PCA							0.0038***
							(5.9399)
MmS_PCA							0.0248***
							(15.0278)
Volume	0.0219***	0.0226***	0.0208***	0.0221***	0.0213***	0.0224***	0.0227***
	(38.9710)	(61.4118)	(23.5073)	(28.2971)	(22.7930)	(38.2882)	(33.3196)
WSJ	-	-	-	-	-0.0013**	-	-
	0.0019***	0.0019***	0.0017***	0.0017***		0.0020***	0.0018***
	(-2.7806)	(-4.5665)	(-3.8770)	(-3.6724)	(-2.4670)	(-5.4811)	(-4.0568)
FF1	-	-	-	-	-	-	-
	0.0026***	0.0019***	0.0018***	0.0020***	0.0014***	0.0014***	0.0017***
	(-11.7955)	(-14.1803)	(-10.5128)	(-12.9558)	(-6.1174)	(-6.3102)	(-7.9092)
FF2	0.0031***	0.0034***	0.0029***	0.0029***	0.0030***	0.0032***	0.0034***
	(12.1840)	(15.1271)	(8.5825)	(12.1932)	(9.9798)	(11.9348)	(15.8107)
FF3	-	-	-	-	-	-	-
	0.0033***	0.0035***	0.0032***	0.0026***	0.0026***	0.0031***	0.0035***
	(-9.0942)	(-11.8679)	(-8.3225)	(-5.2025)	(-4.9995)	(-7.8322)	(-11.0296)
FF4	-	-	-	-	-	-	-
	0.0069***	0.0078***	0.0068***	0.0064***	0.0066***	0.0071***	0.0074***
	(-15.5544)	(-18.2520)	(-10.9025)	(-10.6967)	(-9.5207)	(-13.4346)	(-15.0101)
FF5	0.0136***	0.0147***	0.0134***	0.0131***	0.0132***	0.0132***	0.0143***
	(15.7694)	(19.5966)	(15.2149)	(12.0934)	(13.0858)	(13.4211)	(18.9979)
Constant	-	-	-	-	-	-	-
	0.3595***	0.3614***	0.3324***	0.3580***	0.3390***	0.3557***	0.3587***
	(-38.7835)	(-60.8446)	(-23.1281)	(-27.9950)	(-22.5001)	(-38.9730)	(-30.8695)
Observations	11,372	11,372	11,372	11,372	11,372	11,372	11,372
Num of stockid	54	54	54	54	54	54	54

Table 4. Marginal effect of social media sentiment on return. The table presents the marginal impact of social media on meme stocks' daily return. $S\_Twitter(Stocktwits/$

Reddit). means the sentiment measured from Twitter (Stocktwits/Reddit). Prefix Mm means the corresponding sentiment multiplied by a dummy variable indicating if the stock is a meme stock. Models (1) - (3) are about equally-weighted, followers-weighted, and retweets-number weighted Twitter sentiments, models (4) - (5) are about equally-weighted and followers-weighted Stocktwits sentiments. Model (6) is for equally-weighted Reddit sentiment. Finally, model (7) reports the results of main component of all sentiments from different sources. *Volume* is the daily trading volume. *WSJ* and *MarketRet* are the Wall Street Journal sentiment and market return, respectively. t-statistics are in parentheses. ***, **, and * represent statistical significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
STwitter_1	-0.0073***						
	(-13.3516)						
MmS_Twitter_1	-0.0092***						
	(-13.9704)						
STwitter_2		-0.0191***					
		(-118.7260)					
MmS_Twitter_2		0.0153***					
		(97.0768)					
STwitter_3			-0.0020***				
			(-6.8683)				
MmS_Twitter_3			-0.0026***				
			(-8.0058)				
SStocktwits_1				-0.0032***			
				(-10.9896)			
MmS_Stocktwits_1				-0.0027***			
				(-6.1723)			
SStocktwits_2					-0.0010***		
					(-8.7874)		
MmS_Stocktwits_2					-0.0005***		
					(-4.3333)		
SReddit_1						-0.0002	
						(-1.6361)	
MmS_Reddit_1						-0.0016***	
						(-10.0668)	
S_PCA							0.0047***
							(44.2295)
MmS_PCA							-0.0067***
							(-68.6025)
Volume	0.0116***	0.0013***	0.0116***	0.0116***	0.0116***	0.0116***	0.0013***
	(305.8988)	(920.2016)	(201.7366)	(340.6712)	(384.5871)	(260.7405)	(772.5265)
WSJ	0.0066***	-0.0050***	0.0063***	0.0064***	0.0064***	0.0065***	0.0212***
	(74.2768)	(-111.9606)	(23.5152)	(43.5483)	(26.1650)	(26.1458)	(460.4797)
MarketRet	0.0005***	0.0009***	0.0004***	0.0005***	0.0004***	0.0004***	0.0011***
	(39.4674)	(71.5498)	(39.1871)	(31.8869)	(34.6334)	(35.1245)	(99.1528)
Constant	-0.1829***	-0.1945***	-0.1831***	-0.1829***	-0.1835***	-0.1836***	-0.1687***
	(-264.2760)	(-11.9023)	(-178.3267)	(-304.9551)	(-342.6001)	(-232.2836)	(-7.4360)
Observations	11,372	11,372	11,372	11,372	11,372	11,372	11,372
Number of stockid	54	54	54	54	54	54	54

Table 5. Marginal effect of social media on volatility. The independent variables are the same as in table 4.

	Prompt-TSA						FinRoBERTa						LM Dictionary					
	Twitter_1	Twitter_2	Twitter_3	Stocktwits_1	Stocktwits_2	Reddit	Twitter_1	Twitter_2	Twitter_3	Stocktwits_1	Stocktwits_2	Reddit	Twitter_1	Twitter_2	Twitter_3	Stocktwits_1	Stocktwits_2	Reddit
DWAC	1.53	0.47	0.81	0.91	0.17	0.33	0.24	0.05	-0.55	-0.31	-0.46	0.41	1.38	-0.24	0.14	0.33	0.47	-0.20
SOFI	0.31	-0.14	0.17	-0.02	0.02	0.10	0.28	-0.07	-0.30	0.21	0.11	0.26	-0.19	0.06	0.03	0.27	2.72	0.52
DKNG	-0.22	-0.26	-0.16	0.41	0.01	-0.14	-0.35	-0.34	-0.33	-0.24	-0.25	0.29	-0.28	-0.32	-0.26	-0.04	-0.23	-0.20
BB	0.97	-0.05	0.09	0.17	0.60	-0.12	-0.27	-0.38	-0.27	-0.24	-0.10	0.00	0.00	-0.31	-0.32	0.03	-0.09	-0.24
HOOD	-0.06	-0.54	0.17	0.48	0.48	0.28	0.75	-0.38	-0.46	-0.30	0.08	2.20	-0.39	-0.17	-0.35	-0.46	-0.36	-0.29
ROKU	0.42	0.18	0.07	-0.02	-0.17	0.76	-0.19	-0.20	0.04	-0.35	-0.28	1.02	-0.05	-0.19	-0.18	-0.11	-0.21	-0.13
AFRM	1.77	0.44	-0.20	0.01	0.14	1.89	-0.26	-0.41	-0.16	-0.04	-0.16	1.95	-0.39	-0.26	-0.40	-0.16	0.75	-0.05
UPST	-0.22	0.29	0.99	-0.12	-0.38	0.63	0.65	-0.24	-0.24	-0.17	-0.22	-0.27	-0.32	-0.35	-0.29	-0.23	-0.30	-0.18
LCID	-0.16	-0.34	0.37	-0.22	-0.19	-0.42	1.21	0.88	1.14	1.57	0.16	0.50	-0.47	-0.16	-0.31	0.47	0.16	0.27
TDOC	0.07	0.14	-0.14	0.17	-0.17	0.31	-0.40	-0.18	-0.36	-0.22	-0.37	-0.26	-0.11	-0.04	-0.24	-0.19	0.01	-0.18
AMD	-0.14	0.17	0.25	0.13	0.24	-0.14	0.00	-0.16	-0.18	-0.21	-0.29	-0.25	-0.01	-0.24	-0.23	0.07	-0.03	-0.21
NET	-0.21	-0.27	0.28	0.57	-0.11	0.07	-0.22	-0.27	-0.34	-0.31	-0.28	0.10	0.01	-0.30	0.11	-0.03	-0.24	-0.23
CLF	-0.07	0.58	-0.14	-0.11	-0.12	1.05	-0.07	-0.27	-0.08	-0.19	0.13	0.33	0.06	-0.26	0.17	1.01	0.18	-0.06
SQ	0.45	-0.03	-0.13	-0.04	-0.11	0.02	0.09	-0.08	-0.29	-0.19	-0.20	0.22	-0.02	-0.07	-0.20	-0.28	-0.18	-0.15
RBLX	0.57	-0.15	-0.19	-0.16	-0.22	0.64	-0.12	-0.09	-0.30	-0.15	0.25	-0.25	-0.28	-0.23	-0.21	0.26	-0.20	-0.24
WISH	0.26	-0.29	-0.24	-0.19	-0.26	-0.07	-0.10	-0.25	0.12	-0.34	-0.34	0.18	0.15	0.00	-0.06	-0.21	-0.01	-0.27
BYND	-0.14	0.27	1.33	-0.19	0.17	0.10	-0.17	0.01	-0.16	-0.38	-0.33	-0.14	-0.28	-0.32	-0.32	0.06	-0.22	-0.22
RIVN	-0.32	0.49	-0.27	-0.57	1.84	-0.65	-0.62	-0.33	-0.24	-0.07	2.08	1.14	-0.47	0.96	0.30	-0.20	-0.51	-0.06
AMC	-0.03	-0.19	-0.03	-0.32	-0.27	-0.33	-0.31	-0.35	-0.35	-0.55	-0.30	0.83	-0.35	-0.35	-0.35	-0.35	-0.99	-0.31
COIN	-0.15	-0.19	0.26	-0.11	-0.03	0.38	0.23	0.04	-0.05	0.08	0.06	0.05	0.07	0.20	0.03	-0.05	0.06	-0.19
MSTR	-0.17	0.22	0.00	0.41	-0.15	0.74	-0.83	-0.80	-1.01	1.55	1.23	0.03	-0.39	-0.36	-0.39	1.19	-0.27	-0.28
SNAP	0.30	-0.15	1.69	1.07	0.26	-0.17	-0.17	-0.04	-0.24	0.60	0.58	0.11	0.25	0.04	-0.26	-0.03	0.40	0.45
PLTR	-0.15	-0.12	0.15	0.23	0.32	-0.28	0.14	-0.09	0.04	0.63	0.37	0.00	-0.22	-0.12	-0.28	-0.17	-0.21	-0.32
GME	2.41	-0.25	0.25	-0.31	-0.09	-0.33	-0.35	-0.35	-0.35	-0.52	-0.35	-0.35	-0.34	-0.48	-0.36	-0.36	11.22	-0.36
Average	0.29	0.01	0.22	0.09	0.08	0.19	-0.04	-0.18	-0.21	-0.01	0.05	0.34	-0.11	-0.15	-0.18	0.03	0.50	-0.13
p_value_1	0.082	0.055	0.002	0.466	0.696	0.348												
p_value_2	0.010	0.056	0.002	0.592	0.418	0.016												

Table 6. Comparison of returns derived from different sentiments for meme stocks. The table compares the average daily return (in percentage) of meme stocks for strategies based on sentiments issued from Prompt-TSA, FinRoBERTa, and LM dictionary. *Twitter_1*, *Twitter_2*, and *Twitter_3* are strategies based on equally-weighted, followers-weighted, and retweets-number weighted Twitter sentiments. *Stocktwits_1* and *Stocktwits_2* are strategies based on equally-weighted and followers-weighted Stocktwits sentiments. *Reddit_1* is strategy based on equally-weighted Reddit sentiment. *p_value_1* is the p_value for the test of mean equality between Prompt-TSA and FinRoBERTa. *p_value_2* is the *p_value* for the test of mean equality between Prompt-TSA and LM Dictionary.

	Prompt-TSA						FinRoBERTa						LM Dictionary					
	Twitter_1	Twitter_2	Twitter_3	Stocktwits_1	Stocktwits_2	Reddit	Twitter_1	Twitter_2	Twitter_3	Stocktwits_1	Stocktwits_2	Reddit	Twitter_1	Twitter_2	Twitter_3	Stocktwits_1	Stocktwits_2	Reddit
AXP	-0.01	0.07	0.03	0.10	-0.01	0.09	0.05	-0.05	-0.05	0.18	0.13	0.21	-0.02	0.05	-0.07	-0.05	-0.04	0.05
AMGN	0.21	-0.04	0.11	-0.07	0.02	0.00	-0.01	0.00	-0.12	0.05	-0.09	-0.19	-0.05	-0.10	-0.04	-0.07	-0.18	-0.05
AAPL	0.03	-0.18	0.14	0.54	-0.08	0.21	-0.13	-0.07	-0.04	0.00	-0.11	-0.04	-0.09	-0.10	-0.08	-0.14	-0.09	-0.13
BA	-0.18	0.02	0.02	-0.08	0.09	-0.03	-0.11	-0.16	-0.04	-0.14	-0.13	-0.04	-0.09	-0.06	-0.11	-0.12	-0.01	-0.07
CAT	-0.07	-0.09	-0.03	-0.12	-0.08	0.05	-0.18	-0.15	-0.11	-0.09	-0.11	0.04	-0.06	-0.11	-0.07	-0.11	-0.11	0.00
CSCO	-0.12	-0.07	-0.13	-0.13	-0.02	-0.09	-0.05	-0.03	-0.04	-0.10	-0.12	-0.08	-0.01	-0.07	-0.06	-0.10	-0.01	0.01
CVX	-0.06	0.11	-0.06	-0.08	-0.09	-0.08	0.01	-0.02	0.06	-0.13	-0.14	-0.05	-0.05	0.01	-0.08	-0.05	-0.06	-0.01
GS	-0.14	0.12	-0.07	0.05	-0.15	-0.05	-0.07	-0.04	-0.11	-0.07	-0.08	-0.03	-0.08	-0.08	-0.08	-0.03	-0.07	-0.07
HD	-0.15	-0.14	-0.15	0.01	-0.10	0.00	-0.11	-0.02	-0.11	-0.04	-0.03	-0.08	-0.09	-0.12	-0.14	-0.05	-0.10	-0.11
HON	0.04	-0.03	-0.03	-0.02	-0.07	-0.06	-0.08	-0.01	-0.14	0.00	0.06	-0.14	-0.04	-0.09	-0.05	-0.02	-0.07	-0.09
IBM	0.04	-0.05	-0.11	0.00	-0.03	0.04	-0.05	-0.11	-0.08	-0.06	0.01	-0.08	-0.09	-0.10	-0.11	0.01	-0.10	-0.02
INTC	-0.05	0.06	0.14	-0.14	-0.01	-0.04	-0.16	-0.16	-0.15	-0.09	-0.08	-0.13	-0.10	-0.06	-0.15	-0.05	-0.12	0.02
JNJ	-0.19	-0.16	-0.16	-0.17	-0.16	-0.14	-0.04	-0.02	-0.05	0.01	0.02	-0.06	-0.03	-0.07	-0.06	-0.03	-0.06	-0.07
KO	-0.05	-0.06	-0.05	-0.13	-0.11	-0.08	-0.08	-0.11	-0.07	0.00	-0.04	0.05	-0.05	-0.05	-0.05	-0.07	-0.11	-0.08
JPM	-0.01	-0.13	0.14	-0.12	0.09	0.06	-0.12	-0.04	-0.06	0.00	-0.05	-0.07	-0.12	-0.07	-0.11	-0.07	0.00	-0.08
MCD	-0.12	-0.17	-0.14	-0.18	-0.11	-0.10	-0.11	-0.11	-0.10	-0.05	-0.05	-0.02	-0.05	-0.06	-0.07	-0.07	-0.04	-0.06
MMM	-0.01	-0.08	-0.14	-0.13	-0.14	-0.13	-0.13	-0.10	-0.09	-0.08	-0.09	-0.02	-0.09	-0.12	-0.09	-0.08	-0.05	0.03
MRK	-0.04	-0.04	-0.07	-0.02	-0.09	-0.09	-0.12	-0.07	0.01	-0.03	-0.03	0.03	-0.08	-0.08	-0.12	-0.06	-0.04	0.02
MSFT	-0.14	-0.10	0.02	-0.05	-0.15	-0.09	-0.09	-0.04	-0.06	-0.09	-0.07	-0.07	-0.03	-0.07	-0.04	-0.11	-0.03	-0.09
NKE	-0.01	-0.09	0.01	0.11	-0.03	-0.03	-0.09	-0.15	-0.11	-0.02	-0.05	-0.06	-0.09	-0.06	-0.07	-0.16	-0.09	-0.10
PG	-0.10	-0.03	0.00	-0.03	-0.13	-0.08	-0.08	-0.10	-0.10	0.01	0.03	0.07	-0.04	-0.08	-0.08	-0.03	-0.03	-0.07
TRV	-0.42	-0.17	-0.47	-0.36	-0.49	-0.32	-1.08	0.02	-0.38	-0.36	-0.36	0.27	-0.30	-0.08	-0.25	0.09	-0.09	0.00
UNH	0.06	0.04	-0.12	0.01	-0.01	-0.05	0.12	0.18	0.20	-0.03	-0.02	0.04	-0.06	-0.09	-0.08	-0.02	-0.07	0.02
CRM	0.29	0.04	-0.07	0.19	0.08	0.16	-0.19	0.09	-0.11	-0.18	-0.09	0.10	-0.11	-0.09	-0.04	-0.09	-0.11	-0.05
VZ	-0.06	-0.13	-0.08	-0.06	-0.03	-0.08	-0.02	-0.02	0.02	0.04	0.00	0.00	-0.07	-0.04	-0.05	-0.07	-0.01	-0.04
V	-0.06	-0.04	-0.15	0.04	-0.07	-0.05	-0.02	-0.07	-0.05	0.01	0.02	-0.03	-0.07	-0.01	-0.13	-0.10	-0.10	-0.05
WBA	0.00	0.03	0.03	-0.06	0.08	0.03	0.06	0.08	-0.01	-0.04	0.00	0.03	-0.10	-0.07	-0.06	-0.14	-0.13	-0.16
WMT	-0.17	-0.12	-0.11	-0.17	-0.13	-0.01	-0.09	-0.09	-0.10	-0.05	-0.09	-0.12	-0.11	-0.11	-0.08	-0.10	-0.02	-0.08
DIS	-0.08	-0.18	0.10	0.00	-0.07	0.00	-0.10	-0.08	-0.05	-0.09	-0.06	-0.03	-0.06	0.00	0.02	-0.08	0.04	-0.08
DOW	-0.09	-0.06	-0.02	-0.12	-0.19	-0.01	-0.23	-0.17	-0.29	-0.25	-0.27	-0.20	-0.10	-0.06	0.01	-0.02	-0.07	-0.13
Average	-0.06	-0.06	-0.05	-0.04	-0.07	-0.03	-0.11	-0.05	-0.08	-0.06	-0.06	-0.02	-0.08	-0.07	-0.08	-0.07	-0.07	-0.05
p_value_1	0.087	0.955	0.221	0.562	0.583	0.764												
p_value_2	0.308	0.435	0.108	0.409	0.771	0.405												

Table 7. Comparison of returns derived from different sentiment for DJ30 stocks. The table compares the average daily return (in percentage) of DowJones30 stocks for strategies based on sentiments issued from Prompt-TSA, FinRoBERTa, and LM dictionary. The same as in table 6, *Twitter_1*, *Twitter_2*, and *Twitter_3* are strategies based on equally-weighted, followers-weighted, and retweets-number weighted Twitter sentiments. *Stocktwits_1* and *Stocktwits_2* are strategies based on equally-weighted and followers-weighted Stocktwits sentiments. *Reddit_1* is strategy based on equally-weighted Reddit sentiment. *p_value_1* is the p_value for the test of mean

equality between Prompt-TSA and FinRoBERTa. *p_value_2* is the p_value for the test of mean equality between Prompt-TSA and LM Dictionary.

List of Figures

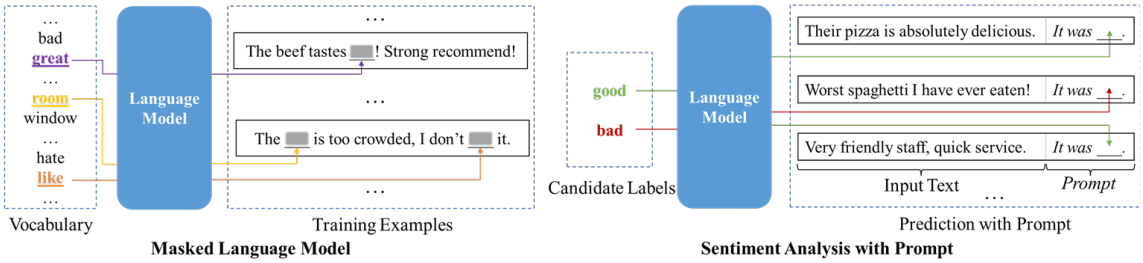

Figure 1. Example of prompting method. The left illustrates the masked language model (MLM) pretraining: we randomly mask words for a given training example, and let the PLM predict which words in the vocabulary are most likely to be the masked words. The right illustrates performing sentiment analysis with prompt based on the PLM: for an input text, we append a prompt *"It was ___."* to its end, and let the PLM fill in the empty slot with the most probable word from the candidate labels. The word chosen can be converted to the prediction about the sentiment of the input text.

Text	Target	Label
Apple beats Microsoft.	Apple	positive

Original Data

Text	Prompt	Label
Apple beats Microsoft.	The sentiment towards *Apple* is **positive**.	True
	The sentiment towards *Apple* is **neutral**.	False
	The sentiment towards *Apple* is **negative**.	False

Text-Prompt Pairs with Binary Label

Figure 2. Construction of prompts from the original labeled training data. The prompt is formatted as an explicit statement of targeted sentiment, "The sentiment towards [*target*] is [**label**].", with clozes filled by the target (shown in italic) and the candidate sentiment labels (shown in bold).

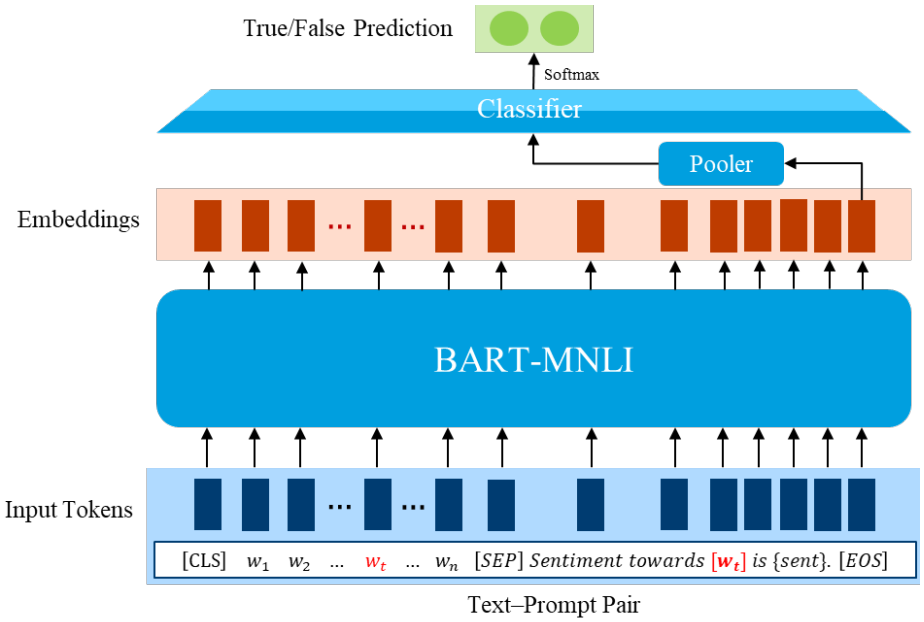

Figure 3. Illustration of the prompt-based targeted sentiment model. The backbone model is BART-MNLI. The last layer of the classifier is modified to make binary instead of 3-class classification. The input of to the model is in the form of text-prompt pair joined by special tokens, displayed in token level. The token in red color represents the targeted word / phrase, which appears in both the input text and the prompt formed. The embedding of the EOS token is inputted into the classifier to make prediction.

Text	Target	Label Vote
Apple beats Microsoft.	Apple	positive: 4 neutral: 1 negative: 0

Original Data

Text	Prompt	Probability
Apple beats Microsoft.	The sentiment towards *Apple* is **positive**.	0.8
	The sentiment towards *Apple* is **neutral**.	0.2
	The sentiment towards *Apple* is **negative**.	0

Text-Prompt Pairs with Soft Label

Figure 4. Prompts construction with soft-label based on the full original annotation data. The "Label Vote" column denotes how many people voted for each label. The probability score is calculated as the number of votes divided by total number of annotators.

Chapter 3
Transformer Model for Subjectivity of Financial Text

Abstract[18]

Distinguishing between subjective and objective statements is an important topic in natural language processing. However, little research has explored this topic in the context of finance. In this paper, we present a novel transformer model that measures the subjectivity of financial texts. We first create a dataset of financial texts labeled for subjectivity by expert annotators. We then design a prompt-based transformer model that specifically incorporates financial vocabulary and is pre-trained on financial semantics tasks. Finally, we finetune the model on our labeled dataset, achieving high test accuracy and F1 score.

[18] This essay is coauthored with Gilles Caporossi (gilles.caporossi@hec.ca) and Hongping Tan (hongping.tan@mcgill.ca). It is an methodological extension from another paper that I coauthored with Hongping Tan and Yaping Zheng (Liu, P., Tan, H., & Zheng, Y. (2022). *Does Opinion Pay off? Evidence from Analyst Report Subjectivity Using Machine Learning*.), which was presented at FMA 2022.

3.1 Introduction

Subjective statements reflect personal opinions of the writer, taking various form such as judgment, feelings, desires, views, beliefs or conclusions. In contrast, objective statements are descriptions of events and facts that can be proven or verified. Subjectivity analysis is a fundamental task in natural language processing (NLP) that involves distinguishing between subjective and objective statements in text data. It is crucial to discern subjectivity in text for many NLP applications, including document summarization, opinion mining, and question answering. As a result, subjectivity analysis has long been a subject of active research in the field of NLP (Mäntylä et al., 2018).

Like many NLP tasks, subjectivity analysis is domain-dependent, because the interpretation of a text's subjectivity can vary depending on the specific context in which it appears. Previous research has demonstrated that different domains may have their unique lexicons of words that indicate subjectivity (Dehkharghani et al., 2012), as well as distinct usage patterns of common subjectivity terms (Karamibekr & Ghorbani, 2013). In finance, for example, certain words that are irrelevant to subjectivity in other domains, such as "bear" and "bull", may be considered indicative of opinions; while some common objective financial expressions such as "earnings surprise", "analyst forecasts" and "market expectation" may contain words that could be strong clues of subjectivity in other contexts. Furthermore, the definition of subjectivity itself — i.e., which types of statements are considered opinions or facts — can also be nuanced and complex depending on the specific context. Therefore, a domain-specific subjectivity analysis model is often necessary to achieve the best performance in a specialized domain such as finance.

There is a significant and expanding body of research in finance and accounting that establishes the value of textual data in providing additional insights beyond quantitative data. Previous studies have investigated a range of textual aspects including sentiment (Antweiler & Frank, 2004; Huang et al., 2014; Kothari et al., 2009; Tetlock, 2007), topics (Dyer et al., 2017; Huang et al., 2017), readability (De Franco et al., 2015; Lo et al., 2017), and occurrences of terms concerning a particular concept (S. Huang et al., 2022; Loughran et al., 2009).

More recently, there has been an increased focus on the subjectivity present in various types of financial documents. Kogan & Meursault (2021) investigate the different functions of facts and opinions in earnings call transcripts. Liu et al. (2022) examine the impacts of financial analysts' subjectivity in their research reports. While these studies delve into the financial consequences of subjectivity, they primarily concentrate on the financial implications rather than the methodology of subjectivity analysis. The former trains a convolutional neural network (CNN), a traditional machine learning model that is not optimized for financial text analysis; and the later employs OpinionFinder, a subjectivity model off-the-shelf developed for general purpose. However, there is a lack of specialized financial subjectivity models, due in part to the challenge of data labeling that requires financial expertise, and the limited research dedicated to developing NLP methodology tailored to the finance domain.

In this paper, we aim to address the gap by developing a novel, powerful NLP model that is specifically designed to handle financial texts based on expert-annotated subjectivity data in financial texts. We adopt the prompting paradigm with a large pretrained language model (PLM) as the backbone model, which is the state-of-art NLP model architecture.

We engineer the model to take into account financial vocabulary by adding common financial terms into the tokenizer and embedding layer. We also initialize the added embeddings based on the natural language definitions of these terms. We further construct an auxiliary task of financial term distinction to train the model to better digest the added terms and understand financial semantics.

Based on the wide dissemination and usage in the financial industry, we choose financial analyst reports as our targeted text genre. We draw detailed guidelines on how to annotate subjectivity in this context. With finetuning on this data, our model achieves over 90% test accuracy and F1 score.

3.2 Background and Related Work

3.2.1 Subjectivity Analysis

Riloff & Wiebe, 2003; Wiebe et al. (2004) pioneered the subjectivity analysis research in NLP by developing a process for automatically extracting clues of subjectivity from news texts and evaluating the utility of these clues using labeled datasets. They later published OpinionFinder, a multi-stage subjectivity classifier (Wiebe et al., 2011; Wiebe & Riloff, 2005; Wilson et al., 2005). This model first uses a rule-based classifier to generate a labeled training dataset, which is then used by a pattern learner to extract subjectivity clues. These clues are incorporated back into the rule-based classifier to enhance its recall rate. Finally, a naïve Bayes classifier based on the learned clues and other linguistic features is trained to categorize sentences as subjective or objective.

Besides general subjectivity, researchers have also investigated subjectivity in specific domains. Li et al. (2008) and Bjerva et al. (2020) both explored subjectivity in the context

of online question answering (QA). The former trained a support vector machine (SVM) model on ~1K question texts labeled as subjective or objective. The latter constructed a dataset of over 10K questions and answers annotated with subjectivity to build a subjectivity-aware QA model. Karamibekr & Ghorbani (2013) studied subjectivity in the social domain and proposed an unsupervised method that uses lexical and syntax knowledge specific to the domain to classify sentences as subjective or objective. Recently in the finance domain, Kogan & Meursault (2021) trained a convolutional neural network (CNN) model on ~3.7K sentences from earning calls tagged as opinion, fact, or irrelevant.

Our work differs from Kogan & Meursault (2021) in several key aspects. Firstly, we use a PLM-based prompt model, which is far more advanced for NLP than the vanilla CNN model. A PLM model is pretrained on vast amount of raw texts so that it has a foundation knowledge on language, and the attention mechanism-based architecture makes it very efficient in capturing long-range dependency and complex relationship between words. By comparison the CNN model possesses no prior language knowledge, and can only capture simple local dependency in the input text due to its fixed-size receptive window. Moreover, we also specifically build our model to handle financial vocabulary more effectively while the CNN model was taken out-of-the-box. Secondly, our labeled data consists of ~20K sentences, which is much larger and therefore likely to improve model performance and generalization ability. Lastly, the targeted text genre and the nuanced definition of subjectivity in our study are different from those of Kogan & Meursault (2021). We target financial analyst reports and focus on distinguishing opinions and facts conveyed by financial analysts who are representative of outsider investors' perspective,

while they study corporate disclosure and focus on subjectivity delivered by the management who are company insiders.

3.2.2 Domain adaptation of pretrained language model

When a PLM is directly applied to a specialized domain, it may not perform optimally due to the domain's distinct content and terminology that may differ from the general corpora on which the PLM was pretrained. A typical way to improve domain adaptation is to specifically pretrain the model on large corpora from the target domain. For example, Alsentzer et al. (2019) and Lee et al. (2019) separately pretrain BERT model for clinical text and biomedical text. However, this method can be prohibitively expensive in terms of computation. As an alternative, some recent studies explored cheaper methods for domain adaptation by augmenting the vocabulary of the PLM with additional domain-specific terms. Zhang et al. (2020) extend a LM's vocabulary to include the 5K to 10K most frequently occurring words in the domain that were out-of-vocabulary (OOV). They randomly initialized the embeddings of the added words and then pretrained the model with an auxiliary in-domain task, resulting in improved performance on several IT domain tasks. Poerner et al. (2020) augment the vocabulary with all OOV in-domain words and initialize their embeddings with transformed Word2Vec embeddings learned from an in-domain corpus, leading to performance gain on multiple biomedical tasks. Sachidananda et al. (2021) added 10K domain-specific words, and initialize their embeddings using either the mean of their subword embeddings from the original model (subwords are common word pieces that can be combined into infrequent words, e.g., "neurological" can be decomposed into subwords "neurolog" and "ical"), or the transformed Word2Vec

embeddings similar to those in Poerner et al. (2020). Both methods show similar performance improvements when applied across multiple domains.

In our paper, we augment the vocabulary based on expert-made domain terminology dictionaries instead of the empirical word distribution of a domain corpus. We also initialize the embedding of each term using the mean embeddings of its definition sentence to capture better in-domain meanings compared to using merely subwords of the term itself, which may be domain irrelevant and sometimes meaningless (e.g., Nasdaq would be split into "Nas" and "daq"). Our approach requires minimal computation compared to Word2Vec embedding.

3.3 Data

In the context of financial analyst reports, we define subjectivity as the expression of the analysts' personal opinions on the company or economy, including judgments, feelings, views, beliefs, desires, or conjectures. On the other hand, we define objectivity as description of events or facts including observation, information, or data that can be proven true or false or verified with external sources. We further clarify the subjectivity definition for the following nuanced cases:

a) Analysts' forecasts. If a sentence contains an analyst's forecast, such as recommendation rating, performance estimation, and target price, but does not provide underlying analysis or justification, we define it as objective since it's simply a release of information. In contrast, if the forecast is accompanied by supportive arguments, we consider it subjective because it reveals the analyst's internal thinking and reasoning.

   Examples:

Objective: “After the announcement, we decide to raise our rating on the firm’s shares to strong buy from hold.”

Subjective: “We expect an in-line revenue as we incorporate the market data of last quarter and new exchange rate estimates.”

b) Analysts' assessments and interpretations. Analysts would interpret a result or status by providing reasons, which may themselves be based on facts. Even so, we still label these sentences as subjective because the reasons could have been selectively chosen by the analysts, and other equally informed individuals might find different facts as reasons. Example:

“The weak operating results have been driven by significant declines in the non-residential construction market, as well as an excess of high cost inventory.”

c) Opinions from a third party. Analysts would cite opinions from other information sources including management, guidance (which usually refers to management or company guidance), media, company communication, 3rd party report, etc. We define such sentences as objective since they do not represent the analysts’ own opinions.

Example:

“The zircon market remains extremely tight according to the management, with the greatest threat being substitution due to product availability.”

d) Key risks identified by the analysts. We label them as subjective, because similar to b), although the risks themselves could be facts, the attribution of importance is open to debate.

Example:

> "ABC Software Inc. key investment risks 1) slow growing aggregate end-market…"

We recruit ten research assistants with adequate business education backgrounds to annotate the sentences. To ensure labelling accuracy and consistency, we train those annotators three rounds with the definition guidelines and up to 1,000 exemplary sentences randomly drawn from the universe of analysts reports that we have downloaded. We divide these annotators into five groups, with each group having two annotators to first code 4,000 sentences independently and then reconcile their different labelling afterwards. We require each group to label the first 300 sentences and discuss their labeling with the author who is in charge of the labelling before they label the rest of the assigned sentences. For the sentences that can't achieve a consensus in labelling from all the five groups, we designate one annotator to label the sentences in consultation with the author. Since we deliberately allocate some duplicate sentences to different groups to ensure labeling consistency, we end up with 19,951 annotated sentences to indicate subjectivity or objectivity, including the 1,000 annotated sentences used for the training purpose.

After dropping sentences that are textually corrupted or unable to annotate, the data contains 19,756 labeled sentences. Table 1 shows its summary statistics.

[Insert Table 1 here]

3.4 Methods

3.4.1 Prompt Based Model

We use prompting, a novel NLP paradigm for effectively leveraging large PLMs (Liu et al., 2021). A prompt is a piece of text that is added to the PLM input, allowing the original task to be reformulated to imitate a task that the LM has already been pretrained on. As the backbone PLM for our prompt-based model, we use BART-MNLI (Lewis et al., 2020), a pretrained transformer model combining the encoder of BERT and decoder of GPT, and is further trained on the natural language inference (NLI) task dataset MNLI (Williams et al., 2018). BART-MNLI is known for its strong ability of language understanding and domain generalization. To adapt BART-MNLI for our specific task, we download the pretrained model from HuggingFace[19], and modify the last layer of its three-class classification head to have a binary output, while keeping all other pretrained parameters unchanged. We take the output embedding of the EOS (end-of-sentence) token as the representation of the whole input sentence pair and feed it into the classification head.

Following the prompting paradigm, we reformat our subjectivity analysis task to mimic NLI task. For any input sentence to be classified, we append a prompt text *"The statement is ___."* with the blank to be filled by either *"a fact"* or *"an opinion"*. The model will then predict the probability of each candidate prompt being true given the input text as the premise, and label the sentence as subjective or objective based on which prompt has a higher probability.

3.4.2 Extension of Financial Vocabulary

We download two professionally compiled financial terms dictionaries from public online resources, one from the NYS Society of CPAs[20] consisting of ~1.1K terms with

[19] https://huggingface.co/facebook/bart-large-mnli

[20] https://www.nysscpa.org/professional-resources/accounting-terminology-guide, accessed on 2022-11-05.

definitions, and another ~6.7K from Investopedia[21] which is a financial media website. After merging and processing, 7300 terms are kept (including abbreviations as independent terms), and each has a definition of one or a few sentences. Among them, 6701 terms are OOV for the BART model.

To incorporate those OOV financial terms into the model's vocabulary, we follow the following steps: 1) Extend the model's tokenizer to include the OOV terms. 2) For each term added, retrieve the BART embeddings of the tokens in their definition sentences. 3) Average the retrieved embeddings to create an embedding for the term that captures its meaning from its textual definition. 4) Append the terms' embeddings to the BART model's embedding layer as initial weights.

3.4.3 Auxiliary Task of Financial Term-Definition Matching

It has been shown that adding auxiliary synthetic in-domain tasks can help the PLM model transfer to downstream tasks (Zhang et al., 2020). To enhance the model's understanding of financial semantics, we utilize the Investopedia financial terms dictionary data at our disposal to construct a task in which the model must correctly match terms with their definitions. Specifically, given a definition D and two candidate terms T_1 and T_2, the model is asked to choose the correct term for D solely based on their texts. To add to the difficulty of the task, we select the incorrect term from the same category (as defined by Investopedia, including categories such as corporate finance, economy, etc.) as the correct term, so that T_1 and T_2 are more likely to be similar. Additionally, we remove any direct reference to the term itself from the definition to avoid trivial shortcuts for the model.

[21] https://www.investopedia.com/dictionary, accessed on 2022-11-06. We extract the first section of the "KEY TAKEAWAYS" part as the concise definition of each term. We also delete the reference to the term itself that appears in the beginning of the definition. For instance, the original definition of the term "ADR" is "An ADR is a certificate issued by a U.S. bank that represents shares in foreign stock.", we change it to "A certificate issued by a U.S. bank that represents shares in foreign stock.".

The auxiliary task is trained using the same prompting method as the main task. We design the prompt P as *"It is the definition of ___."*, with the blank to be chosen between T_1 and T_2. The model predicts $Probability(P(T) \mid D)$ for $T\ in\ [T_1, T_2]$ and choose the correct term accordingly. The model is trained on this task after extending and initializing financial vocabulary, and before finally being trained on the analyst subjectivity analysis data.

3.5 Results

3.5.1 Experimental Setups

Our financial vocabulary enhanced model is first trained with the auxiliary financial term-definition matching task. We constructed ~6.6K $[D, T_1, T_2]$ triplets as task data from the Investopedia financial terms dictionary data. The data is split into 80% training and 20% evaluation sets, we choose the optimal training steps according to the evaluation performance, which achieves 96.6% accuracy after about 2 epochs.

Continuing from the model trained on the auxiliary task, we finetune it on our financial analyst report subjectivity data to perform the main task of subjectivity analysis. We use 5-fold cross-validation to evaluate the model's performance to ensure the robustness of the result.

For comparison, we also train on our data a vanilla BART-MNLI model without enhancing the vocabulary or auxiliary training, under the same setups. In addition, we apply OpinionFinder to our data and measure its performance.

The following hyperparameters are fixed for all experiments involving the BART-MNLI model. We use a linear learning rate scheduler with a $base\ learning\ rate\ =\ 5e-06$ and a $warmup\ ratio\ =\ 0.1$. A $label\ smoothing\ factor$ is set to 0.2. The models are

trained on $4 * NVidia\ A100\ GPU$ servers[22], with an $effective\ batch\ size = 64$ using gradient accumulation.

3.5.2 Results

Table 2 summarizes the test performance results on our financial analyst report subjectivity data. Unsurprisingly, the prompt-based BART models trained on our data outperform OpinionFinder, the classical model designed for general sentiment analysis data, by a large margin. Also consistent with our expectations, the model with enhanced financial vocabulary and auxiliary task training shows a performance gain over the original model.

[Insert Table 2 here]

For comparison, we have also tested a version of FinBERT by A. H. Huang et al. (2022) that is a BERT model pretrained on financial texts including corporate filings, financial analyst report, and earnings call transcripts. We finetune and test the FinBERT model on our data for subjectivity. The results show that our model with enhanced financial vocabulary perform in par with the FinBERT model which takes computationally expensive pretraining on large finance-specific texts. This proves that our methods of domain adaptation through dictionary enhancement can effectively capture domain specificity with orders of magnitude less computational cost.

We apply our finetuned model to classify a sample of 100K sentences randomly sampled from around one million analyst reports. Among them 41 554 are predicted as subjective

[22] The computation is supported by Calcul Québec Narval (https://docs.alliancecan.ca/wiki/Narval/en) and the Digital Research Alliance of Canada (alliancecan.ca).

and 58 446 as objective. Figure 1 shows the different keywords for subjective versus objective sentences. We can see that those sentences predicted as subjective talk more about aspects such as risk and growth that have more room for discretionary judgement, and use more adjectives for evaluation, by contrast those objective sentences involve more factual terms like revenue, eps, price, etc.

[Insert Figure 1 here]

3.5 Conclusion

In this paper, we introduce a prompt-based transformer model that is specifically designed to measure financial text subjectivity. We carefully design a labeling scheme and create a high-quality dataset for subjectivity in analyst reports, a typical financial document genre. We leverage professional financial terminology dictionaries to enhance the model's ability to understand financial vocabulary and semantics. The results of our experiments show that large PLM models with prompting methods can very effectively perform subjectivity analysis in the context of finance. On top of that, our methods to enhance the model with finance-specific vocabulary improve the performance even further. Potential future work includes experimenting with our methods on different types of financial texts, including informal texts such as social media posts. Also, the idea of enhancing PLM with existing domain terminology dictionaries could be generalized to other specialized domains beyond finance.

List of Tables

List of Figures

Figure 1. Comparative word-cloud of subjective (in red) vs. objective (in blue) keywords.

References

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., & McDermott, M. (2019, June). Publicly Available Clinical BERT Embeddings.*Proceedings of the 2nd Clinical Natural Language Processing Workshop* Minneapolis, Minnesota, USA.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, *59*(3), 1259-1294.

Bjerva, J., Bhutani, N., Golshan, B., Tan, W.-C., & Augenstein, I. (2020, November). SubjQA: A Dataset for Subjectivity and Review Comprehension. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online.

De Franco, G., Hope, O.-K., Vyas, D., & Zhou, Y. (2015). Analyst Report Readability. *Contemporary Accounting Research*, *32*(1), 76-104.

Dehkharghani, R., Yanikoglu, B., Tapucu, D., & Saygin, Y. (2012, 10-10 Dec. 2012). Adaptation and Use of Subjectivity Lexicons for Domain Dependent Sentiment Classification. 2012 IEEE 12th International Conference on Data Mining Workshops, Brussels.

Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, *64*(2-3), 221-245.

Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2017). Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach. *Management Science*, *64*(6), 2833-2855.

Huang, A. H., Wang, H., & Yang, Y. (2022). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research*.

Huang, A. H., Zang, A. Y., & Zheng, R. (2014). Evidence on the Information Content of Text in Analyst Reports. *The Accounting Review*, *89*(6), 2151-2180.

Huang, S., Tan, H., Wang, X., & Yu, C. (2022). Valuation uncertainty and analysts' use of DCF models. *Review of Accounting Studies*.

Karamibekr, M., & Ghorbani, A. A. (2013, 17-20 Nov. 2013). Sentence Subjectivity Analysis in Social Domains. 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Atlanta.

Kogan, S., & Meursault, V. (2021). Corporate Disclosure: Facts or Opinions? *SSRN*.

Kothari, S. P., Li, X., & Short, J. E. (2009). The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis. *The Accounting Review*, *84*(5), 1639-1670.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020, July). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online.

Li, B., Liu, Y., Ram, A., Garcia, E. V., & Agichtein, E. (2008). Exploring question subjectivity prediction in community QA. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore.

Liu, P., Tan, H., & Zheng, Y. (2022). *Does Opinion Pay off? Evidence from Analyst Report Subjectivity Using Machine Learning*. Working paper.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. arXiv e-prints, arXiv:2107.13586.

Lo, K., Ramos, F., & Rogo, R. (2017). Earnings management and annual report readability. *Journal of Accounting and Economics*, *63*(1), 1-25.

Loughran, T., McDonald, B., & Yun, H. (2009). A Wolf in Sheep's Clothing: The Use of Ethics-Related Terms in 10-K Reports. *Journal of Business Ethics*, *89*(1), 39-49.

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16-32.

Poerner, N., Waltinger, U., & Schütze, H. (2020, November). Inexpensive Domain Adaptation of Pretrained Language Models: Case Studies on Biomedical NER and Covid-19 QA.*Findings of the Association for Computational Linguistics: EMNLP 2020* Online.

Riloff, E., & Wiebe, J. (2003). *Learning extraction patterns for subjective expressions* Proceedings of the 2003 conference on Empirical methods in natural language processing,

Sachidananda, V., Kessler, J., & Lai, Y.-A. (2021). Efficient Domain Adaptation of Language Models via Adaptive Tokenization. Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing, Virtual.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, *62*(3), 1139-1168.

Wiebe, J., Akkaya, C., Conrad, A., Choi, Y., Hoffmann, P., Ihrig, C., Kessler, J., Somasundaran, S., Wilson, T., Riloff, E., Patwardhan, S., Cardie, C., Breck, E., & Choi, Y. (2011). *Documentation for OpinionFinder 2*. https://mpqa.cs.pitt.edu/opinionfinder/opinionfinder_2/opinionfinder_2_0/opinionfinder_2_0_README.txt

Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. International conference on intelligent text processing and computational linguistics, Berlin.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning Subjective Language. *Computational Linguistics*, *30*(3), 277-308.

Williams, A., Nangia, N., & Bowman, S. (2018, June). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference.*Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). *OpinionFinder: A system for subjectivity analysis* Proceedings of HLT/EMNLP Demonstration Abstracts, Vancouver.

Zhang, R., Reddy, R. G., Sultan, M. A., Castelli, V., Ferritto, A., Florian, R., Kayi, E. S., Roukos, S., Sil, A., & Ward, T. (2020). Multi-Stage Pre-training for Low-Resource Domain Adaptation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online.

General Conclusion

This dissertation tackles the challenge of methodically extracting valuable, actionable information from unstructured financial textual data, by advancing both methodological tools and the empirical significance of their application.

The first essay presents a financial sentiment analysis model pretrained with finance-specific texts, and apply it to explore the connection between social media activities and the cryptocurrency market. The second essay further refines financial sentiment measurement by developing a model adept at measuring fine-grained sentiment towards targeted entities in the text, and validates its superior performance while studying the impact of social media sentiment on stocks. The third essay addresses the under-explored area of subjectivity analysis in financial NLP by devising a prompt-based transformer model that enriched with financial vocabulary and semantics, enabling the model to achieve high performance.

Potential future work could involve refining and enhancing the existing models to capture more nuanced aspects of financial texts, extending beyond sentiment and subjectivity to elements such as uncertainty, risk perception, and stress perception, etc. By incorporating these additional dimensions, researchers can develop a more comprehensive understanding of financial texts and their implications on market behavior. Furthermore, exploring additional information extraction such as financial event detection and categorization, entity relationship identification, etc., could prove to be valuable research directions in advancing the field of financial NLP. Another emerging topic is to explore the implications of today's rapidly evolving super large language models such as ChatGPT, within the field on financial NLP. It is important to understand the boundaries of their capabilities, their reliability, and transparency, in the context of financial domain. There could also be straightforward ways to leverage these models to facilitate current financial NLP efforts. For instance, super large models could serve as cost-effective alternatives or supplements to human labeling for many tasks, as they have demonstrated the ability to provide fairly accurate annotations for tasks that do not require highly specialized, domain-specific expertise.