RESEARCH ARTICLE

JASIST | WILEY

SEntFiN 1.0: Entity-aware sentiment analysis for financial news

Ankur Sinha[1] | Satishwar Kedas[1] | Rishu Kumar[1] | Pekka Malo[2]

[1]Production and Quantitative Methods, IIM, Ahmedabad, India

[2]Department of Information and Service Economy, Alto University, Espoo, Finland

Correspondence
Satishwar Kedas, Production and Quantitative Methods, IIM, Ahmedabad, India.
Email: p15satishwark@iima.ac.in

Abstract

Fine-grained financial sentiment analysis on news headlines is a challenging task requiring human-annotated datasets to achieve high performance. Limited studies have tried to address the sentiment extraction task in a setting where multiple entities are present in a news headline. In an effort to further research in this area, we make publicly available SEntFiN 1.0, a human-annotated dataset of 10,753 news headlines with entity-sentiment annotations, of which 2,847 headlines contain multiple entities, often with conflicting sentiments. We augment our dataset with a database of over 1,000 financial entities and their various representations in news media amounting to over 5,000 phrases. We propose a framework that enables the extraction of entity-relevant sentiments using a feature-based approach rather than an expression-based approach. For sentiment extraction, we utilize 12 different learning schemes utilizing lexicon-based and pretrained sentence representations and five classification approaches. Our experiments indicate that lexicon-based N-gram ensembles are above par with pretrained word embedding schemes such as GloVe. Overall, RoBERTa and finBERT (domain-specific BERT) achieve the highest average accuracy of 94.29% and F1-score of 93.27%. Further, using over 210,000 entity-sentiment predictions, we validate the economic effect of sentiments on aggregate market movements over a long duration.

1 | INTRODUCTION

Financial sentiment analysis (FSA) addresses the problem of *automatically* extracting sentiments from financial text such as documents, stock message boards, news headlines, and so on (Kearney & Liu, 2014) with minimal human intervention. Using FSA systems, market participants seek to gain edge over competitors toward profitable investment or trade decisions. Of significant interest is the sentiment analysis of financial news flow which is a timely and reliable source of information on market events. The news flow presents novel information to the financial market participants, *for ex.* enabling the traders to conduct company or industry directed trades (Von Beschwitz et al., 2015) and investors to actively manage portfolios (Seo et al., 2004). Over the last decade, FSA has also attracted significant research attention driven by the abundance of data, the utility of sentiment information, and the nuances of sentiment bearing expressions (Das & Chen, 2001; Malo et al., 2014; Van de Kauter, Desmet, & Hoste, 2015). In our study, we focus on fine-grained FSA, in which the task is to extract sentiments for the identified entities in the financial text (Schouten & Frasincar, 2015). Our sentiment classification framework is directed toward the extraction of sentiment toward the identified entities in the news text.

Fine-grained sentiment analysis or aspect-based sentiment analysis (ABSA) entails two different tasks—the recognition of entities (also aspects or topics) and the extraction of sentiments corresponding to the identified

entities. Researchers in fine-grained FSA often focus exclusively on the sentiment extraction task, often using named entity recognition (NER) systems for entity-recognition or utilizing entity-annotated datasets. In the context of financial news flow, entities are often the companies or organizations toward which the news text expresses sentiment. Research in financial sentiment extraction task has led to three streams of literature—(a) sentiment lexicon (Loughran & McDonald, 2011; Malo et al., 2013; Wilson et al., 2005), (b) sentiment-annotated datasets such as the JRC Corpus (Balahur et al., 2013), Financial PhraseBank (Malo et al., 2014), SemEval 2017 Task 5 SubTask 2 Headlines Dataset (Cortis et al., 2017), and Financial Opinion Mining and Question Answering (FiQA) 2018 Task 1 Sentiment Scoring Dataset (Maia et al., 2018), and (c) annotation schemes such as the polarity expressions (Van de Kauter, Breesch, & Hoste, 2015; Van de Kauter, Desmet, & Hoste, 2015) and MPQA Opinion Corpus (Deng & Wiebe, 2015). With the advent of language models based on deep learning architectures the focus has shifted from dictionary-based approaches to dataset-driven approaches.

Most of the financial news datasets, unfortunately, are only effective in cases where there is a single entity in the news headlines. FSA systems trained on these datasets fail to achieve high performance in the presence of multiple entities especially with conflicting sentiments in a news headline. Considering the limitations of the existing datasets, through our work, we make publicly available SEntFiN 1.0, a human-annotated financial news dataset containing 10,753 news headlines annotated for the entities present and their corresponding classes (positive, negative, and neutral, hereafter referred to as sentiments) with an average sentence length of 9.91 words. The dataset contains 2,847 headlines with at least two entities and 1,233 headlines containing conflicting sentiments. To our best knowledge, SEntFiN 1.0 is the largest publicly available financial news dataset annotated for multiple entities and their corresponding sentiments.

Sentiment analysis based on news headlines is a difficult task. News headlines contain high signal content and have a low risk of spurious information; however, they often have short text span. If multiple entities are mentioned in the news headlines, the syntactic structure capturing the interactions between the entities and their sentiment bearing expressions is required for effective sentiment extraction. The machine-learning based NER systems currently available for public use such as the Stanford CoreNLP NER (Manning et al., 2014) are trained on multiple corpora—the CoNLL 2003 (Sang & De Meulder, 2003), MUC-6 (Grishman & Sundheim, 1996), MUC-7 (Chinchor & Robinson, 1997). As such, they suffer from two primary limitations—first, the

models are probabilistic in nature and therefore the errors in entity recognition can lead to wrong economic decisions, and second, the training data suffers from temporal and geographical limitations and therefore may not support extensive economic analysis. The annotation schemes such as polarity expressions are limited in their scalability and the NER system (Stanford CoreNLP NER) indicated multiple errors in recognizing entities and entity spans as discussed in section 3.2. Therefore, to conduct our experiments on identifying economic value of news-based sentiments, we developed an entity database covering 1,009 financial entities (companies, sectors) and their various forms of representations in news media amounting to a total of 5,070 phrases. Using this entity database, we were able to recognize the entities mentioned in the news headline without errors. We address the task of recognizing relevant entities using a method of annotating entities with two features—**Target** and **Other** resulting in multiple instances of the same sentence with different feature annotations. In each instance, the entity for which the sentiment needs to be extracted is replaced with **Target** and the other entities are replaced with **Other** as described in Figure 1. With these features, we utilize the sentence as it is, and avoid the use of dependency parsing, part-of-speech tagging, polarity expression, and other methods used in prior studies (Ding et al., 2014; Malo et al., 2014; Van de Kauter, Desmet, & Hoste, 2015). Our entity database adds value by providing a list of financial entities and their entity spans as officially recognized, and indicates the requirement of developing specialized NER systems for use in finance.

In the SemEval 2017 SubTask 4 on sentiment analysis in Twitter, systems that utilized deep learning based architectures such as recurrent neural networks (RNN; Rumelhart et al., 1986) and long-short term memory (LSTM; Hochreiter & Schmidhuber, 1997; Graves & Schmidhuber, 2005) with pretrained word-embeddings such as GloVe (Pennington et al., 2014) were placed among the top performing systems (Cabanski et al., 2017; Ghosal et al., 2017; Mansar et al., 2017; Moore & Rayson, 2017). Further, recent research utilizing pretrained context-based representations of text such as the bidirectional encoder representations from transformers (BERT; Devlin et al., 2018) fine-tuned to the financial domain using Financial PhraseBank (Malo et al., 2014) outperform previously best performing models in sentiment analysis (Araci, 2019; Hiew et al., 2019). In this work, we utilize three models based on BERT–finBERT (Araci, 2019), DistilBERT (Sanh et al., 2019), and RoBERTa (Liu et al., 2019). Overall, we evaluate 12 different learning schemes and the results are reported in section 4. As baselines, we utilize the linearized phrase structures (LPS) representations (Malo et al., 2014) with support
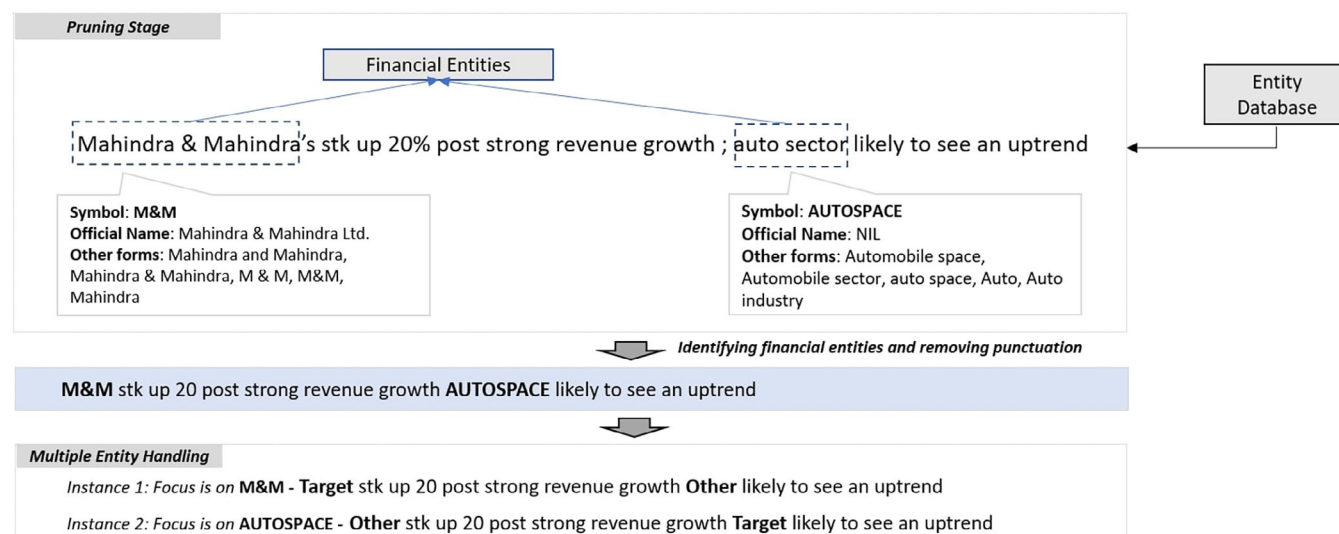
FIGURE 1 The news headline is first processed to recognize the relevant entities using the entity database. The entity recognition step is performed prior to the preprocessing step. Preprocessing step includes the removal of punctuation and special characters. Handling of multiple entities using two features—Target and Other, with the entity in focus replaced with the Target feature and the entities not in focus replaced with the Other feature. Each news headline can lead to multiple instances of text each with a different Target–Other feature annotation

vector machines (SVM) and gradient boosting machines (GBM). Our experiments indicate that RoBERTa achieves highest average accuracy of 94.29% and finBERT, a domain-specific BERT model achieves the highest average F1-score of 93.27%. Lexicon-based N-gram ensemble representations with multi-layer perceptrons (MLP) can achieve above par performance of GloVe-based representations with deep neural architectures. To further the claim that the entity agnostic sentiment analysis is not sufficient for sentiment extraction, we provide a comparison of the proposed entity-based sentiment extraction method against the nonentity-based sentiment extraction approaches, like, NLTK Vader and HuggingFace Sentiment system with Twitter-roBERTa-base model in section 4.4.

We utilize the best-performing sentiment extraction model identified through our experiments to extract the sentiments from over 470,000 news headlines from "The Economics Times" on a daily basis from the years 2002–2017.[1] We construct an aggregate market sentiment index in correspondence to the broad-based market equity index NSE 500, and study the predictive value contained in the sentiments. In the duration 2012–2017, our experiments indicate that the sentiments of the news flow generated in the after-market hours, have an impact on the after-market log price returns and percentage price returns of the NSE 500 index. We observe that the relationship holds significant across multiple years. Our study highlights the sentiments as potential underlying mechanisms through which market participants exchange and act upon information

related to market events. We conclude our study with a discussion on the potential future research directions. The study also holds implications for commercial news-based sentiment analysis solution providers such as Bloomberg, Thomson Reuters, and RavenPack whose sentiment analysis systems are designed for high speed, high volume, large variety, and best accuracy. At high speeds, due to the large volumes of news flow on a large basket of financial entities, it is beneficial to operate with short news text for the reasons of scalability and efficiency, and our proposed and validated framework serves as a robust approach for extracting entity-specific sentiments from news.

To summarize, our contributions are threefold:

- We build and release the SEntFiN 1.0, the largest human-annotated financial news dataset, annotated for multiple entities in the news headlines and their corresponding sentiments, containing 10,753 news headlines with 14,404 entity-sentiment annotations. We also provide an entity database of over 1,000 financial entities and their recognized entity spans in news media aiding the entity recognition task.[2]
- We conduct multiple experiments comparing 12 learning schemes based on two different forms of sentence representations—lexicon-based versus pretrained. Our findings indicate that RoBERTa outperforms all other learning schemes by a wide margin and that certain lexicon-based approaches achieve a comparative performance to deep-learning approaches.

- Our experiment on sentiments accumulated during the after-market hours and their effects on opening market prices, confirms the hypothesis that news sentiments hold economic value, and the relation holds over the long-term. We also provide access to over 470,000 news headlines with timestamps, entities, and sentiments pertaining to the Indian Economy.

The rest of the paper is structured as follows. In section 2, we discuss the related work on the annotated financial news datasets, the financial sentiment lexicon, and the applications of deep learning approaches and lexicon-based approaches for sentiment extraction. In section 3, we introduce SEntFiN 1.0, the sentiment-entity annotated financial news dataset that we release through this paper, and use for our experiments and discuss the annotation process, interannotator agreement, and dataset statistics. Followed by the discussion on the learning schemes, experiments, and results of our modeling approaches in section 4. In section 5, we discuss the construction of our experiment and the evaluation of hypothesis on the economic value of sentiments. We conclude our paper in section 6, where we discuss the potential directions of future research.

2 | RELATED WORK

2.1 | Financial sentiment analysis

The first decade of the 21st century marked the advent of the Information age driven by the adoption of Internet as the information exchange medium. Increasingly research has been directed toward tracking financial and economic developments in real time through automated methods utilizing news flow and machine learning techniques. Wysocki (1998) was the first to explore the relation between daily stock message posting volumes, and future stock returns and trading volumes. His observations on 3,000 stocks listed on Yahoo! message boards indicated that the over-night message-posting volumes affect the next day abnormal stock returns and trading volumes.

Consequently, computational linguistics techniques were applied to opinions on various stocks discussed on the stock message boards—Yahoo! Finance (Antweiler & Frank, 2004; Das & Chen, 2001) and RagingBull.com (Antweiler & Frank, 2004) to identify whether the messages could be translated to financial sentiments—buy, sell, or hold decisions. Das and Chen (2001) were among the first to utilize natural language processing (NLP) techniques and human annotated datasets along with various machine learning based classification techniques

to automate the information extraction process. Their work also revealed the significance of negation effects on the overall sentiment of the sentence, and the ambiguity inherent in human classification of financial news. Based on the analysis of Morgan Stanley Technology (MSH) index, they concluded that there exists a strong link between aggregate sentiments and market movements. Antweiler and Frank (2004) found evidence that stock messages helped predict market volatility based on a bullishness index derived from more than 1.5 million stock messages about 45 companies in Dow Jones Industrial Average and Dow Jones Internet Index.

While the previous research focused on opinions on stock message boards, Tetlock (2007) hypothesized that news columns from major business media firms might contain information that can have significant impact over future company returns and earnings. Based on news data from Wall Street Journal and Dow Jones News Service from 1980 to 2004, and the word polarities (positive, negative and neutral) defined in the General Inquirer (GI) (Stone et al., 1966), he concluded that quantification of financial news presents novel information that can be utilized along with traditional financial information to predict company earnings and stock returns suggesting that the automated news analysis should provide valuable information for traders and analysts. The academic and industrial interest in analyzing news flow and financial text increased after the Global Financial Crisis of 2007–2008, which led to the development of the specialized field of fine-grained or aspect-based FSA.

As the sentiment analysis field grew researchers identified that the sentiments are often expressed toward certain aspects, topics or entities which led to the development of the area of aspect-based sentiment analysis or fine-grained sentiment analysis (Schouten & Frasincar, 2015). The aspect-based sentiment analysis (ABSA) corresponds to two tasks—entity (aspect or topic) recognition and subsequently, sentiment extraction for the entity. The entity recognition task led to the development of named entity recognition (NER) systems which are typically conditional random fields models which learn the boundary spans of named entities such as locations, organizations, and companies from given training data (Lafferty et al., 2001). Currently, the state-of-the-art publicly available NER systems are trained on three datasets—the CoNLL 2003 (Sang & De Meulder, 2003), MUC-6 (Grishman & Sundheim, 1996), and MUC-7 (Chinchor & Robinson, 1997); however, there are no specialized NER systems publicly available for use in finance. The problem of aspect based sentiment analysis has often been formulated as "given an entity, extract the sentiment corresponding to the entity" leading to first SemEval 2016 Task 5 on ABSA which introduced 39 datasets across eight languages (Pontiki et al., 2016). In finance, SemEval

2017 Task 5 subtask 2 and FiQA 2018 Task 1 corresponded to the fine-grained FSA on financial news with training datasets of 1,647 and 1,750 news texts annotated for entities and their sentiment scores. The FSA research community has focused on the sentiment extraction task, leading to developments in financial lexicon and annotated corpora, and machine learning based approaches for sentiment extraction. We provide a review of the two aspects of fine-grained FSA below.

2.2 | Financial lexicon and datasets

Lexicon-based methods for sentiment analysis utilize sentiment lexicon which contain expressions annotated for their sentiments. These expressions may range from words called opinion words (Ding et al., 2008) to polarity expressions which can be both subjective and objective expressions (Van de Kauter, Breesch, & Hoste, 2015). Generally, the lexicon-based methods extract the sentiment from the text using the semantic orientation of the constituting words or expressions (Taboada et al., 2011). The lexicon-based methods utilize dictionaries of these opinion words or polarity expressions annotated for their sentiment scores or semantic orientations (see also Pang et al., 2008; Tong, 2001). There exist multiple sentiment lexicon for general sentiment analysis—Harvard General Inquirer (Stone et al., 1966), MPQA (Wiebe et al., 2005), and SentiWordNet (Baccianella et al., 2010; Esuli & Sebastiani, 2006).

Loughran and McDonald (2011) were the first to study the effects of domain-dependence on prior polarities of words, especially in the financial domain in which the words have interpretations that differ from their general usage. After a detailed study of the Harvard Dictionary, they provide a finance-focused word dictionary with six classifications (positive, negative, uncertain, litigious, strong modal and weak modal). Wilson et al. (2005) noted that the contextual polarities of words in a sentence may not be the same as their prior polarities. They proposed various word-level, sentence-level, and document-level features along with machine learning techniques to perform a phrase-level sentiment analysis.

Building on this research, Malo et al. (2013) identified that representation of financial concepts as direction-dependent phrases (ex. profit as "Positive-if-up," loss as "Negative-if-up," etc.) is essential to capture the sentiment expressed in the financial text. To illustrate, the financial concept—"profit" expresses neutral sentiment in "The profit stands at $ 100 million," and a positive sentiment in "The profit increased by 10%." The directional word "increased" when combined with the word "profit" makes the sentiment positive in the second sentence.

Recognizing that there are multiple ways in which sentiments are expressed in financial text, Van de Kauter, Desmet, and Hoste (2015) developed an annotation scheme to differentiate between two sentiment-bearing expressions—the private state expressions which are explicit expressions (ex. *dissapointing performance*) and the polarity fact expressions which are objective expressions that express sentiment (ex. *facing action from US federal*). Based on their annotation scheme, they provide a bi-lingual annotated corpus of financial text with 4,790 sentences annotated for sentiment bearing expressions and their sentiments and corresponding target entities.

Balahur et al. (2013) released the JRC Corpus consisting of 1,592 quotes from newspaper articles in English for opinion mining. Malo et al. (2014) released a collection of around 5,000 phrases/sentences manually annotated by 16 subject experts. These were further organized into sets of news headlines with varying interannotator agreement. However, these sentences are not annotated for the entity name, making them unusable for entity-aware sentiment extraction. Two other publicly available datasets are from the SemEval 2017 Task 5 Subtask 2 (Cortis et al., 2017) and FiQA Task 1 Sentiment Scoring Dataset (Maia et al., 2018) both of which provide a set of 1,000–2,000 news headlines annotated for multiple entities and their sentiment scores ranging between −1 and 1. SemEval 2017 Task 5 Subtask 2 and FiQA Sentiment Scoring dataset offers annotations for multiple entities and sentiment scores including conflicting sentiments; however, the dataset sizes are small and contain limited set of entities. Van de Kauter, Breesch, and Hoste (2015) provide an annotated corpus of company-specific news for four companies, annotating each sentence for its polarity. Their focus however remains on the suitability of their annotation scheme to the financial context. Though they consider news headlines with multiple companies, they provide the annotations of sentiment-bearing expressions and identify the relevant entities. In SEntFiN, we provide sentiment annotations toward the entities in the news headlines and we do not annotate for sentiment-bearing expressions.

2.3 | Approaches for sentiment extraction

Research in financial news analysis, especially in the sentiment analysis, started in the 2000s with increasing interest in quantifying and learning the mechanisms through which investor sentiments affect stock movements. Traditional approaches to modeling information from news flow include the count of news items in a certain duration (Goonatilake & Herath, 2007; Tetlock,

2007), weighted term vector models (Schumaker & Chen, 2008, 2009; Wang et al., 2012), word-based and sentence-based sentiment analysis (Li, Wang, et al., 2014) and syntax analysis through recognition of actors and their roles (Ding et al., 2014, 2015). Sentiment analysis-based models have shown to perform better than news count and weighted term vector models in explaining and predicting market movements (Li, Xie, et al., 2014). Over time, lexicon-based methods and phrase-level sentiment analysis (Bruce & Wiebe, 1999; Hu & Liu, 2004; Riloff & Wiebe, 2003) which assume that the words have prior and contextual polarities that can be quantified have shown to exhibit better performance than word-based approaches.

Malo et al. (2014) study the problem of learning context through the use of lexicon and propose the use of a linearized phrase structure (LPS) model to learn the syntactic structure which performed better than previous baselines. Van de Kauter, Breesch, and Hoste (2015) are among the first ones to consider the use of annotation tools to identify the polarity bearing expressions, which they term as fine-grained sentiment analysis and show that it performs better than sentence-level sentiment analysis. However, these methods do not address the task of extracting entity-relevant sentiments in the presence of multiple entities. Pivovarova et al. (2018) study the problem of extracting entity relevant sentiments; however, their model fails to perform well in the presence of multiple entities with conflicting sentiments. Our approach of tagging entities based on relevance is able to achieve high level of accuracy in extracting entity-relevant sentiments.

As classification models, SVMs were predominantly used in research (Antweiler & Frank, 2004; Das & Chen, 2001; Li, Xie, et al., 2014; Ranco et al., 2015) until the SemEval 2017 Task 5 Subtask 2—the fine-grained sentiment analysis of news headlines, in which some of the best performing systems utilized RNNs and LSTMs (Cabanski et al., 2017; Ghosal et al., 2017; Mansar et al., 2017; Moore & Rayson, 2017). With introduction of BERT (Devlin et al., 2018), we see an uptick in use of language models for sentiment analysis (Gao et al., 2019; Sun et al., 2019; Xu et al., 2019).

3 | DATASETS

In this section, we provide an overview of the two key datasets that we release with this paper. The first dataset that we discuss is our human-annotated financial news dataset with over 14,000 sentiment-entity annotations. We provide a comprehensive overview of the annotation process, discuss the interannotator agreement, and provide the dataset statistics. The second dataset is the entity database covering over 1,000 entities and their news representations. The entity database is particularly useful in our work, as our final goal is to evaluate economic value of sentiments for which we are required to map sentiments to entities and construct sentiment indices.

3.1 | SEntFiN 1.0

3.1.1 | Annotation process

In total, 10,753 headlines were sampled from all the financial and economic news pertaining to the duration 2011–2015 from Indian business news provider The Economics Times.[3] The annotation scheme and guidelines were prepared by one of the primary authors, who also acted as the curator of the annotation process. For the annotation process, we identified three human annotators—students pursuing a master's degree in business administration and management and with a strong academic performance in courses relevant to financial markets and macroeconomics. These three human annotators independently approached the curator with their interest in the annotation process, and the curator ensured that the annotators were not informed of the presence of the other annotators until the preparation of the final dataset. Each of the three annotators were provided with the 10,753 news headlines with the entities and their entity spans identified in each news headline by the curator (and validated by the other authors). The annotators were asked to assign a sentiment class—positive, negative, or neutral to the entities identified in the news headlines. For the annotation process, the news headlines were provided to the annotators as a *csv* file with two columns—"news headline," "entity-sentiment" (ex. {"news headline": *Negative on Chambal, Advanta: Mitesh Thacker*, "entity-sentiment": {'Chambal': '', 'Advanta': ''}}).

Further, the following guidelines were issued to the annotators prior to the annotation process:

- The annotators were asked to think like investors and avoid any speculation based on their prior knowledge. The financial sentiments were to be derived from the information that was explicitly available in the headline.
- The classes were to be chosen from among one of the three sentiments defined below:

1. Positive: A positive sentiment is expressed by the phrase or subphrase in relevance to the recognized entity.
2. Negative: A negative sentiment is expressed by the phrase or subphrase in relevance to the recognized entity.

3. Neutral: A neutral sentiment or no sentiment is expressed by the phrase or subphrase in relevance to recognized entity or the entity refers to a regulatory organization, a government body, etc. for which the concept of a financial sentiment does not apply.

Each annotator was provided with a sample of 35 news headlines annotated for the entities and sentiments by one of the primary authors, who also acted as the curator of the annotation process. The annotation process was held over a span of 2 months, with intermittent interventions by the curator to track the progress and identify potential errors. After the completion of the annotation process by individual annotators, the final dataset preparation process proceeded based on a consensus-driven mechanism as we intended to prepare a single gold-standard. The final decision on the sentiment annotation was driven by a consensus between the three annotators and in the situation where a consensus was not achieved, the final decision on the sentiment annotation was provided by the curator. However, such cases were less than 2% of the final annotations.

Overall, each news headline was processed for entities and relevant financial sentiments by three human annotators. Each headline was therefore annotated for the sentiments for identified entities, amounting to a total of 14,404 entities and sentiment annotations.

3.1.2 | Interannotator agreement

Ambiguity is inherent in human decision making, and especially arises when the same dataset is being annotated by multiple researchers. In order to reduce ambiguity during the annotation process, a sample of 100 annotated headlines, and list of 15 typical ambiguous cases with the rationale for the final recommendation was provided to the annotators as a part of training. The cues for the sources and reasons for ambiguity were derived from the observations in Malo et al. (2014). After each news headline was annotated by each annotator, and prior to the consensus and final decision making process, the average interannotator agreements for each pair of classes were calculated. We observed that there is a high agreement when differentiating negative from positive sentiments (98.26%) and negative from neutral sentiments (96.85%). However, as observed in previous studies, there is a lower agreement when differentiating neutral from positive sentiments (80.36%).

3.1.3 | Dataset

The output of the annotation process is the SEntFiN 1.0: Sentiment and Entity annotated Financial News dataset.

The dataset contains 10,753 news headlines annotated for the entities and their relevant financial sentiments. Of these, 2,847 news headlines have more than two entities among which 1,233 news headlines contain conflicting sentiments. The news headlines with multiple entities contribute ~6,500 entity-sentiment annotations, amounting to an average of ~2.3 entities per headline. The average word length of sentences in the entire dataset is 9.91 words and sentences with multiple entities tend to have a higher average word length of 10.39 words. SEntFiN identifies 14,404 entities and their sentiments of which positive, negative, and neutral sentiments are 35.23, 26.48, and 38.29%, respectively, indicating low class imbalance as shown in Table 1.

3.2 | Entity database

During the preparation of the dataset for the annotation process, it was observed that the representations of the financial concepts (ex. "stock" appears as "stk," "stocks," "shares"), and representations of a same company, sector or industry varied with headlines (ex. automotive industry appears as "auto," "auto ind," "auto space"; Larsen & Toubro Ltd. appears "L&T," "Larsen," "Larsen & Toubro," "Larsen and Toubro"). Traditionally, financial entities corresponding to sectors and industries such as "pharmaceutical sector," "automotive sector," "steel industry" are not recognized as named entities by the NER systems. However, news headlines often express sentiments toward sectors or industries. The performance of the Stanford CoreNLP NER on a set of 250,000 news headlines showed multiple errors as indicated in Table 2.

To effectively solve the problem of entity recognition and sentiment attribution, we realized that a database that provides information on such representations are necessary, as they are not available as public resources. We therefore manually studied 920 companies which were included in the NSE500 broad-based market index (represents over 95% free float market capitalization in the Indian equity market) in the duration 2002 to 2017 for which the news headlines were available, and 89 other entities (sectors, commodities, and currencies) that appeared in the news headlines (appearing in the SEntFiN dataset) provided by The Economic Times[4] and Moneycontrol[5] and identified their various media representations amounting to 5,070 phrases. Due to our focus on the companies which were publicly traded and were a part of the NSE500 index the entity database is not an exhaustive list of companies; however, it provides a representative list of the equity market in India for the chosen duration.

Each database entry contains three values—first, the symbol assigned to the entity (stock ticker for a company);

—WILEY | **1321**

TABLE 1   Summary statistics of the SEntFiN 1.0 dataset

			Entity-sentiment annotations			
	Headlines	**Average sentence length**	**Positive**	**Negative**	**Neutral**	**Overall**
Single entity	7,906	9.74	2,837 (35.88%)	2,376 (30.05%)	2,693 (34.06%)	7,906
Multiple entity	2,847	10.39	2,237 (34.43%)	1,438 (22.13%)	2,823 (43.44%)	6,498
Overall	10,753	9.91	5,074 (35.23%)	3,814 (26.48%)	5,516 (38.29%)	14,404

Note: The dataset provides 14,404 entity-sentiment annotations in 10,753 news headlines. The distribution of entity-sentiment annotations indicates low class imbalance. The 2,847 news headlines with multiple entities contain 6,498 entities indicating an average of 2.28 entities for each news headline.

TABLE 2   Major categories of errors identified in the named entity recognition process on 250,000 news headlines using Stanford CoreNLP NER

News headline	Entity recognized	Error
Abbott India Q3 net up 26 pc at Rs. 85.56 crore Abbott India Q1 net up 71% to Rs. 51 crore	Abbott India Q3 Abbott India Q1	Q3 and Q1 indicate quarters – leading to wrong entity span
Hold Abbott India Ltd., target Rs. 4,480.0: Centrum Broking Buy Apollo Hospitals Ltd. with a target of Rs. 1,366: Sandeep Wagle	Hold Abbott India Ltd. Buy Apollo Hospitals Ltd.	Sentiment bearing expressions wrongly recognized as entities
Accumulate Tata Steel, Hindalco on declines: Ashit Suri Accumulate Tata Steel, Shree Cement stocks: JV Capital Services	Accumulate Tata Steel Accumulate Tata Steel	Wrong entity span as well as missing entities
Indian IT to battle IBM, Accenture, Hewitt over $12 bn deals Accumulate Coal India & Bharti Airtel on correction: Sajiv Dhawan After Airtel & Idea, Vodafone announces prepaid free calling plans Spectrum Auction: Bharti Airtel & Idea Cellular outperforms broader market CAG reports on Air India & RIL trigger fears of policy paralysis	Accenture, Hewitt Accumulate Coal India & Bharti Airtel Airtel & Idea Bharti Airtel & Idea Cellular Air India & RIL	Multiple entities recognized as a single entity; in a few cases sentiment bearing expressions also identified as entity
Anand Mahindra & Rajiv Bajaj chasing growth in automobile sector with contrasting strategies	Anand Mahindra & Rajiv Bajaj	Persons recognized as organizations, multiple entities recognized as a single entity
Corporation, Andhra & Dena Bank hot picks on M&A street	Andhra & Dena Bank	Multiple entities recognized as a single entity
Technologies like ArtificiaI Intelligence, big data impacting power sector: Tata Power	ArtificiaI Intelligence	Wrong entity recognition

Note: The errors majorly correspond to wrong entity spans, multiple entities being recognized as single entities, persons recognized as organizations, and missed entities.

second—the official name of the entity if present; and third—the list of all phrases that refer to the entity in news media. For example, the company *State Bank of India* has the following database entry: {"SBIN" - {*Official Name*: State Bank of India Ltd.}, {*Other forms*: State Bank, SBI, State Bank of India}}. We make this database publicly available along with this work. In our future work, we aim to include more entities to the database. The various representations of financial concepts such as "stk,"

"stks," etc. were added to the financial lexicon used in the study.

4 | SENTIMENT EXTRACTION

In this section, we describe our choices of learning schemes which is a combination of sentence representations and classifier methods, and the hyperparameter

choices utilized in our experiments. Further, we present and interpret the results of test performance of all learning schemes.

4.1 | Sentence representations

Each instance of the headline generated after the recognition of entities and removal of noise are then transformed into input feature vectors based on their sentence representations. We categorize our choice of the sentence representations based primarily on the approach taken to extract features which is either through financial lexicon or through pretrained representations. We discuss these further below.

4.1.1 | Lexicon-based representations

We use MPQA (Wiebe et al., 2005), GI (Stone et al., 1966), and LM (Loughran & McDonald, 2011) dictionaries to capture the prior polarities. If an overlap is encountered, the prior polarity as defined by LM dictionary is preferred over other dictionaries. We utilize the financial concepts and directionality words list developed by Malo et al. (2013). Contextual valence shifters (negators) are derived from GI. Numbers are identified and represented by "Number" feature.

We note that lemmatization of words typically results in loss of features majorly pertaining to the prior polarities and directional dependencies. For example, in "Infosys' stock rallies," "rallies" is used to express upward movement in the "Infosys' stock price" indicating a *positive* sentiment toward "Infosys," while in "Nifty rally ends," "rally" describes an action already in place and sentiment for "Nifty" is *neutral*.

After an investigation of the SEntFiN 1.0, a custom dictionary containing 970 words commonly used in the financial news were prepared. Of the 970 words, 895 words are present in the LM dictionary with differing annotations, and the remaining 75 words do not appear in any dictionary. For the purpose of feature extraction, we define Lexicon as including LM, MPQA, GI, Malo et al., and our custom dictionary as shown in Table 3. The feature annotations based on the custom dictionary[6] are given preference over the LM dictionary during the feature extraction step.

Building upon the example in Figure 1, we depict the lexicon-based feature extraction for the first instance in Figure 2. We do not employ stop-word removal, as the number of tokens in the entire dataset is not large to ignore the frequencies of stop-words. The vector of features is then modeled based on one of the two approaches mentioned below.

Linearized phrase structure (LPS) models were introduced by Malo et al. (2014) as a simple and effective technique to capture syntactic structure of news headlines. The LPS model, captures the syntactic structure of the sentence utilizing a sequence of literals framework, where in the literals are the word-features. The final sequence is then converted into a bit vector using a binary coding scheme.

UBT N-gram based models have been among the first techniques to be utilized to capture the probability of co-occurrence of words in text and have been shown to perform reliably well in text classification tasks (Cavnar & Trenkle, 1994). We note that, N-grams in the context of sentiment extraction from news headlines have the ability to capture the syntactic structure and we apply the N-gram model after extracting word-level features, rather than on words itself. We define UBT as the Uni-gram + Bi-gram + Tri-gram ensemble vector of term frequencies based on word-features.

4.1.2 | Pretrained representations

For this study, we have used both context-free representations and context-aware representations, elaborated further below.

Features	LM	MPQA	GI	Malo et al.	Custom	Overall
Positive	408	2,718	55		276	3,457
Neutral	82,319	572	144		21	83,056
Negative	2,404	4,911	82		85	7,482
Up	4		128		233	365
Down			122		213	335
Positive-if-up				69	100	169
Negative-if-up				28	32	60
Negator			31		10	41

TABLE 3 Lexicon has eight types of word-level features and annotations for 94,965 words

Note: In case of an overlap, the preference order is as follows: direction-dependency > directionality > negation > prior sentiment. Lexicon = LM + MPQA + GI + Malo et al. + Custom (our additions and modifications).
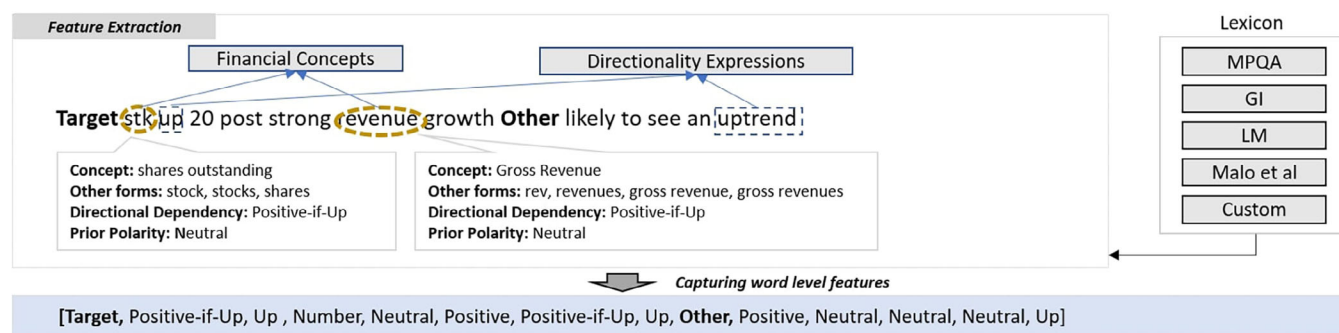
FIGURE 2 The lexicon-based feature extraction step utilizes lexicon to generate a vector of features corresponding to the tokens in the sentence. We do not employ stop-word removal in our preprocessing step

Context-independent representation: We use GloVe (Pennington et al., 2014) for context-independent representation of words. For our study, we use the GloVe word-embeddings dataset with 840 billion tokens, 2.2 million vocabulary and each word/token is represented as vector of 300 dimensions. GloVe embeddings are derived based on the optimization of the word embedding vectors in the word analogy task.

Context-aware representation: Bi-directional encoder representations from transformers (BERT) is a deep bi-directional general language model consisting of stacked layers of a set of transformer encoders. BERT uses either 12 or 16 layers of transformers which generates a context-aware representation for words. The training data used for BERT is 16 GB corpus of books and Wikipedia. For our study, we utilize finBert (Araci, 2019), a financial domain focused BERT implementation which is fine-tuned using Financial PhraseBank (Malo et al., 2014) and FiQA Task 1 sentiment scoring dataset. Further, we compare two other versions of BERT–RoBERTa (Liu et al., 2019) which is an optimized BERT and trained over a 160 GB corpus (144 GB additional corpus than BERT) and DistilBERT (Sanh et al., 2019) which has less model parameters and therefore a smaller version of BERT. We use the source code of finBERT released by Araci (2019). For RoBERTa and DistilBERT, we implement models publicly made available on HuggingFace (Wolf et al., 2019), using *roberta-base-openai-detector* and *distilbert-base-uncased* implementations respectively.[7]

4.2 | Classification algorithms

Support vector machines (SVMs) are a family of kernel-based binary classification methods originally introduced by Cortes and Vapnik (Cortes & Vapnik, 1995; Vapnik, 2013), which maximize the margin defined by the distance between the nearest points (support vectors) to a hyperplane that separates the two classes in the feature space. We implement SVM with a linear kernel using the scikit-learn package (Pedregosa et al., 2011).

Gradient boosting machines (GBMs) are a family of ensemble boosting classifier methods, which formulate the classification problem as a loss minimization problem, and approach the solution using gradient descent approach (Friedman, 2001). The base learners of GBMs are typically classification and regression trees (CARTs) (Breiman, 2017) which recursively partition the feature space and fit simple regression/classification models to each partition. This tree type model is implemented with XgBoost package (Chen & Guestrin, 2016).

Multi-layer perceptrons (MLPs) are feed-forward artificial neural network. Each node in an MLP acts as a nonlinear activation function, except in case of nodes in input layer. In this study we implement MLP with scikit-learn package (Pedregosa et al., 2011).

Recurrent neural networks (RNNs) are successive layers of artificial neurons proposed by Rumelhart et al. (1986). RNNs build over MLPs and add a hidden state which is the output of the previous time step, therefore, are suitable for data with sequential dependencies such as text.

Long-short term memory (LSTM) (Hochreiter & Schmidhuber, 1997) is a version of RNN which aims to learn long-term dependencies of a given sequence and solve the vanishing gradient problem of vanilla RNNs. Bi-directional LSTM (BiLSTMs) (Graves & Schmidhuber, 2005) are similar to BiRNNs in their construction.

We construct uni/bi-directional RNN and LSTM model using the Keras library in Python (Chollet et al., 2018) with Tensorflow backend (Abadi et al., 2016).

4.3 | Experiments

As a first step, we processed the headlines in SEntFiN datasets to handle multiple entities using the features **Target** and **Other**. The processed dataset contains

14,404 headlines with relevant entities identified as **Target** with their corresponding sentiment label, and the irrelevant entities identified as **Other** if present.

For experiments with the lexicon-based and GloVe-based approaches, we divided the processed dataset using a 80/20 train/test split, leading to ~11,200 entity-sentiment annotations in the training set and ~2,800 entity-sentiment annotations in the testing set. We train our learning schemes on a training set, compare the performance on a test set, and evaluate based on Accuracy and F1-score for each sentiment class separately. Given that the distribution of news headlines in the dataset is not uniform with respect to headlines with multiple entities and single entities, we chose to create 31 different randomly sampled train/test datasets instead of a *k*-fold cross-validation approach. We report the results as the median performance across the 31 different train/test splits in the following section. The average and standard deviation of the results are reported in Appendix A1. For the experiments with the BERT-based models, given that the language models are large with over millions of parameters, we utilized a fivefold cross-validation approach with a 80-10-10 train-validation-test split.

Implementation details: We use default parameters for SVC (scikit-learn) and XgBoost Classifier. We constructed the MLP as a 10×10 network with other parameters for the model set to default. The uni/bi-directional layer implementations of LSTM and RNN each have 100 nodes. All four implementations of LSTM and RNN are optimized with *ADAM* (Kingma & Ba, 2014), with a default learning rate of *0.01* and uses sigmoid activation function in the output layer with 10 epochs. A dropout probability of $p = .2$ is applied after the embedding layer and LSTM/RNN layers.

We use finBERT (Araci, 2019) and further fine-tuned the model using our SEntFiN 1.0 dataset for training. For our implementation of finBERT, we use a warm-up proportion of 0.2, a dropout probability of $p = .2$, a learning rate of 2e-5, a mini-batch size of 64, and a maximum sequence length of 30 tokens. For RoBERTa, after a comparison of various hyperparameter choices, we report the results of two models—RoBERTa (A) with a dropout

probability of $p = -.1$, weight decay of 0.02, and a learning rate of 2e-5, and RoBERTa (B) with a dropout probability of $p = .2$, weight decay of 0.01, and a learning rate of 2e-5. For both RoBERTa models, we use a mini-batch size of 64 and a maximum sequence length of 30 tokens. For DistilBERT, after a comparison of various hyperparameter choices, we report the results of two models—DistilBERT (A) with a dropout probability of $p = -.1$, weight decay of 0.02, and a learning rate of 2e-5, and DistilBERT (B) with a dropout probability of $p = .2$, weight decay of 0.01, and a learning rate of 2e-5. We use a five-fold approach with 10 epochs each for RoBERTa and DistilBERT to ensure that a wide combination of data for the train-validate-test split of the dataset is considered.

4.4 | Results and discussion

In Table 4, we present the results of the NLTK Vader and HuggingFace Sentiment Analysis systems on the SEntFiN news headlines dataset. In such kind of nonentity aware systems, a single sentiment is assigned to a headline, and in case of multiple entities within the headline, the same sentiment is assumed for all the entities. Since these systems are not trained for entity-aware sentiment extraction task, the accuracies and F1-scores are clearly quite low.

In Table 5, we present the results of the experiments in which the models have been trained for the entity-aware sentiment extraction task. In case of the lexicon-based sentence representations, we observe that the ensemble UBT approach of capturing features performs better than the previous established benchmark of LPS approach across both SVM and GBM classification methods with prominent improvements across positive and neutral classes. GBM achieves better classification performance on both the lexicon-based representations with 3–5% increments in both Accuracy and F1-score. With the use of MLPs we do not achieve increments in performance as expected and we also observe that UBT + GBM learning scheme performs better than the current UBT + MLP learning scheme. It is however possible

TABLE 4 A comparison of nonentity aware sentiment analysis approaches on the SEntFiN dataset

	Positive		Negative		Neutral	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
NLTK Vader	65.17	64.09	74.02	52.68	39.70	77.61
HuggingFace sentiment	74.90	76.30	82.35	82.02	58.79	60.59

Note: NLTK Vader utilizes a lexicon-based approach with simple count-based features. For HuggingFace Sentiment system, we utilized the Twitter-roBERTa-base model which is trained on ~58 million tweets and finetuned for sentiment analysis with TweetEval benchmark.

TABLE 5 Performance summary of the 12 learning schemes

Approach	Learning scheme	Positive		Negative		Neutral	
		Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
Lexicon	LPS + SVM	79.17	69.58	85.00	71.48	73.69	66.81
	LPS + GBM	82.45	74.72	87.41	76.45	77.41	71.18
	UBT + SVM	83.79	77.24	87.24	75.85	79.38	73.07
	UBT + GBM	85.00	78.73	88.48	77.81	80.52	74.91
	UBT + MLP	82.38	75.58	86.83	74.13	78.24	72.29
Pretrained	GloVe + RNN	83.66	76.28	86.49	74.53	79.03	71.71
	GloVe + Bi-RNN	82.99	75.41	85.83	72.88	78.21	71.20
	GloVe + LSTM	84.73	78.19	87.19	76.16	80.60	73.99
	GloVe + Bi-LSTM	84.78	78.49	86.85	75.69	80.78	74.32
	finBERT	90.58	92.80	93.22	**95.11**	89.45	**91.90**
	DistilBERT (A)	94.34	92.20	94.48	89.20	90.89	88.10
	DistilBERT (B)	94.55	92.40	94.34	89.00	90.82	88.00
	RoBERTa (A)	95.24	93.40	**95.31**	91.00	92.34	89.90
	RoBERTa (B)	**95.38**	**93.60**	95.10	90.50	**92.41**	90.10

Note: The lexicon-based approaches utilize the custom dictionary consisting of 94,965 word-feature annotations. The numbers in bold correspond to the highest performance on a metric in a specific class. RoBERTa (B) achieved highest accuracy in Positive and Neutral classes, while RoBERTa (A) achieved highest accuracy in Negative class.

that with appropriate hyperparameter configuration, MLPs may perform better than GBMs.

In case of the learning schemes based on pretrained representations, the GloVe word embeddings with LSTMs perform better than the RNNs with improvements across both accuracy and F1-score; however, the differences are not pronounced. The BERT based learning schemes outperforms all GloVe based learning schemes with average increments of over 7% in accuracy and 15% in F1-score compared to the best performing GloVe-based learning scheme. Among the BERT based learning schemes, RoBERTa achieves the highest accuracy scores across all the sentiment classes, with above 95% accuracy in positive and negative classes, and ~92% accuracy in neutral class. finBERT achieves the best F1-score across negative and neutral classes, with a marginally lower F1-score compared to RoBERTa-based learning schemes in the positive class.

We observe that UBT + GBM learning scheme achieves comparable performance with GloVe based learning schemes, which is indicative of the potential of N-gram models in short text. The ability to achieve the performance of deep-learning approaches through feature-engineering based on specialized domain knowledge is noteworthy. The training process for generating the GloVe word embeddings includes the use of both the local and global context optimized toward achieving best performance on word analogies (Manning et al., 2014). As such, these unsupervised models learn from large

datasets such as Wikipedia, Twitter, among others where each word appears in multiple contexts with multiple meanings (also known as polysemy). Since, the training process optimizes for analogies across multiple contexts, the embeddings learn a large number of weights. However, we observe that the contextual embeddings fail to perform effectively in a domain specific task as in the case of financial sentiment analysis. Even with large number of weights, the GloVe embeddings likely suffer similar disadvantages of a general sentiment lexicon (Loughran & McDonald, 2011). The high performance of the UBT models are likely to arise from two key reasons. First, UBT models are effective on short text and in the study we utilized UBT on the features rather than on words likely leading to higher contextual performance. Second, the lexicon used in the study is customized to the financial sentiment analysis task. In addition to changes to the existing lexicon, we have added 75 new words to the financial lexicon. We believe that the lexicon provided good word-feature annotations which the UBT model utilized effectively.

Most of the learning schemes report highest performance on *negative* class, lowest performance on *neutral* class and medium performance on *positive* class. This indicates that the models are better able to differentiate between the negative sentiments from positive or neutral sentiments, which has been observed by Malo et al. (2014) as well. However, RoBERTa based models achieve high accuracy scores on positive and neutral classes

unlike other learning schemes. BERT language models utilize a masking strategy in which hidden sections of text are learnt from the datasets, through which the contextual embeddings of tokens are generated. Therefore, fine-tuning BERT to the task-at-hand using the approach as performed in this study provides high performance in a domain-specific task, here financial sentiment analysis. RoBERTa is learnt on much larger dataset than BERT including a novel Common Crawl CC-News dataset (Mackenzie et al., 2020) containing over 44 million news documents. This learning procedure is likely to have produced an improved language model and improved performance in the financial sentiment extraction task.

5 | INFORMATION CONTENT OF NEWS FLOW

The information related to the value of securities traded publicly is reflected in the prices set by the forces of supply and demand on the exchanges that the securities trade on. The market-based pricing mechanism constantly consumes the publicly available information related to the historical performance and the expected future performance of the underlying asset. This mechanism was shown to be profound in emerging markets (Morck et al., 2000) where the stock price variations were not much affected by fundamentals. With respect to stocks, major public announcements such as earnings releases, changes in management, new product launches are bound to have an impact on stock prices, and have been well studied and documented (Chambers & Penman, 1984; Lee & Chen, 2009; Warner et al., 1988). While, these are the major approaches undertaken by companies to release company related information to the public, there exists an unstructured form of news flow which involves opinions, discussions, daily events regarding companies, which may not necessarily involve company's participation. The news released by media houses fall into this category of news flow, where the source (the editor) determines the content pertaining to the reported market events. Often, such information is utilized by High Frequency Traders in extremely short durations (orders of 10^{-9} to 10^{-6} s) to conduct directed trades. However, it is of interest to identify whether such news flow has economic value over longer time horizons.

5.1 | Modeling information from news flow

The most prominent approach to modeling information from news flow previously has been to compute a measure for sentiment in a certain duration based on the number of buy and sell decisions derived from the news headlines (Das & Chen, 2001; Zhang & Skiena, 2010). The sentiment measure is constructed using the following formula.

The sentiment score (s) for a specific entity E, and for a certain duration T, is measured by the following formula:

$$s^1_{T,E} = \frac{\text{Pos}_{T,E} - \text{Neg}_{T,E}}{\text{Pos}_{T,E} + \text{Neg}_{T,E}},$$

where $\text{Pos}_{T,E}$, $\text{Neu}_{T,E}$ are the number of Positive and Negative recommendations.

However, this measure does not account for the news flow that contains neutral sentiment. Neutral sentiments in a certain duration are likely to attenuate the effects of positive and negative sentiments therefore leading to reduced movements in market prices. Given this inverse relationship of price movements and neutral sentiments, we construct an alternative sentiment measure as the difference between the frequencies of buy and sell decisions relative to all the decisions (buy, sell, and hold) generated with respect to an entity in a certain duration. This construction takes into account the probability that the news source generates neutral sentiments, and attenuates the effects of positive and negative sentiments in a neutral environment,

$$s^2_{T,E} = \frac{\text{Pos}_{T,E} - \text{Neg}_{T,E}}{\text{Pos}_{T,E} + \text{Neu}_{T,E} + \text{Neg}_{T,E}},$$

where $\text{Pos}_{T,E}$, $\text{Neu}_{T,E}$, $\text{Neg}_{T,E}$ are the number of Positive, Neutral, and Negative recommendations.

5.2 | Datasets

We utilize two datasets for our experiments, one corresponding to the aggregate market movements, and the other corresponding to the information from news flow, as detailed below.

5.2.1 | NSE 500 index prices

The NSE 500 index is the broad equity market index based on free-float market capitalization tracking more than 95% of the market capitalization at any point of time.[8] The index has a base value of 1,000, established on January 1, 1995. The index is considered for reconstitution semi-annually, and more than 400 companies have

been changed since inception. We collect the daily opening and closing values of index from January 1, 2002 to December 31, 2017.

5.2.2 | News flow

We captured the news headlines along with the timestamp for the duration January 1, 2002 to December 31, 2017 relevant for the NSE 500 index. Overall, we collected 418,703 news headlines relevant to the time duration, which led to 576,191 instances of news text after our entity-recognition step using the entity database.[9] We extracted sentiments for the recognized entities information utilizing the best performing RoBERTa-based model.

In Figure 3, we juxtapose the 30-day moving averages of the daily news-based sentiment index and the NSE 500 index. We can observe that there is an apparent correlation between the long-term movements in both indices, with sentiments movements preceding the market movements. Moreover, we find that the temporal distance between the movements in sentiments and markets, has reduced over time, likely due the technological progress leading to improvements in market efficiency. Nevertheless, there appears to be relationship between the sentiments and market movements, and therefore, we developed an experiment to evaluate the economic value of sentiments.

5.3 | Construction of the experiment

The purpose of the experiment is to identify whether the information derived from news flow has predictive power in terms of the ability to affect price changes/updates. As shown in Figure 4, the National Stock Exchange (NSE) is open for the duration 9:15 a.m. to 8:30 p.m., during which there is a bi-directional information flow between the domestic market and domestic media.[10] However, during the after-market hours, since the market is closed, the information flow is unidirectional from the market events happening on other exchanges to the domestic media, and there are no domestic market events. The novel information accumulated during the after-market period is reflected in the corresponding price update on the opening hours of the next market day. We therefore construct the experiment to measure the predictive power of the news flow during the after-market period. To capture the nature of the relationship over long periods, we considered a 6 year duration from 2012 to 2017, which corresponds to three peaks and four troughs in the 30-day moving average of the Nifty 50 index. Overall, 175,575 news headlines were utilized for the experiment, which led to 212,828 instances of news text after the entity recognition step. Therefore, we utilized over 210,000 entity-sentiment predictions for the experiment.
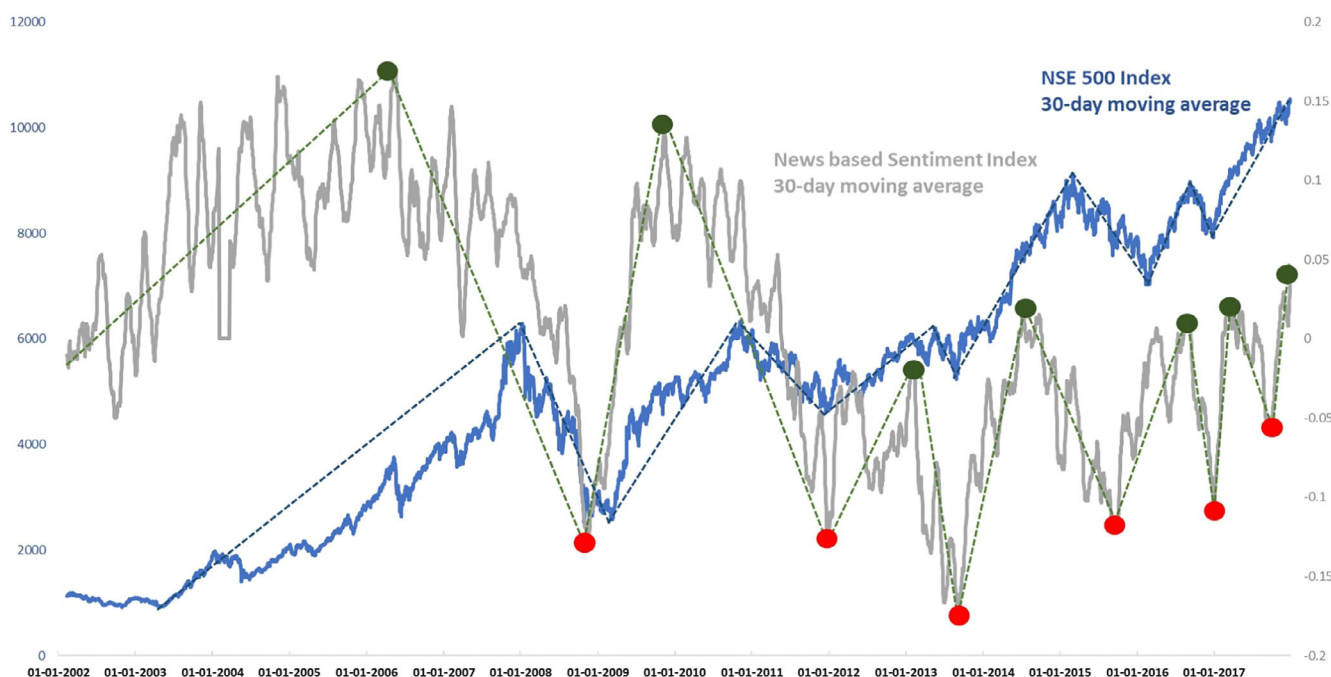

FIGURE 3 A snapshot of the 30-day moving averages of the NSE 500 and daily News-based Sentiment indices over the duration January 1, 2002–December 29, 2017. The graph is derived from the data corresponding to 5,113 market days and 576,191 instances of news headlines. The graph indicates that the movements in the news-based sentiment index and the NSE 500 index correlate with the troughs and crests in the sentiments
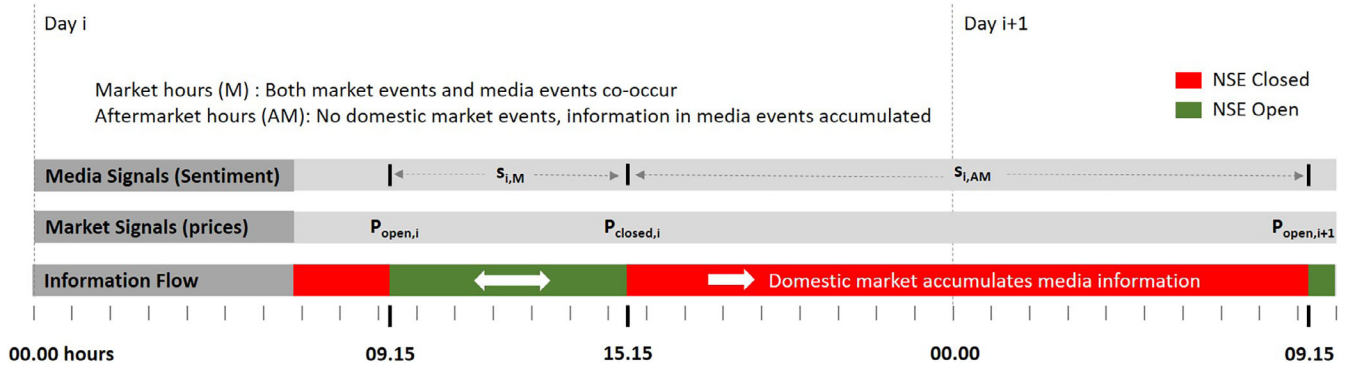
FIGURE 4 The interactions between market and media over the duration of an open market day indicate that the day can be divided into two durations—market hours and after-market hours. In the market hours, there is two-way exchange of information across the market and the news media. However, in the after-market hours, there are no domestic market events and therefore there is only way of information flow which is the media capturing information from other markets and their events. The information captured by the news media in the after-market hours is likely to reflect in the opening prices of the assets the next open market day

5.4 | Variables

1. Sentiment scores

We utilize the best performing RoBERTa-based learning scheme discussed in Section-4.4, to get the time-stamped sentiments for the entities present in the news headlines. We adjust for the changes in the composition of the NSE 500 index, for all years from 2012 to 2017. The final sentiment scores are calculated for two durations

- $s_{iM,\text{NSE500}}$ – market hours (9:30 a.m., ith day to 3:30 p.m., ith day)
- $s_{iAM,\text{NSE500}}$ – after-market hours (3:30 p.m., ith day to 9:30 a.m., $i + 1$th day)

- Daily log price after-market returns

We calculate the daily after-market log price returns for the NSE 500 index, using the opening and closing prices as,

$$d_{i,\text{NSE500}} = \log \frac{P_{\text{open},i+1,\text{NSE500}}}{P_{\text{close},i,\text{NSE500}}},$$

$$d_{i,\text{NSE500}} = \log P_{\text{open},i+1,\text{NSE500}} - \log P_{\text{close},i,\text{NSE500}}.$$

5.5 | Hypothesis

We hypothesize that the sentiment information derived from the news flow in the after-market hours has a causal effect on the daily after-market price returns. We use a linear regression model to test our hypothesis,

$$d_{i,\text{NSE500}} = \alpha + \beta * s_{iAM,\text{NSE500}} + \varepsilon,$$

with $d_{i,\text{NSE500}}$, and $s_{iAM,\text{NSE500}}$ as defined above. The null and alternate hypotheses are as follows:

H_0: On a certain day, the sentiment score of the after-market hours during the previous day has no effect on the log price after-market returns with respect to the previous day ($\beta = 0$).

H_A: On a certain day, the sentiment score of the after-market hours during the previous day has a significant effect on the log price after-market returns with respect to the previous day ($\beta \neq 0$).

5.6 | Results and discussion

The results of the regressions are presented in Table 6. For the primary sentiment measure, we observe that there exists a statistically significant relation between the log price after-market returns and the sentiment for the after-market duration across the years 2014–2017 at a 10% significance level. For the alternate sentiment measure, we observe that there exists a statistically significant relation between the log price after-market returns and the sentiment for the after-market duration across all the years at a 10% significance level. On an average we also observe that the alternate sentiment measure is able to capture the variations in NSE 500 index prices more effectively than the primary sentiment measure. In Appendix A2, we also present an analysis with percentage price returns as the dependent variable.

Overall, we notice that the information derived from unstructured news flow has predictive value over longer time horizons and also that the predictive value holds across multiple years. While there seems to be economic value in the news information, trading based on such

TABLE 6 Results depicting the economic value of sentiments across the years 2012–2017

| Year | # observations | Primary sentiment measure (s^1_{AM}) | | Alternate sentiment measure (s^2_{AM}) | |
		beta	*r*-squared	beta	*r*-squared
	261	0.00065	0.0073	0.00136*	0.0108
2013	260	0.00104	0.01	0.00242**	0.0159
2014	260	0.00242***	0.1093	0.00504***	0.1304
2015	249	0.00105**	0.0166	0.00192**	0.0168
2016	260	0.00181**	0.0175	0.00385**	0.0206
2017	260	0.00076*	0.0145	0.00142*	0.0139

Note: The primary and alternate sentiment measures show significant effects across the years 2014–2017. The alternate sentiment measure takes into account the neutral sentiments and captures more variation than the primary sentiment measure. In doing so, the alternate sentiment measure exhibits significant effects in the years 2012 and 2013.

*$p < .1$; **$p < .05$; ***$p < .01$.

information may not be economically viable if we consider the transaction costs.

In Appendix A3, we also present a vector auto-regression (VAR) based analysis of the daily sentiments and NSE 500 index log price returns across the duration 2012–2017 to understand the lags and leads in the time-series correlations.

6 | CONCLUSION AND FUTURE WORK

Financial sentiment analysis has a rich history of research with specialized dictionaries and datasets. However, most of the studies till date addressed the problem of extracting a single sentiment from the news headlines. One of the key impediments to multiple sentiment extraction is the lack of availability of datasets. To our best knowledge, there are no datasets that provide annotations for multiple entities and sentiments. We therefore make publicly available SEntFiN 1.0 containing 10,753 news headlines with 14,404 entity-sentiment annotations. Using SEntFiN, we address the task of extracting entity-relevant sentiments in a general setting where multiple entities are present in the news headline.

Through our work we provide a comparison of lexicon-based and pretrained approaches for sentence representations. We definitely conclude that deep bidirectional pretrained language models such as BERT fine-tuned to SEntFiN outperforms all other learning schemes by a significant margin. We also note that lexicon-based approaches utilizing domain specific task related knowledge are at and above par with generic deep-learning based approaches. Future work can explore the experimentation with varied feature and model architecture choices to better understand the contribution of features and models. We encourage researchers to experiment

with other model architectures that are gaining prominence such as GRUs, and other pretrained language models such as GPT3. Future work can also explore the usage of lexicon-based sentence representations with deep learning architectures to evaluate explainable AI. In our work, we made a conscious design choice to not utilize methods such as dependency parsing to limit our model complexity and focus on the final goal of evaluating the economic value of sentiments. However, the use of graph-based dependency parsing holds promise for extracting contextually relevant phrases and we consider it to be a potential future research direction.

We utilize the model to extract sentiments from news headlines for over 900 Indian companies and evaluate the economic value of the sentiments at an aggregate market level. Our findings show that the sentiments accumulated in the after-market period have an effect on the market opening prices. We motivate researchers to consider our results and evaluate sentiment-driven trading strategies, and the implications of such a strategy. While we have considered only companies and sectors as entities, we urge future researchers to extend our work and datasets to entities such as commodities like gold and silver (Sinha & Khandait, 2021), currencies, futures, and other financial entities.

Our work is grounded in sentiment analysis of news headlines and as such, our methods are focused on the analysis of short text. In order to build comprehensive financial sentiment analysis systems, we urge researchers to consider the task of financial sentiment analysis on large text such as news articles and press releases. The task will entail the use of methods such as coreference resolution, which is a difficult task given that the current publicly available NER systems are not accurate in the financial context. A dictionary-based approach that we have utilized in the study served the study purpose of evaluating economic value of sentiments. However, such

an approach is not scalable toward a generalized financial sentiment analysis system, with the ever changing nature of financial markets and growing number of financial entities across geographies. We suggest that future researchers consider this research gap and develop financial domain specific language processing systems and libraries. Such financial NER systems can encourage research in other areas of financial text analysis such as question answering, summarization, robo-advisory. Even with such text processing methods in place, it is likely that the annotated corpora for sentiment analyses on large texts might be required due to the structural complexity and changing themes across the text. Topic models combined with entity-aware sentiment analysis approaches are likely to be effective for large financial text.

ACKNOWLEDGMENT

ORCID

Satishwar Kedas https://orcid.org/0000-0002-0032-9346

ENDNOTES

[1] Available upon request from the corresponding author.

[2] The datasets are currently available at https://github.com/pyRis/SEntFiN

[3] https://economictimes.indiatimes.com/

[4] https://economictimes.indiatimes.com/

[5] https://www.moneycontrol.com/

[6] Available upon request from the corresponding author.

[7] https://www.huggingface.co/transformers/pretrained_models.html

[8] NSE India official website.

[9] We cannot release the data publicly due to legal reasons. Available upon request for academic research.

[10] https://www.nseindia.com/resources/exchange-communication-holidays#

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M. & Ghemawat, S. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

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

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

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

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., & Belvaeva, J. (2013). Sentiment analysis in the news. arXiv preprint arXiv: 1309.6202.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Bruce, R. F., & Wiebe, J. M. (1999). Recognizing subjectivity: A case study in manual tagging. *Natural Language Engineering*, 5(2), 187–205.

Cabanski, T., Romberg, J., & Conrad, S. (2017). Hhu at semeval-2017 task 5: Fine-grained sentiment analysis on financial data using machine learning methods. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 832–836). Association for Computational Linguistics.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval* (pp. 161–175).

Chambers, A. E., & Penman, S. H. (1984). Timeliness of reporting and the stock price reaction to earnings announcements. *Journal of Accounting Research*, 22(1), 21–47.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). Association for Computing Machinery.

Chinchor, N., & Robinson, P. (1997). MUC-7 named entity task definition. In *Proceedings of the 7th conference on message understanding* (Vol. 29, pp. 1–21). Association for Computational Linguistics.

Chollet, F. et al. (2018). Keras: The python deep learning library. Astrophysics Source Code Library, ascl-1806.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., & Davis, B. (2017). *Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news*. Association for Computational Linguistics.

Das, S., & Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association annual conference (APFA)* (Vol. 35, p. 43). Asia Pacific Finance Association.

Deng, L., & Wiebe, J. (2015). MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 1323–1328). Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 231–240). Association for Computing Machinery.

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2014). Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 conference on empirical methods*

in natural language processing (EMNLP) (pp. 1415–1425). Association for Computational Linguistics.

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. In Twenty-fourth international joint conference on artificial intelligence. AAAI Press.

Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of the fifth international conference on language resources and evaluation (LREC'06) (Vol. 6, pp. 417–422). European Language Resources Association (ELRA).

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Gao, Z., Feng, A., Song, X., & Wu, X. (2019). Target-dependent sentiment classification with BERT. IEEE Access, 7, 154290–154299.

Ghosal, D., Bhatnagar, S., Akhtar, M. S., Ekbal, A., & Bhattacharyya, P. (2017). IITP at SemEval-2017 task 5: An ensemble of deep learning and feature based models for financial sentiment analysis. In Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017) (pp. 899–903). Association for Computational Linguistics.

Goonatilake, R., & Herath, S. (2007). The volatility of the stock market and news. International Research Journal of Finance and Economics, 3(11), 53–65.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5–6), 602–610.

Grishman, R., & Sundheim, B. (1996). Design of the MUC-6 evaluation (Technical Report). Deptartment of Computer Science, New York University.

Hiew, J. Z. G., Huang, X., Mou, H., Li, D., Wu, Q., & Xu, Y. (2019). BERT-based financial sentiment index and LSTM-based stock return predictability. arXiv preprint arXiv:1906.09024.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining (pp. 168–177). Association for Computing Machinery.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. International Review of Financial Analysis, 33, 171–185.

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

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceeding of the Eighteenth International Conference on Machine Learning, (pp. 282–289). Association for Computing Machinery.

Lee, R. P., & Chen, Q. (2009). The immediate impact of new product introductions on stock price: The role of firm resources and size. Journal of Product Innovation Management, 26(1), 97–107.

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. (2014). The effect of news and public mood on stock movements. Information Sciences, 278, 826–840.

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. Knowledge-Based Systems, 69, 14–23.

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., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance, 66(1), 35–65.

Mackenzie, J., Benham, R., Petri, M., Trippas, J. R., Culpepper, J. S., & Moffat, A. (2020). CC-News-En: A large English news corpus. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 3077–3084). ACM.

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. In Companion Proceedings of the The Web Conference 2018 (pp. 1941–1942). Association for Computing Machinery.

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.

Malo, P., Sinha, A., Takala, P., Ahlgren, O., & Lappalainen, I. (2013). Learning the roles of directional expressions and domain concepts in financial news analysis. In 2013 IEEE 13th international conference on data mining workshops (pp. 945–954). Institute of Electrical and Electronics Engineers (IEEE).

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System demonstrations (pp. 55–60). Association for Computational Linguistics.

Mansar, Y., Gatti, L., Ferradans, S., Guerini, M., & Staiano, J. (2017). Fortia-FBK at SemEval-2017 task 5: Bullish or bearish? Inferring sentiment towards brands from financial news headlines. arXiv preprint arXiv:1704.00939.

Moore, A., & Rayson, P. (2017). Lancaster A at SemEval-2017 Task 5: Evaluation metrics matter: Predicting sentiment from financial news headlines. arXiv preprint arXiv:1705.00571.

Morck, R., Yeung, B., & Yu, W. (2000). The information content of stock markets: Why do emerging markets have synchronous stock price movements? Journal of Financial Economics, 58(1–2), 215–260.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12, 2825–2830.

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) (pp. 1532–1543). Association for Computational Linguistics.

Pivovarova, L., Klami, A., & Yangarber, R. (2018). Benchmarks and models for entity-oriented polarity detection. In Proceedings of the 2018 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 3, pp. 129–136). Association for Computational Linguistics.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., & Hoste, V. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation* (pp. 19–30). Association for Computational Linguistics.

Qin, D. (2011). Rise of VAR modelling approach. *Journal of Economic Surveys*, *25*(1), 156–174.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of twitter sentiment on stock price returns. *PLoS One*, *10*(9), e0138441.

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (pp. 105–112). Association for Computational Linguistics.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Schouten, K., & Frasincar, F. (2015). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, *28*(3), 813–830.

Schumaker, R. P., & Chen, H. (2008). Evaluating a news-aware quantitative trader: The effect of momentum and contrarian stock selection strategies. *Journal of the American Society for Information Science and Technology*, *59*(2), 247–255.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, *27*(2), 12.

Seo, Y.-W., Giampapa, J., & Sycara, K. (2004). *Financial news analysis for intelligent portfolio management* (Technical Report). Carnegie-Mellon University.

Sinha, A., & Khandait, T. (2021). Impact of news on the commodity market: Dataset and results. In *Future of Information and Communication Conference* (pp. 589–601). Springer.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). *The general inquirer: A computer approach to content analysis*. M.I.T. Press.

Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. arXiv preprint arXiv:1903.09588.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307.

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

Tong, R. M. (2001). An operational system for detecting and tracking opinions in online discussion. In *Working notes of the ACM SIGIR 2001 workshop on operational text classification* (Vol. 1). Association for Computing Machinery.

Van de Kauter, M., Breesch, D., & Hoste, V. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications*, *42*(11), 4999–5010.

Van de Kauter, M., Desmet, B., & Hoste, V. (2015). The good, the bad and the implicit: a comprehensive approach to annotating explicit and implicit sentiment. *Language Resources and Evaluation*, *49*(3), 685–720.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer.

Von Beschwitz, B., Keim, D. B., & Massa, M. (2015). First to "read" the news: News analytics and high frequency trading. In *Paris December 2015 finance meeting EUROFIDAI-AFFI*. SSRN.

Wang, B., Huang, H., & Wang, X. (2012). A novel text mining approach to financial time series forecasting. *Neurocomputing*, *83*, 136–145.

Warner, J. B., Watts, R. L., & Wruck, K. H. (1988). Stock prices and top management changes. *Journal of Financial Economics*, *20*, 461–492.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, *39*(2–3), 165–210.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 347–354). Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac P., Rault T., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. (2019). HuggingFace's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

Wysocki, P. D. (1998). *Cheap talk on the web: The determinants of postings on stock message boards* (Working Paper 98025). University of Michigan Business School.

Xu, H., Liu, B., Shu, L., & Yu, P. S. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. arXiv preprint arXiv:1904.02232.

Zhang, W., & Skiena, S. (2010). Trading strategies to exploit blog and news sentiment. In *Proceedings of the international AAAI conference on web and social media* (Vol. 4). AAAI Press.

APPENDIX A

A.1 | Entity-sentiment extraction results
In Tables A1 and A2, we present the average and standard deviation of the performance metrics Accuracy and F1-Score for all the learning schemes across the 31 runs.

A.2 | Economic value of sentiments: Percentage price returns
Alternatively, we calculate the daily after-market percentage price returns for the NSE 500 index, using the opening and closing prices as

$$d_{i,\text{NSE500}} = \frac{P_{\text{open},i+1,\text{NSE500}} - P_{\text{close},i,\text{NSE500}}}{P_{\text{close},i,\text{NSE500}}}.$$

The results are presented in Table A3. On an average, the alternate sentiment measure which accounts for neutral sentiments expresses higher explanatory power as measured by multiple R^2, and a stronger significant relationship with percentage price returns across all the years considered.

A.3 | Vector autoregression analysis of sentiments and market index time-series
Vector autoregression (VAR) models are stochastic process models utilized to evaluate the relationships between multivariate time-series. VAR models have risen to prominence in economics and natural sciences (Qin, 2011) leading to data-driven theories. In the current context, the dependent variable time-series is the daily NSE500 index percentage price returns (d_{NSE500}), and the independent

TABLE A1 Average of the performance metrics across the 31 runs

Learning scheme	Positive		Negative		Neutral	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
LPS + SVM	81.92	75.09	85.95	79.00	77.93	71.73
LPS + GBM	82.61	75.99	86.47	80.29	78.16	71.34
UBT + SVM	84.16	79.16	87.70	81.18	80.13	73.25
UBT + GBM	84.54	79.87	87.57	81.64	80.93	73.18
UBT + MLP	82.15	76.78	87.20	75.25	79.20	73.10
GloVe + RNN	83.56	76.23	86.46	74.21	79.01	71.75
GloVe + BiRNN	83.03	75.44	85.78	72.76	78.29	71.20
GloVe + LSTM	84.73	78.04	87.08	76.10	80.51	74.05
GloVe + BLSTM	84.73	78.27	86.75	75.59	80.79	74.26
finBERT	90.93	87.02	92.63	85.88	88.87	85.72

TABLE A2 Standard deviation of the performance metrics for 31 runs

Learning scheme	Positive		Negative		Neutral	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
LPS + SVM	0.80	1.30	0.98	1.46	0.80	0.96
LPS + GBM	0.75	1.15	0.83	1.17	0.81	1.04
UBT + SVM	0.97	1.42	0.86	1.37	0.96	1.07
UBT + GBM	0.91	1.23	0.84	1.22	1.11	1.42
UBT + MLP	0.75	1.21	0.90	1.24	1.05	1.35
GloVe + RNN	0.64	1.12	0.54	0.92	0.63	1.54
GloVe + BiRNN	0.32	1.00	0.50	0.82	0.42	1.02
GloVe + LSTM	0.36	0.75	0.38	1.62	0.44	0.83
GloVe + BLSTM	0.50	0.85	0.38	0.80	0.40	0.83
finBERT	0.20	0.28	0.15	0.30	0.20	0.27

TABLE A3 Results depicting the economic value of sentiments across the years 2012–2017. The primary and alternate sentiment measures show significant effects across the years 2014–2017. The alternate sentiment measure takes into account the neutral sentiments and captures more variation than the primary sentiment measure. In doing so, the alternate sentiment measure exhibits significant effects in the years 2012 and 2013 as well

Year	# observations	Primary sentiment measure (s_{AM}^1)		Alternative sentiment measure (s_{AM}^2)	
		beta	*r*-squared	beta	*r*-squared
	261	0.00150	0.0073	0.00314*	0.0108
2013	260	0.00240	0.01	0.00557**	0.0158
2014	260	0.00557***	0.1088	0.00116***	0.1298
2015	249	0.00240**	0.0166	0.00441**	0.0168
2016	260	0.00414**	0.0177	0.00882**	0.0209
2017	260	0.00176*	0.0145	0.00326*	0.0139

*$p < .1$; **$p < .05$; ***$p < .01$.

time-series is the daily sentiment measure (*s*). The VAR (p_1, p_2) model can be elaborated as

$$d_{\text{NSE500}}(t) = \alpha_1 * d_{\text{NSE500}}(t-1) + \alpha_2$$
$$* d_{\text{NSE500}}(t-2) + ... + \alpha_{p_1}$$
$$* d_{\text{NSE500}}(t-p_1) + \beta_1 * s(t-1) + \beta_2$$
$$* s(t-2) + ... + \beta_{p_2} * s(t-p2) + \varepsilon.$$

In our analysis, we test for $p_1 = 0, 1, 2, 3$ and $p_2 = 0, 1, 2, 3$, for the years 2012–2017. For sentiment, we utilize the alternate sentiment measure which in our experiments expressed higher explanatory power compared to the primary sentiment measure. In Table A4 below we present the models which expressed statistical significance for at least one independent variable. The VAR models were implemented in R using the *dynlm* package

TABLE A4 VAR model summary across the years 2012–2017

Year	VAR	d_{t-1}	d_{t-2}	d_{t-3}	s_{t-1}	s_{t-2}	s_{t-3}	BIC
	(0,2)				0.0014	0.0018+		−2596.30
2012	(0,3)				0.0015	0.0019+	−0.0016	−2582.01
2012	(1,2)	0.0777			0.0012	0.0017+		−2592.31
2012	(3,1)	0.0837	0.1244+	−0.0788	0.0013			−2578.15
2012	(1,3)	0.0877			0.0013	0.0017+	−0.0017	−2578.43
2012	(2,3)	0.078	0.1147		0.0014	0.0015	−0.0019	−2576.32
2013	(3,0)	−0.0886	−0.0514	−0.1293+				−1242.86
2013	(3,1)	−0.0892	−0.052	−0.1275+	−0.0001+			−1238.07
2013	(3,2)	−0.0904	−0.0531	−0.129+	−0.0001+	−0.0001		−1233.33
2013	(3,3)	−0.0895	−0.0518	−0.1278+	−0.0001+	−0.0001	0.0001	−1228.70
2014	(3,0)	0.2252**	−0.0242	0.1558+				−2601.36
2014	(2,0)	0.2287**	0.0085					−2609.64
2014	(1,0)	0.23**						−2626.18
2014	(1,1)	0.27***			−0.0016+			−2624.99
2014	(2,2)	0.269***	0.0189		−0.0016*	0.0002		−2603.08
2014	(1,2)	0.2734***		−0.0016*	0.0003		−2608.55	
2014	(3,1)	0.2621***	−0.0087	0.1589+	−0.0016*			−2600.38
2014	(3,2)	0.2621***	−0.0086	0.1589+	−0.0016*	0		−2594.83
2014	(3,3)	0.2623***	−0.0168	0.1727*	−0.0016+	0.0001	−0.0005	−2589.76

TABLE A4 (Continued)

Year	VAR	d_{t-1}	d_{t-2}	d_{t-3}	s_{t-1}	s_{t-2}	s_{t-3}	BIC
2014	(1,3)	0.271***			−0.0016+	0.0003	0.0001	−2593.63
2014	(2,3)	0.266***	0.0218		−0.0016+	0.0002	0.0001	−2588.19
2015	(3,0)	0.0455	0.0235	−0.1368				−2423.70
2015	(2,2)	0.0275	0.0438		0.0019+	−0.0021**		−2432.36
2015	(1,2)	0.0281			0.0019*	−0.002*		−2437.43
2015	(3,1)	0.018	0.0167	−0.146+	0.0017+			−2421.83
2015	(3,2)	0.027	0.0479	−0.1396+	0.0021*	−0.002*		−2421.53
2015	(3,3)	0.0173	0.0522	−0.1252	0.0021*	−0.0018*	−0.0009	−2417.08
2015	(1,3)	0.0148			0.002*	−0.0017*	−0.0013+	−2423.52
2015	(2,3)	0.0136	0.0507		0.002*	−0.0018*	−0.0013+	−2418.62
2015	(0,2)				0.002*	−0.002*		−2442.79
2015	(0,3)				0.002*	−0.0017*	−0.0013+	−2429.02

$+p < .1$; $*p < .05$; $**p < .01$; $***p < .001$.

and the standard errors were calculated using the *coeftest* package. The results show the α, β corresponding to the lagged time-series with values rounded to four decimal digits. For all the models results, kindly contact the corresponding author.

We observe that the lags in sentiments and index returns differ in their effects across the years. In 2012 and 2014, sentiments lagged by 2 days showed positive and negative statistically significant relationships with the daily percentage price returns respectively. In 2013, sentiments lagged by 1 day showed negative and statistically significant relationship with the daily percentage price returns. In 2015, lowest BIC value model indicates that the daily percentage price returns is affected by the sentiments of 1 day and 2 day lags, however, with differing relationships positive and negative, respectively. For the years 2016 and 2017, we found that no models exhibited statistical significance.