Buy, Sell or Hold: Entity-Aware Classification of Business News

Ankur Sinha
Satishwar Kedas
Rishu Kumar
Production and Quantitative Methods,
Indian Institute of Management Ahmedabad, India
Pekka Malo
Department of Information and Service Economy,
Aalto University, Finland

W.P. No. 2019-04-02
April 2019

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD – 380015
INDIA

Buy, Sell or Hold: Entity-Aware Classification of Business News

Ankur Sinha
Satishwar Kedas
Rishu Kumar
Production and Quantitative Methods,
Indian Institute of Management Ahmedabad, India
Pekka Malo
Department of Information and Service Economy,
Aalto University, Finland

Abstract

Financial sector is expected to be at the forefront of the adoption of machine learning methods, driven by the superior performance of the data-driven approaches over traditional modelling approaches. There has been a widespread interest in automatically extracting information from financial news flow as the signals might be useful for investment decisions. While quantitative finance focuses on analysis of structured financial data for investment decisions, the potential of utilizing unstructured news flow in decision making is not fully tapped. Research in financial news analytics tries to address this gap by detecting events and aspects that provide buy, sell or hold information in news, commonly interpreted as financial sentiments. In this paper, we develop a framework utilizing information theoretic concepts and machine learning methods that understands the context and is capable of extracting buy, sell or hold information contained within news headlines. The proposed framework is also capable of detecting conflicting sentiments on multiple companies within the same news headline, which to our best knowledge has not been studied earlier. Further, we develop an information system which analyzes the news flow in real-time, allowing users to track financial sentiments by company, sector and index via a dashboard. Through this study we make three dataset related contributions - firstly, a training dataset consisting of more than 12,000 news headlines annotated for entities and their relevant financial sentiments by multiple annotators, secondly, an entity database of over 1,000 financial and economic entities relevant to Indian economy and their forms of appearance in news media amounting to over 5,000 phrases and thirdly, make improvements in existing financial dictionaries. Using the proposed system, we study the effect of the information derived from daily news flow in the years 2012 to 2017, over the Indian broad market equity index NSE 500, and show that the information has predictive value.

1 Introduction

The rate at which data is being generated continues to grow at an exponential pace. Given the significant enhancements in the storage capacities, almost every sector of the global economy is aiming at capturing and storing this vast resource. Artificial intelligence/Machine Learning (AI/ML) is today's buzzword in financial markets that is capable of providing some compelling solutions to financial services providers, investors and analysts. In fact, the vast amount of data that is generated in this sector coupled with the recent success of machine learning methods is already driving a number of solutions in banking, finance and insurance [Hussain and Prieto, 2016]. The applications include risk measurements and assessments in auto insurance [Viaene et al., 2005, Guelman, 2012, Šubelj et al., 2011], credit ratings in mortgages and loans [Khandani et al., 2010, Tsai and Chen, 2010], asset pricing in real estate [Plakandaras et al., 2015], algorithmic trading in financial markets [Shah, 2007, Kearns and Nevmyvaka, 2013, Gerlein et al., 2016], and robo-advisory for retail and institutional investors [Salampasis et al., 2017], to name a few. While risk measurements, credit ratings and pricing models have been studied for decades and are still evolving, the applications of machine learning methods in algorithmic trading and robo-advisory is a relatively new phenomenon. The investors and analysts are often overwhelmed with too much of information and often struggle with huge quantities of structured and unstructured data. While data can be owned easily, the real value lies in the ability to make sense out of this data, which along with unknown patterns, trends or signals, contains a significant amount of noise.

The area of quantitative finance deals with the analysis of the structured quantitative information that is produced in the financial markets to make informed investment decisions. However, both analysts and investors are known to follow a variety of strategies in making investment decisions, for instance, focusing on the unstructured information like news is another key ingredient that users seldom ignore. Given a large amount of news flow, it often becomes difficult to track the information contained in this unstructured data. The cognitive limitations of focusing only on a small sample of news items and the opinions contained within poses a threat of biases. [Park et al., 2013] show that human investors often succumb to confirmation bias while consuming information from such unstructured data, with the degree of confirmation bias increasing with the degree of perceived knowledge and prior beliefs concerning a specific company.

To tackle such difficulties that come with financial news analysis, we propose a model that is able to recognize the entities and extract entity-relevant information (financial sentiments - buy, sell or hold) from news headlines. We utilize the model as the backbone of an information system aimed at presenting information contained in large number of financial news to users in real-time. To enable the model to perform entity-aware sentiment extraction, we provide a dataset of over 12,000 news headlines annotated for the entities and their respective sentiments by multiple annotators. To the best knowledge of the authors, datasets annotated with multiple entities are not available in the area of Finance, nor there exist research articles that attempt to extract sentiments for multiple entities from financial news. Overall, we make a three-fold contribution in this paper as mentioned below:

1.1 Entity-aware sentiment recognition

We focus on entity-aware sentiment recognition task, where multiple entities occur in a single headline, and the system extracts sentiments for each entity. For instance, consider the following news headline, *Bullish on L&T, but BHEL and SBI seem weak*. The headline contains multiple sentiments that are conflicting, i.e. *positive (buy)* for L&T and *negative (sell)* for BHEL and SBI. We learn the phrase structure responsible for assigning sentiments to various entities in a news headline using a target-based approach embedded in an N-gram model. We compare the performance of the proposed approach with the state-of-the-art methods in the literature. In addition to identifying the sentiments, we are also able to identify the reasons for the sentiment with the help of the financial concepts that might have been discussed in the news headline.

1.2 Establishment of a phrase-bank for an emerging economy

To further research on emerging economies, we provide a hand-annotated dataset of more than 12,000 news headlines derived from "The Economics Times" database. These headlines are annotated for the financial sentiments with respect to the entities present in each headline. Out of the 12,000 news headlines, 1,450 news headlines contain multiple entities; i.e. multiple sentiments in the same news headline. The phrase bank is utilized as the training and testing corpus for the models and machine learning methods supporting the information system proposed in this paper.

It is often in the interest of researchers to study the economy at various levels - individual entities such as stocks or assets to a group of entities such as a sector or index. To build an information system that can support large variety of queries, there exists a need for a database of various entities and their multiple forms of occurrence in news media. We manually created a database of 907 listed companies, 11 sectors, and 92 other entities (commodities, currencies etc.) that appeared in the Indian financial news in the past decade. This amounts to a total of 5,070 phrases, which cover the various ways the entities were represented in the news headlines. Further, we build on the lexicon provided by [Loughran and McDonald, 2009] and modify the feature annotations of 895 words which are commonly used in financial text.

1.3 Information Content and Visualization

We present a dashboard that presents the information derived from news flow in a readily consumable form through visual elements. Further, we tailor the information system to the Indian economy by using the datasets developed in this paper, and extract financial sentiments for daily Indian financial news flow generated in the years 2012 to 2017. After adjusting for the information transfer between domestic market and media, we find strong evidence that news flow has economic value.

The paper is organized as follows. In Section 2, we provide a survey on financial text analysis, which includes various lexicons, methods and advancements made in the area. Section 3 introduces the framework for extracting entity-relevant information from news flow, followed by Section 4 that provides a discussion on the creation of training dataset and its annotation process. Section 5 provides the results of the models in terms or various performance measures. Section 6 evaluates the information produced from the proposed framework in terms of predictive value and also provides the design for the dashboard. Finally, we provide the conclusions and future research ideas in Section 7.

2 Financial Text-Analysis Survey

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. With increasing flow of news and opinions over Internet, analysis of this unstructured qualitative information became a hot research topic especially with respect to financial markets. Increasingly research has been directed towards tracking financial and economic developments in real time through automated methods utilizing news flow and machine learning techniques. [Wysocki, 1998] was the first one 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 and Frank, 2004, Das and Chen, 2001] and RagingBull.com [Antweiler and 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. They also are among the first ones to recognize the importance 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. Specifically, the impact of negative words on firms' future earnings and stock prices was found to be prominent; suggesting that the automated news analysis should provide valuable information for traders and analysts.

[Loughran and McDonald, 2009], 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 6 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., 2014, Malo et al., 2013], identified that representation of financial entities as directionality-dependent words (ex. profit as "Positive-if-up", loss as "Negative-if-up" etc.) is essential to capture the sentiment expressed in the financial/economic text. To illustrate, the financial entity - "profit" expresses neutral sentiment in *"The profit stands at $ 100 million"*, and a positive sentiment in *"The profit increased by 10%"*. The directionality word "increased" when combined with the word "profit" makes the sentiment positive in the second sentence. To the best of our knowledge, they are the first ones to apply phrase level sentiment analysis models to the buy, sell or hold classification task. They argue that the overall sentiment of the phrase depends on the word level features, rather than the words itself, and use various structures - Linearized Phrase Structure (LPS), Constituency trees and Dependency trees to model phrase-level sentiment. Their major contributions are a publicly available human-annotated datasets of 5,000 news headlines on Finnish companies and a list of financial entities with their directional dependencies. However, their proposed model only extracts a single sentiment from sentences without identifying the entity to which the sentiment corresponds to. This limits the application of their framework in building an information system which is capable of identifying entity-specific sentiments, particularly being able to extract conflicting sentiments on entities within the same sentence.

Over the last decade, with explosion in social media usage, media houses, market participants, and general public have adopted Twitter as the primary medium for information exchange. Various studies have been directed towards extracting information from Twitter feeds and measuring its predictive value [Zhang et al., 2011, Bollen et al., 2011, Mao et al., 2012, Ranco et al., 2015]. Most of the studies till date, have been directed towards developed economies and majorly the firms relevant to the S&P 500 index or Dow Jones Industrial Average index, arguably because of the market capitalization of the companies tracked. Limited research has been directed towards preparation of datasets for emerging economies and identifying the relevant modifications required.

News analytics has seen mainstream adoption across companies in the financial services industry. Especially, sentiment monitoring tools like RavenPack are being utilized by High Frequency Traders, to gain a competitive advantage and conduct company or industry oriented directional trades [von Beschwitz et al., 2015]. However, such commercial tools provide limited understanding of the methodology and framework that underlies the sentiment extraction process. Limited research efforts have been directed towards presenting a formal design for an information system for extracting and visualizing insights from unstructured news flow that can be adapted by traders, investors and public at large.

3 System Design

The system for entity-aware sentiment recognition (Figure-1) proposed in this paper works in three stages - first, Pre-processing stage that involves data cleaning and tagging, second, modelling stage that involves using entity-aware sentence representation models to convert the unstructured textual data into a vector form of word-level and phrase-level features, and third, classification stage that involves application of a supervised classification technique to extract entity relevant financial sentiments. The recommendations thus derived are stored in a database, which is further utilized for presenting information via a dashboard. The three stages are discussed in detail in this section.

3.1 Pre-processing

The first stage of the system involves recognition of entities present in the news headline and pruning the headline for the feature extraction process. The following steps are undertaken,

3.1.1 Entity Tagging :

Financial news is different from general news, as the news headlines often contain information directed towards specific forms of entities, which comprise named entities such as companies, associations, regulatory organizations etc., and financial and economic entities such as sectors, indices, commodities, currencies etc. To achieve high accuracy in extraction of entity relevant sentiments, it is necessary to have a process that is effective in recognizing entities in a news headline. Our preliminary experiments with the Stanford CoreNLP Toolkit [Manning et al., 2014, Finkel et al., 2005], have shown poor results in recognizing entities from the news headlines derived from Indian business news providers. For example, in the headline - *Pharma companies like Sun Pharma, Dr Reddy's Labs to do well*, the CoreNLP Named Entity Recognizer fails to recognize *Dr Reddy's Labs* as an organization and *Pharma* as a sector.

The bottleneck in achieving high recognition accuracies is the lack of a database with entities and their various forms of appearance in news media. This gap has been addressed for developed economies, especially driven by high academic research interest. For entity recognition, we have developed an Entity Database (discussed in Section-4) covering entities relevant to the Indian economy. In the entity tagging step, we scan through the news headline, tag the entities identified, and replace the entities with the symbols assigned as per the database. The following step is then undertaken to prepare the headline for feature extraction process.

3.1.2 Text Pre-processing :

Often, the news headlines extracted from the news databases contain special characters, stop words, and HTML tags. These objects (commonly referred to as grammar) do not provide information relevant to the financial sentiments contained in the sentence. Research efforts have shown that the text pre-processing step significantly improves the sentiment recognition performance [Haddi et al., 2013]. Post entity tagging, we scan through the headline and remove the identified grammar objects. This pre-processing step does not affect the lexical organization of the news headline, therefore conserving its syntactic structure.

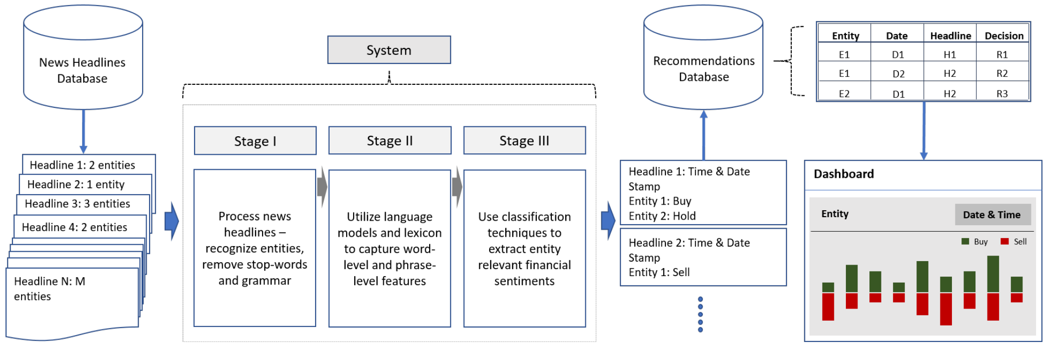

Figure 1: An overview of the information system proposed in this work. The news headlines are captured real-time from the News Headlines database, which go through the 3 stages - pruning, modelling and extraction of the system. The recommendations derived are stored in the Recommendations database in the format depicted above. The dashboard presents the information thus extracted in a visual form in real-time

3.2 Entity-aware sentence representation modelling

Traditional language models for the sentiment classification task have used word-based methods such as bag-of-words models to represent text, and weighting schemes based on frequency or presence to quantify the information contained in the text [Pang et al., 2002, Pang et al., 2008, Liu, 2012]. Lexicon-based methods [Taboada et al., 2011, Bruce and Wiebe, 1999, Hu and Liu, 2004, Riloff and Wiebe, 2003] which assume that the words have prior polarities that can be quantified have shown to exhibit better performance than word-based approaches, and have led to the development of various language resources for sentiment analysis [Wiebe et al., 2005, Baccianella et al., 2010, Cambria et al., 2014, Deng and Wiebe, 2015].

In the financial context, research efforts by [Loughran and McDonald, 2009] indicate that the financial concepts are often misinterpreted in the general context, and establish a finance-focused sentiment analysis dictionary (LM lexicon). Financial concepts like *liability, risk* are assigned *negative* prior polarities as per the General Inquirer (GI) [Stone et al., 1966], while in the financial context they carry *neutral* prior polarities. Further, [Malo et al., 2013] establish that the financial sentiments are often determined by the change in financial concepts mentioned in the sentence, and develop a dictionary of financial concepts with their directional dependencies. For example, the expression *profit widens by 10%* contains a *positive* sentiment, derived from an upward movement, *widens* in the financial concept *profit*. An application of the LM lexicon would suggest that *profit, widens* are *neutral* words, leading to an overall *neutral* sentiment. In their subsequent work [Malo et al., 2014], proposed the Linearized Phrase Structure (LPS) model based on sequence-of-literals framework to capture the underlying context in the sentence.

The second stage termed as the modelling stage, involves two phases, first, a curated financial lexicon is utilized to extract word-level features without affecting the phrase structure, and second, the phrase-structure models extract the context capturing the interactions among the word-level features present in the headline.

3.2.1 Word-level Feature Extraction

We utilize five types of word-level features :

(i) Prior Polarities - To capture prior semantic orientation of the word. If the word does not belong to the lexicon utilized in the experiments, it is deemed neutral prior polarity,

(ii) Directional Dependency - To capture the dependency of the word or entity on directionality expressions. The semantic orientation is determined based on the relevant directionality expression (e.g. Profit - *positive-if-up*),

(iii) Directionality: To capture the movement induced by the word's semantics on directionality dependent words or entities. We consider two categories of movement - upward and downward (e.g. sales declined - sales *Down*),

(iv) Contextual valence shifters: To capture the influence of the word on the phrase level sentiment. Among the various classes, negators are the most relevant for our experiments (e.g. despite losses - *negate* losses), and

(v) Financial Entities: To capture the financial concepts present in the headline and their relevant prior polarities and directional dependencies (e.g stock - *price per share*)

In cases of conflict, we used the following order of feature importance - Financial Entity > Contextual Valence Shifter > Directionality > Directional Dependence > Prior Polarity.

Lexicon : We use MPQA [Wiebe et al., 2005], GI [Stone et al., 1966] and LM [Loughran and McDonald, 2009] 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.

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 *bullish (positive)* sentiment towards "Infosys", while in "Nifty rally ends", "rally" describes an action already in place and overall sentiment for "Nifty" is *bearish (negative)*. After an investigation of the news headlines training dataset developed in this paper, a custom dictionary containing 970 words commonly used in the financial news has been 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 evaluating the custom dictionary, we define two lexica - *Lexicon A* which includes LM, MPQA, GI and Malo et al, and *Lexicon B* which includes Lexicon A and our custom dictionary. In Lexicon B, the feature annotations based on the custom dictionary are given preference over the LM dictionary. We compare the results of models trained separately on Lexicon A and Lexicon B in Section-5

A summary of the overall financial lexicon statistics are presented in the Table-1

Table 1: Financial Lexicon Statistical Summary

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
Lexicon A = LM + MPQA + GI + Malo et al.						
Lexicon B = Lexicon A + Custom						

Handling multiple entities : In cases of headlines with multiple entities, the information about the syntactic structure that capture the interactions between the entities and their relevant sentiment bearing expressions becomes important for effective sentiment extraction. Research efforts towards capturing the syntactic structures have typically used constituency trees and dependency trees [Klein and Manning, 2004, Culotta and Sorensen, 2004], which work well in the case of well

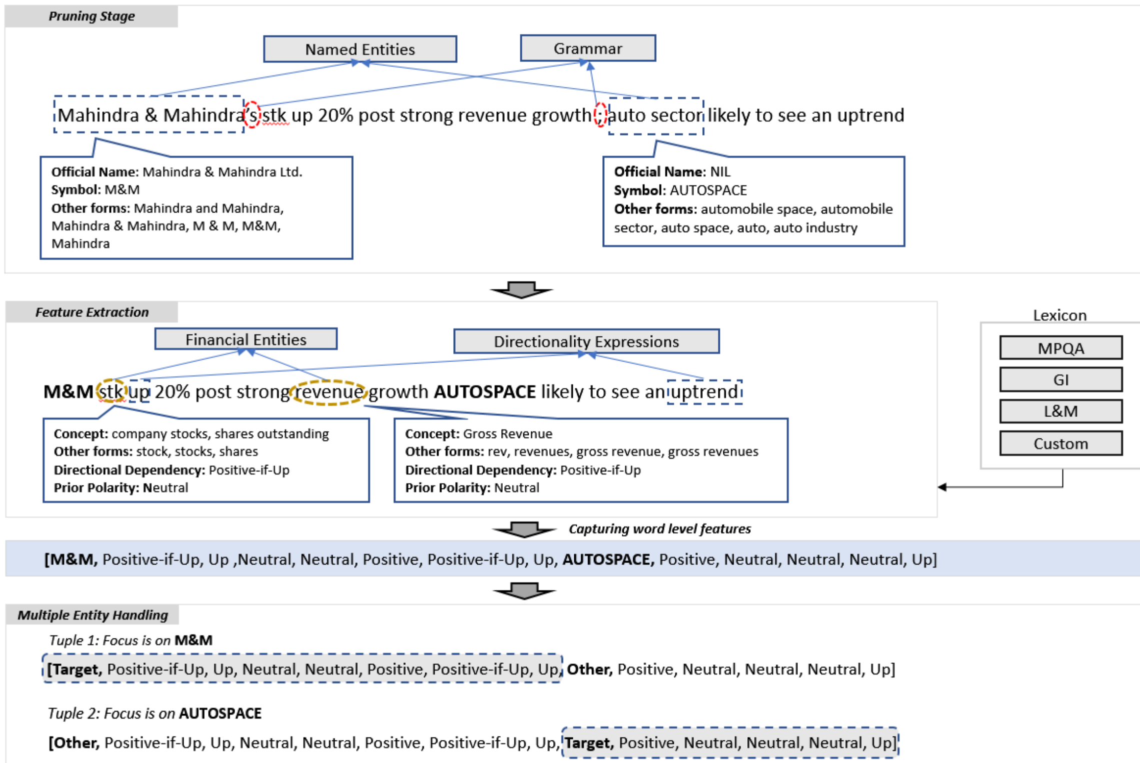

Figure 2: An example depicting the first two stages of the system and the steps undertaken in each of the stages. A news headline is converted into a tuple of word-level features

structured textual data such as paragraphs and documents. However, the news headlines are often short and the contextual information is highly localized. These characteristics enable us to solve the entity-based sentiment extraction task efficiently by introducing two word-level features - "Target" and "Other" used to represent the entities present in the news headline. If there exists only a single entity in the news headline, we represent it with the "Target" feature, and "Other" feature is not used. However, if there exist multiple entities as determined in the entity-tagging step of pruning stage, we create multiple instances of the news headline matching the number of entities present in the news headline. In each instance, the entity for which the sentiment has to be extracted is replaced with "Target" feature, and the other entities are replaced with "Other" feature.

Post pruning stage, the word-level feature extraction step is operationalized by traversing the phrase along the tokens and identifying the relevant lexicon entry in the financial lexicon. The tokens are then replaced by the lexicon entries, maintaining the sequence in which the tokens appear in the phrase. The resulting tuple is checked for presence of multiple entities as determined by the entity tagging step, and if present, the multiple entity handling process generates tuples for individual entities using "Target" and "Other" word-level features. As shown in the example in Figure. 2, in the pruning stage - the headline is processed for grammar, and the entities are identified. In the modelling stage, the headline is traversed in sequence and the lexicon entries are matched with the tokens, capturing the word-level features. Since there are two entities *M&M* and *AUTOSPACE*, two tuples are generated, one with *M&M* and the other with *AUTOSPACE*, as the target entities.

Let F denote the collection of word-level features, and let S be the space of possible phrases. The ***word-level feature extraction step*** *can be defined as a mapping $f : s \mapsto [f_1, f_2, \ldots, f_n]$, which presents the phrase as a tuple of entities, $f_i \in F, i \in \{1, \ldots, N_f\}$*

3.2.2 Vectorization: Phrase Structure Models

In the vectorization step, the tuple generated in the feature extraction step is converted to a vector of phrase-level features. The phrase-level features are designed to capture the interactions between various word-level features present in the sentence. These interactions are defined by the types, frequencies and positions of the word-level features in consideration. For example, in the tuple *[Target, Positive-if-Up, Up]*, phrase-level features can be the number of occurrences or positions of the word-level features, or the interactions such as "*Positive-if-Up follows Target*" and "*Up follows Positive-if-Up*".

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 et al., 1994]. However, n-grams have typically been used on words rather than features extracted from words. We note that, n-grams in the context of sentiment extraction from news headlines have the ability to capture the syntactic structure and frequencies of interactions of word-level features. Unlike previous approaches, we apply the n-gram model after extracting word-level features, rather than on words itself. We formalize the vectorization step based on the n-gram models below.

Let F be as defined above with cardinality N_f, and consider a mapping f(s) for a specific phrase (s). We define n vectors V_1, V_2, V_3, ..., V_n such that,

$$V_1 = \{n_{f_1}, n_{f_2}, ..., n_{f_{N_f}}\}$$

$$V_2 = \{n_{(f_1,f_1)}, n_{(f_1,f_2)}, ..., n_{(f_{N_f},f_{N_f})}\}$$

$$V_3 = \{n_{(f_1,f_1,f_1)}, n_{(f_1,f_1,f_2)}, ..., n_{(f_{N_f},f_{N_f},f_{N_f})}\}$$

$$...$$

$$V_n = \{n_{(f_1,f_1,f_1,...,f_1)}, n_{(f_1,f_1,f_1,...,f_2)}, ..., n_{(f_{N_f},f_{N_f},f_{N_f},...,f_{N_f})}\}$$

where n denotes the number of occurrences of the tuples corresponding to the co-occurrences of word-level features in the mapping f(s).
We further define the various phrase-structure models based on the n-grams as follows:

Table 2: N-gram based models

Model	Representation
Uni-gram (U)	[V_1]
Uni-gram + Bi-gram (UB)	[V_1, V_2]
Uni-gram + Bi-gram + Tri-gram (UBT)	[V_1, V_2, V_3]
Uni-gram + Bi-gram + ... + N-gram (N)	[V_1, V_2, V_3, ..., V_n]

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. In their approach, the news headline is traversed sequentially to capture the words and their relevant word-level features followed by an entity pruning step, wherein the sequentially repeating features are removed. The final entity sequences is then converted into a bit vector using a binary coding scheme. They also consider the reduced form of LPS (rLPS) wherein the sequentially repeating features are not removed. With the new features "Target" and "Other", the LPS model will be able to capture the various entities and their relevant contexts from the news headline. Further, we look at utilizing Huffman Coding [Huffman, 1952] which has been shown to be effective in coding information into bits. We refer to the two LPS models using Huffman codes as Huffman-based LPS (H-LPS) and Huffman-based rLPS (H-rLPS) models. The Huffman codes are derived from the frequencies of word-level features in the training dataset.

3.3 Classification methods

We model the problem of extracting financial sentiments from news headlines as a supervised multi-class classification task, with sentiments (recommendations) falling into one of the three classes - Buy, Sell and Hold. We utilize a human-annotated dataset of more than 12,000 news headlines, with entities and their respective classes to train the machine learning algorithms used for classification. To capture the sentiment relevant to an entity, we employ three binary classifiers (Buy vs Sell, Buy vs Hold, Sell vs Hold) with a majority-vote. For example, if the three binary classifiers generate *Buy, Buy* and *Hold* recommendations, the final recommendation attributed to the entity is *Buy*.

The binary classifiers are chosen from two families of classification algorithms - Gradient Boosting Machines (boosting methods) and Support Vector Machines (kernel-based classification methods). We intend to identify the family of classifiers most suited for the entity-based sentiment extraction task as formulated above.

Support Vector Machines (SVMs) are a family of kernel-based binary classification methods originally introduced by Cortes and Vapnik [Cortes and Vapnik, 1995, Vapnik, 2013], which maximize the margin defined by the distance between the nearest points (support vectors) to a hyper-plane that separates the two classes in the feature space. The SVM family of classifiers have shown to exhibit very high scalability and adaptability, providing flexibility to the user in the choice of basis kernels and fast learning compared to other methods. The underlying kernel method defines the feature space, and the most commonly used kernel is the Linear kernel.

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.

In our experiments, we compare the results of two classifiers - linear kernel-based SVM with One-vs-One strategy and the Gradient Boosting (GB) Classifier.

4 Dataset

One of the major concerns, regarding conducting machine learning research in the financial context, has been the lack of publicly available and reliable human-annotated datasets. Various research efforts in the last decade, have been directed towards mitigating this issue. We discuss some of the available finance/economics related corpora and dictionaries below.

4.1 Available Financial Text Corpora

MPQA Opinion Corpus: The MPQA Opinion Corpus is a collection of manually annotated 535 news article from various news sources for sentiments and other private states (i.e beliefs, emotions, sentiments, speculations, etc.). [Wilson et al., 2005] explains the contextual polarity annotations and an agreement study.

JRC Corpus of general news quotes: JRC Corpus consists of 1592 quotes from newspaper articles in English for opinion mining. [Balahur et al., 2013] reached an 81% inter annotator agreement between three pairs of 2 annotators.

Financial Phrase Bank: [Malo et al., 2014] provides a collection of 5000 phrases/sentences manually annotated by 16 subject experts. The phrases/sentences were collected from LexisNexis database for Finnish companies.

SemEval-17 Task 5[1]: This task for fine-grained sentiment analysis on financial microblogs and news contained StockTwits Messages and News Statements & Headlines. These messages consist of microblog messages focusing on events in stock markets and inverstor assesments. The news headlines were crawled from different news sources.

Financial Polarity Lexicon: This polarity dictionary provides a means of determining tokens' relevance in financial sentiment analysis. It also has flags for sentiment dictionary such as positive, negative, litigious, uncertainty.

Our Contribution: Along with this paper, we are releasing a dataset of more than 12,000 news headlines annotated for entities present and their relevant financial sentiments. We also provide a database of over 1,000 entities - companies, sectors, indices and other assets relevant to the Indian economy, annotated for their various forms of appearance in media.

4.2 Annotation Process

In total, 12,417 headlines were sampled from all the financial and economic news pertaining to the duration 2011-2015 from the Economics Times database. Three annotators with relevant academic backgrounds were identified and given the tasks of recognizing the entities in the headlines and assigning a financial sentiment (class) to the relevant entities. All annotators were provided with the 12,417 news headlines identified for preparation of the dataset. The annotators were not informed of the presence of other annotators until the preparation of the final dataset. The annotation process was curated by one of the primary authors and the final decision on the class annotation was driven by a consensus between the curator and the three annotators. Overall, each news headline was processed for entities and relevant financial sentiments by four human annotators. Further, following guidelines were issued to the annotators to be considered during the annotation process:

(i) 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.,

(ii) Entities are expressions representing or referring to named entities such as companies and indices and economic entities such as sectors, commodities and currencies. Under this broad notion of entities, various terms such as "Rupee", "Pharmaceutical sector", "auto space" etc. are also considered.

(iii) The classes were to be chosen from among one of the three financial sentiments defined below:

    (a) *Buy*: A positive sentiment is expressed by the phrase or sub-phrase in relevance to the recognized entity

    (b) *Sell*: A negative sentiment is expressed by the phrase or sub-phrase in relevance to the recognized entity

    (c) *Hold*: A neutral sentiment or no sentiment is expressed by the phrase or sub-phrase 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.

Ambiguity: Ambiguity is inherent in human decision making, and especially arises when the same dataset is being annotated by multiple researchers. A sample of 100 annotated headlines, and

[1] http://alt.qcri.org/semeval2017/task5/

list of 15 most typical ambiguous cases with the rationale for the final recommendation was provided to the annotators in the initial training phase. The cues for the sources and reasons for ambiguity were derived from the observations in [Malo et al., 2014]. We intended to create a single dataset which can be utilized as a standard to measure the performance of various models, dictionaries and classification techniques.

4.3 Dataset

For entity recognition, we have manually compiled a list of 5,070 phrases/words that refer to 907 companies listed on NSE 500 index in the duration 2006-2018, 11 sectors and 92 other entities (commodities, currencies, indices etc.) relevant to the Indian economy. The phrases depict the various forms in which the entities have been represented in news headlines across two major Indian business news providers (The Economic Times[2] and Moneycontrol[3]). Each database entry contains three values - first, the symbol assigned to the entity (stock ticker for a company); 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}

The News Headlines Dataset containing 12,417 news headlines annotated for the entities and their financial sentiments. Each dataset entry contains three values - first, the news headline; second - the list of entities appearing in the news headline; and third - the list of financial sentiments attributed to each entity in the order of the list of entities. Of the annotated headlines, 88.32% contain single entity, 11.68% contain mulitple entities. This dataset serves as the training and testing corpus for the proposed framework in this paper to extract multiple financial sentiments from individual news headlines.

5 Model evaluation

In this section, we discuss the results of our comparative analysis on various model, lexicon and classifier combinations discussed below. From here on, we will refer to a combination of a model, a lexicon and a classifier as a learning scheme.

For our experiments, we have considered three modelling approaches - N-gram based approach, LPS based approach, and Bag-of-Words approach. The descriptions of the three approaches are provided below:

(i) N-gram based approach: The models based on the N-gram based approach capture the syntactic structure of news headline using a frequency vector of joint occurrence of word-level features.

(ii) LPS based approach: The LPS based approach utilizes a sequence-of-literals framework to capture the syntactic structure and word-level features from the news headline.

(iii) Bag-of-Words (BoW) approach: In this approach, the news headline is modelled using a vector of weighted-frequencies or presence of words derived from a vocabulary. The vocabulary is learnt from the training dataset, and does not utilize a lexicon. The vectors thus derived do not capture the syntactic structure of the news headline, as only the occurrence of words are considered and not the positions.

[2] https://economictimes.indiatimes.com/
[3] https://www.moneycontrol.com/

We utilize the UB, UBT, LPS, rLPS, H-LPS and H-rLPS models discussed in Section 3.2 for our learning schemes. Further, we compare the performance of the above mentioned models with the BoW model utilizing Tf-Idf weighting scheme [Ramos et al., 2003], which is the most commonly used weighting scheme.

Therefore, overall we have seven different models on which we have conducted our experiments with different lexicon and classification algorithms. For dictionaries, we use Lexicon A and Lexicon B as discussed in Section 3.2. For extraction phase, we utilized two classification algorithms that are described below:

(i) GB Classifier: The implementation of this classifier is based on the XgBoost package [Chen and Guestrin, 2016].

(ii) SVC_L Classifier: The implementation of this classifier is based on Support Vector Classifier (SVC) subpackage which utilizes the "One-vs-One" multi-class classification strategy, in the scikit-learn package [Pedregosa et al., 2011]. We set the regularization parameter to 0.05 in order to ensure better generalization.

We split our analysis into two parts: first, in which the analysis is conducted on the news headlines with single-entities, and second, in which the learning schemes are implemented for news headlines containing single multiple-entities. In both the parts, we report the median performance of 31 runs of each learning scheme in the tables of results. A detailed description of the performance metrics utilized in this study is available in the Appendix.

5.1 Single-Entity

In the single-entity case, it is guaranteed that the financial sentiment extracted from the news headline is relevant to the entity present in the headline. Therefore, we compare the 28 learning schemes derived from seven models, two lexicons, and two classifiers using the single-entity news headlines dataset. Tables 4 - 5 report the performance of various learning schemes with Lexicon A and Lexicon B respectively.

We provide a visual representation of the average performance of all learning schemes in Figure-3. Best performing model across each classifier is represented using a marker. Firstly, we observe that all learning schemes derived from Lexicon B outperform the ones derived from Lexicon A, with an improvement of 7-9% in Accuracy and 9-12% in F1 Score. This indicates that the custom dictionary provides better feature annotations in regards to financial news and leads to an improvement during the word-level feature extraction stage across all models and classifiers.

Secondly, we observe that the N-gram based models (UB and UBT) perform better than the LPS based models and BoW models in most cases. With the SVC_L classifier, both N-gram based and LPS based models outperform the BoW model. With GB classifier, except for the models using the reduced form of LPS, all other models perform better than the BoW model with a difference of 0.5-2% in accuracy and F1 score. In the case of N-grams, increasing the model complexity does not lead to a comparable increase in performance (less than 0.5% improvement in Accuracy in moving from UB to UBT model across GB and SVC_L classifiers). In the case of LPS models, we note that the reduced form of LPS (rLPS) model does not perform better than the LPS model across all classifiers, indicating that the increase in model complexity driven by the number of word-level features has led to a reduction in LPS model performance. Using Huffman Coding for generating the bit sequences has led to a very small decrease (less than 1%) in Accuracy and F1 score when the H-LPS and LPS models are compared.

Thirdly, the GB classifier performs better than the SVC_L classifier in terms of both accuracy and F1 score across all learning schemes. With Lexicon B, the UBT model with GB classifier exhibits an average accuracy of 83.48% compared to 82.73% with SVC_L classifer. Overall, the average

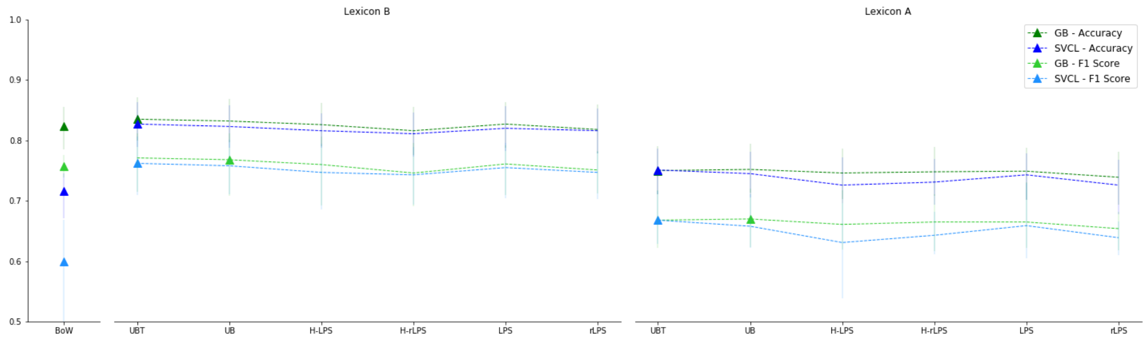

Figure 3: Comparison of various learning schemes on headlines with single entities. The models are trained on the entire News Headlines dataset, and the results indicate their performance on the news headlines with multiple-entities. The scores are measured as an average of the performance on buy, sell and hold classes

performance of GB classifier is higher than the SVC$_L$ classifier. This can be attributed to the ability of GB classifier to partition and effectively utilize the feature space. However, the performance differences are minimal and both family of classifiers are suitable for the sentiment extraction task.

5.2 Multiple-Entity

In the previous subsection, we have established that the Lexicon B provides best performance across all learning schemes and that the N-gram based models perform better than LPS based models in most cases with LPS models performing better than rLPS models. We have not been able to discount the relevance of Huffman Coding scheme, as the performance of the encoding scheme is dependent on the cardinality of the feature set. We therefore conduct our analysis for the multiple-entity news headlines using Lexicon B and since BoW model cannot capture the syntactic structure, we only compare the N-gram based models and LPS based models. The learning schemes are trained on the dataset of news headlines containing both single and multiple entities, however, the results reported in Table-6 show the performance of the learning schemes on the news headlines containing only multiple entities.

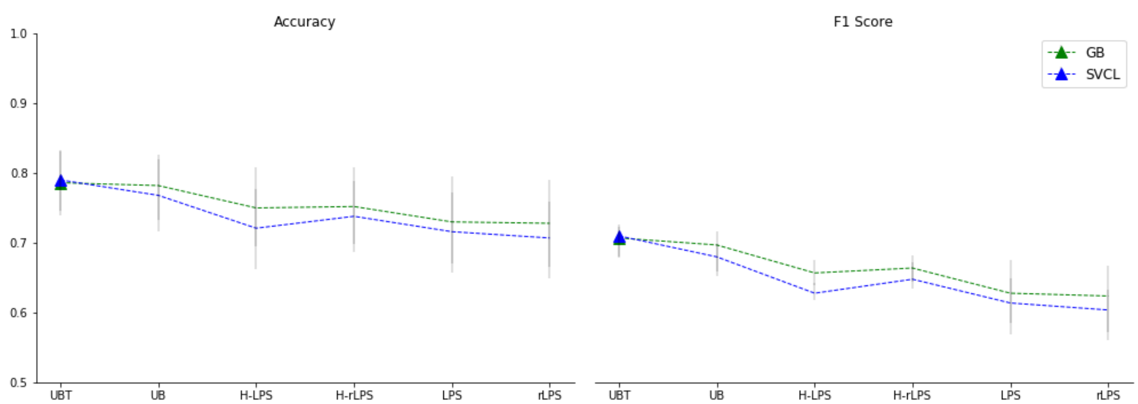

Figure 4: Comparison of various learning schemes on headlines with multiple entities. The scores are measured as an average of the performance on buy, sell and hold classes

As in the single entity case, we present the visual representation of the average performance of the learning schemes and denote the best performing models across each classifier with a marker in Figure-4. The best performing learning scheme is the UBT model and SVC_L classifier with an average Accuracy of 79.04% and F1 Score of 71.01%. Overall, the GB classifier with an average Accuracy of 75.47% and F1 score of 66.29% exhibits better performance than the SVC_L classifiers with an average Accuracy of 73.98% and F1 score of 64.75%. The N-gram based models perform better than the LPS models with GB and SVC_L classifiers. As in the single entity case, we observe that the performance improvements driven by increasing model complexity are insignificant, and in the case of LPS models leads to a reduction in performance. We observe that the Huffman Coding scheme leads to an improvement in performance of LPS models across learning schemes, which is an improvement driven by the increase in the collection of word-level features.

6 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 pricing mechanism constantly consumes the publicly available information related to the historical and expected future performance of the underlying asset. 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 [Lee and Chen, 2009, Chambers and Penman, 1984, 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, the news flow is utilized by High Frequency Traders in extremely short duration (orders of 10^{-9} to 10^{-6} seconds) to conduct directed trades [von Beschwitz et al., 2015]. However, it is of interest to identify whether such news flow contains any novel information that has economic value.

The most prominent approach to modelling 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 and Chen, 2001, Zhang et al., 2010]. We construct the 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 a entity in a certain duration. This construction takes into account the chance that the news source generates neutral sentiments. Based on the measure, we associate bearish and bullish situations to values less than and greater than 0 respectively.

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

$$s_{T,E} = \frac{B_{T,E} - S_{T,E}}{B_{T,E} + H_{T,E} + S_{T,E}}$$

where $B_{T,E}$, $H_{T,E}$, $S_{T,E}$ are the number of buy, hold, and sell recommendations relevant to E in the duration defined by T.

6.1 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. 5 the National Stock Exchange (NSE) is open for the duration 09:15 hrs to 15:30 hrs, during which there is a bi-directional information flow between the domestic market and domestic media. However, during the after-market hours, since the market is closed, the information flow is uni-directional 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. Further, we intend to capture the nature of the relationship over long periods, and hence, we consider a 6 year duration from 2012 to 2017.

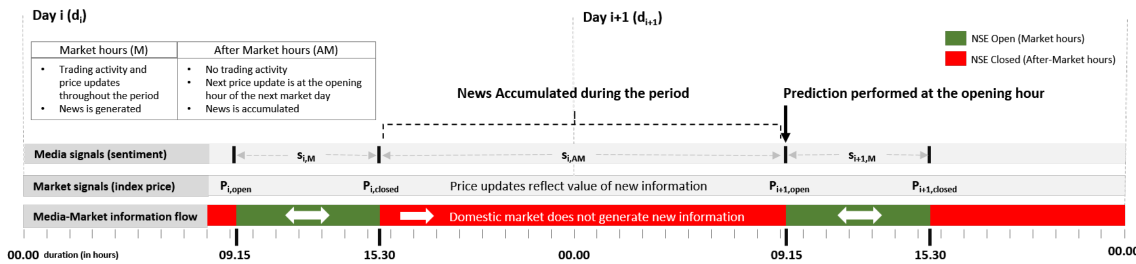

Figure 5: Diagram depicting the information exchange between markets - domestic and non-domestic and media. The duration marked with green colour indicates open market hours, and the duration marked in red indicates after-market hours. The information exchange is unidirectional from non-domestic markets to media during after-market hours

We utilize the best performing learning scheme - UBT model and GB classifier trained on Lexicon B as discussed in Section 5 to get the time-stamped financial 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 - 2017. The final sentiment scores are calculated for two durations :

- s_{iM} - market hours (09:30 AM, i^{th} day to 03:30 PM, i^{th} day)
- s_{iAM} - after-market hours (03:30 PM, i^{th} day to 09:30 AM, $i+1^{th}$ day)

We calculate the daily after-market log price returns as,

$$d_i = \log \frac{P^o_{i+1}}{P^c_i} = \log P^o_{i+1} - \log P^c_i$$

where P^o_i and P^c_i are opening and closing prices, respectively, on the i^{th} day

6.2 Hypothesis

We hypothesize that the sentiment information derived from the news flow in the after-market hours has a predictive value wrt the daily after-market price returns. We frame the hypothesis as below:
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 ($\beta \neq 0$)

6.3 NSE 500 index and news flow

1. The NSE 500 index is the broad equity market index based on free-float market capitalization tracking more than 90% of the market capitalization at any point of time. The index has a base value of 1000 established on 01 January, 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 01 January, 2012 to 31 December, 2017.

2. We capture the news headlines for the duration 01 January, 2012 to 31 December, 2017 relevant for the NSE 500 index, by utilizing the Named Entity Database containing more than 900 companies, various sectors and industries. We also capture the timestamp of the news headline, which is necessary to construct a duration-based sentiment measure. Overall, we have collected over 200,000 headlines relevant for the experiment, and extracted financial sentiment information utilizing the best performing model.

6.4 Results and Discussion

The results of the regressions are presented in Table 6.4. 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 years at a 10% significance level. Over the years 2013 to 2016, the relationship has a stronger significance with p-values below the 5% signficance level. We therefore 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.

Table 3: Regression Summary

Year	**p-value**
2012	0.0938*
2013	0.0422**
2014	0.0000**
2015	0.0383**
2016	0.0205**
2017	0.0577*

* - p-value < 0.1

** - p-value < 0.05

6.5 Dashboard

In the financial markets, investors, analysts and traders are faced with a large amount of information pertaining to thousands of companies, and hundreds of indices which are derived from clusters of sectors or companies. Cognitive limitations do not allow a human reader to analyze and track more than a small subset of the entities present and discussed in the financial markets. Our work has been guided by the idea that a system that can analyze large amounts of unstructured information flow in real-time can be effectively used by all market participants. We further our work by proposing a dashboard that presents the information contained in the news flow in a way that can be readily consumed by users - ranging from financial experts to general public. In the following subsections, we discuss our choice of the visualization elements, and the different views proposed to present the information related to indices, sectors and companies.

6.5.1 Choice of visualization elements

:

We choose the following visualization elements to present the information related to the variables discussed above:

1. *Stacked bar-graphs* to display sentiment relevant to an entity in a certain duration. The stacked graph shows number of buy recommendations as a positive deviation and number

of sell recommendations as a negative deviation from the normal. We calculate and present information on daily, monthly, quarterly, half-yearly and annual basis

2. *Treemaps* to display the weighted composition of various constituents (companies and sectors) of an index under consideration
3. *Wordcloud* to display the entity relevant financial concepts mentioned in the news headlines and their corresponding sentiments in the duration considered
4. Shades:
   (a) *Green* is used to indicate a bullish situation with respect to a certain entity in the treemap, a positive sentiment with respect to a financial concept, and for the buy recommendation in bar-graph, and
   (b) *Red* is used to indicate a bearish situation with respect a certain entity in the treemap, a negative sentiment with respect to a financial concept, and for the sell recommendation in bar-graph

6.5.2 Views

:
The dashboard is organized into two views - sector-based and company-based, giving users the ability to navigate through the information generated from news flow at various levels - indices at an aggregate level, and companies at a granular level. The dashboard contains a header box that presents aggregate information with respect to the index chosen - an indicator presenting the current state - bearish or bullish and a date and time stamp. Further, the header provides the user the choice of the views - sectoral or company based. The layout of both views contains a Treemap for displaying the composition of the index along with the current sentiment with respect to the sectors or companies. The information with respect to an individual company or sector is presented towards the right of the Treemap which is divided into two views - graphical view and textual view. The graphical view presents the sentiment information using a stacked bar-graph and gives users the choice to switch between various durations - daily to annual basis. The textual view provides the users the headlines (as hyperlinks) relevant to the duration, the decisions generated from each news headline. The users are also provided with an overview of the financial concepts discussed in the news in the duration considered with respect to the entity chosen, and their corresponding sentiments in the wordcloud.

In Figures 6 and 7, we present the weighted composition of companies relevant to the Indian equity *Nifty 50* index in terms of companies or sectors based on the view chosen and their current sentiments in the Treemap. The company *Reliance Industries* is chosen in the company-based view and the sector *Information Technology* is chosen in the sector-based view, for which the monthly sentiment is presented for the trailing 24 months. The headlines utilized for sentiment extraction are presented below in a tabular view indicating the date, the title and the decision inferred from each news headline. The wordcloud for *Reliance Industries* indicates that the company has positive revenue and profitability outlook, however the investor confidence and valuations are low, probably driven by low production and poor management. The current sentiment indicators designed as square boxes adjacent to the entity names, show that the *bearish* sentiment is exhibited towards *Nifty 50* and *Information Technology* and a *bullish* sentiment towards *Reliance Industries*.

7 Conclusions

In this paper, we further the work on finance focused sentiment analysis by proposing a system built over machine learning methods that can extract sentiments from news headlines and present it to

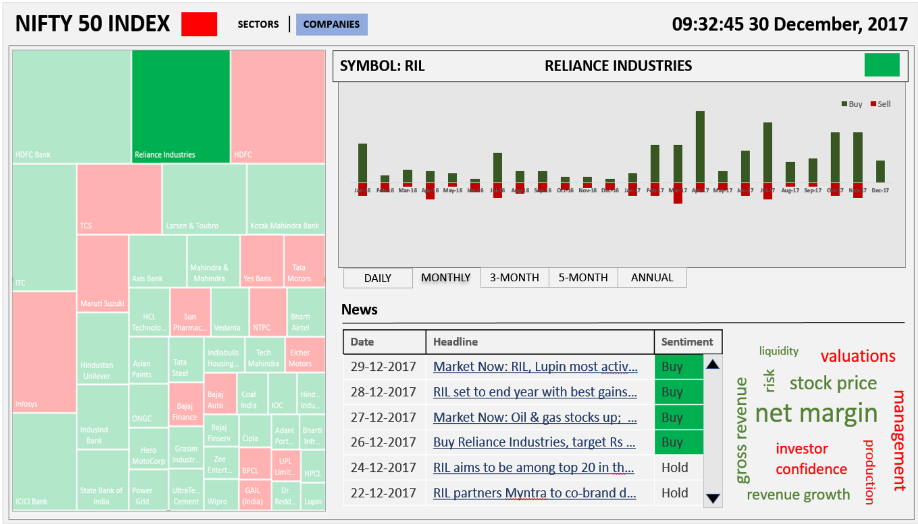

Figure 6: The company-based view of the dashboard

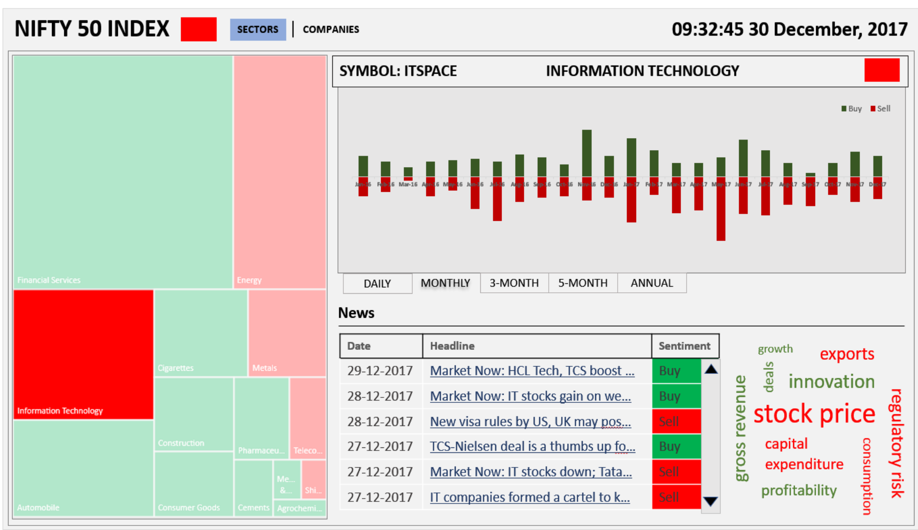

Figure 7: The sector-based view of the dashboard

the user in a visual form that can be readily consumed. Our proposed approach can work with news headlines containing multiple entities with conflicting sentiments and extract those sentiments with reasonably high accuracy.. We make two more major contributions - firstly, a phrase bank of over 12,000 news headlines relevant to an emerging economy annotated by multiple annotators for entities and their relevant financial sentiments and secondly, we provide an entity database of various companies, sector and indices relevant to the Indian economy and their forms of appearance in the news headlines amounting to over 5,000 phrases. We also propose various modifications to the well used Loughran and McDonald financial polarity lexicon, that in our experiments have shown to improve the performance of underlying machine learning models.

We propose a measure for information content in the news flow and construct an experiment to measure the predictive value of the information contained in the news flow. Based on the data obtained in the duration 2012 to 2017, we establish the predictive value of news flow. Based on the proposed framework, we develop a dashboard that enables users to access information derived from news flow at various levels.

Table 4: Single entity results with Lexicon-A

		GB Classifier						SVC_L Classifier					
		UBT	UB	H-LPS	H-rLPS	LPS	rLPS	UBT	UB	H-LPS	H-rLPS	LPS	rLPS
Accuracy													
	Buy	74.79%	74.96%	74.79%	74.75%	**75.66%**	74.23%	**75.59%**	74.66%	70.95%	73.11%	74.84%	73.31%
	Hold	71.14%	**71.34%**	70.38%	70.79%	70.16%	69.30%	**71.00%**	70.58%	69.59%	69.29%	70.10%	67.74%
	Sell	79.10%	**79.42%**	78.71%	78.85%	78.77%	78.07%	**78.60%**	78.13%	77.19%	76.84%	77.83%	76.81%
Precision													
	Buy	72.74%	72.20%	**75.51%**	71.06%	74.15%	70.79%	75.62%	**75.66%**	61.37%	71.43%	71.03%	72.22%
	Hold	59.15%	**59.66%**	57.68%	59.46%	56.07%	56.42%	58.69%	58.20%	**63.14%**	57.30%	57.80%	54.61%
	Sell	69.54%	69.94%	69.16%	69.15%	**73.26%**	72.01%	67.78%	67.42%	67.70%	65.95%	**72.79%**	69.23%
Recall													
	Buy	61.65%	62.70%	56.92%	**63.05%**	56.78%	61.65%	58.49%	56.22%	**72.15%**	57.09%	58.46%	57.62%
	Hold	65.50%	65.13%	67.34%	64.02%	**69.52%**	68.63%	66.97%	67.34%	46.49%	65.87%	66.18%	**69.37%**
	Sell	74.11%	74.74%	**75.37%**	73.49%	73.49%	64.72%	**75.99%**	75.16%	72.23%	71.19%	72.86%	63.26%
F1-score													
	Buy	66.48%	66.73%	64.69%	**66.79%**	64.19%	66.16%	**65.82%**	64.04%	65.76%	63.50%	64.33%	64.11%
	Hold	62.18%	**62.36%**	61.95%	61.61%	62.18%	61.79%	**62.84%**	62.32%	53.96%	61.15%	60.39%	61.02%
	Sell	71.69%	72.03%	71.76%	70.99%	**73.07%**	68.22%	71.65%	71.18%	69.70%	68.18%	**72.87%**	66.52%

Table 5: Single entity results with Lexicon-B and BoW model

		GB Classifier							SVC_L Classifier						
		BoW	UBT	UB	H-LPS	H-rLPS	LPS	rLPS	BoW	UBT	UB	H-LPS	H-rLPS	LPS	rLPS
Accuracy															
	Buy	82.77%	**83.39%**	83.11%	82.64%	81.69%	82.98%	81.68%	73.23%	**82.85%**	82.28%	81.61%	81.33%	82.10%	81.67%
	Hold	78.53%	**79.97%**	79.69%	79.07%	77.57%	78.74%	77.90%	67.10%	**78.93%**	78.83%	78.87%	77.25%	78.28%	77.85%
	Sell	85.59%	**87.09%**	86.86%	86.15%	85.50%	86.31%	85.89%	74.60%	**86.42%**	85.82%	84.46%	84.59%	85.71%	85.32%
Precision															
	Buy	**81.70%**	78.76%	78.51%	77.36%	79.46%	78.92%	79.07%	70.46%	80.86%	80.95%	77.70%	**81.55%**	79.73%	78.24%
	Hold	64.88%	72.26%	72.36%	**73.17%**	67.44%	68.60%	65.88%	53.13%	69.23%	68.88%	**74.26%**	66.96%	68.85%	66.78%
	Sell	**89.22%**	80.04%	79.22%	77.36%	77.41%	82.19%	82.98%	**96.82%**	79.07%	77.82%	72.19%	75.00%	78.19%	80.67%
Recall															
	Buy	73.99%	78.63%	78.98%	**79.16%**	71.80%	76.71%	72.85%	62.72%	73.20%	71.10%	**74.08%**	68.83%	73.91%	72.85%
	Hold	**85.79%**	70.66%	69.37%	64.94%	70.30%	74.54%	77.68%	**91.46%**	73.25%	73.80%	63.47%	71.77%	71.77%	74.54%
	Sell	65.96%	83.30%	83.09%	**84.76%**	82.25%	76.62%	74.32%	30.94%	82.05%	82.88%	**87.68%**	83.30%	81.21%	76.83%
F1-score															
	Buy	77.53%	**78.57%**	78.49%	78.10%	75.14%	78.10%	75.82%	66.12%	**76.98%**	76.15%	75.95%	74.67%	76.51%	75.64%
	Hold	**73.92%**	71.46%	70.86%	69.14%	69.09%	71.02%	71.31%	66.80%	70.97%	**71.12%**	68.57%	69.34%	70.37%	70.34%
	Sell	75.64%	**81.20%**	81.14%	80.63%	79.67%	79.17%	78.32%	46.82%	**80.54%**	80.04%	79.46%	78.85%	79.59%	78.10%

Table 6: Multiple entity results with Lexicon-B

		GB Classifier						SVC_L Classifier					
		UBT	UB	H-LPS	H-rLPS	LPS	rLPS	UBT	UB	H-LPS	H-rLPS	LPS	rLPS
Accuracy													
	Buy	78.64%	**78.72%**	74.74%	74.93%	72.32%	72.87%	**79.45%**	76.79%	72.36%	73.88%	71.81%	71.19%
	Hold	**74.00%**	73.32%	69.45%	69.78%	67.11%	66.58%	**74.56%**	71.70%	66.14%	68.63%	65.78%	64.93%
	Sell	**83.21%**	82.58%	80.85%	80.88%	79.48%	78.93%	**83.12%**	81.86%	77.66%	78.80%	77.13%	75.94%
Precision													
	Buy	72.48%	**72.73%**	66.18%	67.55%	66.83%	68.82%	**72.90%**	70.62%	65.20%	66.02%	66.58%	66.58%
	Hold	**71.35%**	69.08%	66.86%	67.17%	61.58%	60.64%	**71.45%**	68.71%	64.10%	67.92%	61.84%	60.31%
	Sell	**68.24%**	67.17%	63.92%	64.26%	67.40%	65.84%	**68.50%**	65.88%	57.57%	58.86%	59.33%	56.32%
Recall													
	Buy	**72.15%**	70.43%	68.90%	69.04%	58.35%	58.64%	**72.44%**	69.82%	63.41%	68.30%	58.94%	57.77%
	Hold	71.76%	73.78%	64.71%	65.61%	74.72%	**75.14%**	**73.01%**	70.04%	59.34%	59.86%	68.25%	66.61%
	Sell	**67.55%**	63.57%	63.76%	63.76%	51.51%	50.66%	66.10%	64.36%	67.60%	**69.55%**	54.66%	55.74%
F1-score													
	Buy	**72.31%**	71.63%	67.64%	68.15%	62.37%	63.21%	**72.67%**	69.79%	64.24%	67.25%	62.44%	61.89%
	Hold	**71.86%**	71.61%	65.56%	66.55%	67.61%	66.72%	**72.55%**	68.94%	61.80%	63.40%	64.91%	63.31%
	Sell	**68.04%**	65.93%	64.01%	64.40%	58.53%	57.14%	**67.81%**	65.19%	62.48%	63.79%	56.94%	56.07%

Appendix

The performance metrics used to compare different learning schemes are as follows:

(i) *Accuracy* measures the predictive ability of the learning scheme, calculated as the proportion of *identified* true data points (positives and negatives) among all data points considered.

(ii) *Precision* measures the ability of the learning scheme to identify the true positives among all the *identified* true data points.

(iii) *Recall* measures the ability of the learning scheme to identify the true positives among all the *defined* true data points.

(iv) *F1 Score* is the harmonic mean of *precision* and *recall*. It is a composite measure of the ability of learning scheme to identify true positives among both *identified* and *defined* true data points.

Table 7: Performance Matrix

Measure	**Formula**	**Notation**
Accuracy (A)	$\frac{TP+TN}{TP+FP+TN+FN}$	TP - True Positives
Precision (P)	$\frac{TP}{TP+FP}$	FP - False Positives
Recall (R)	$\frac{TP}{TP+FN}$	TN - True Negatives
F1 Score (F1)	$\frac{2*Precision*Recall}{Precision+Recall}$	FN - False Negatives

References

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

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

[Balahur et al., 2013] Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. (2013). Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202*.

[Bollen et al., 2011] Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

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

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

[Cambria et al., 2014] Cambria, E., Olsher, D., and Rajagopal, D. (2014). SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *Twenty-eighth AAAI conference on Artificial Intelligence*.

[Cavnar et al., 1994] Cavnar, W. B., Trenkle, J. M., et al. (1994). N-gram-based text categorization. *Ann Arbor Mi*, 48113(2):161–175.

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

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[Culotta and Sorensen, 2004] Culotta, A. and Sorensen, J. (2004). Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423. Association for Computational Linguistics.

[Das and Chen, 2001] Das, S. and Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, volume 35, page 43. Bangkok, Thailand.

[Deng and Wiebe, 2015] Deng, L. and 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*, pages 1323–1328.

[Finkel et al., 2005] Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

[Friedman, 2001] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

[Gerlein et al., 2016] Gerlein, E. A., McGinnity, M., Belatreche, A., and Coleman, S. (2016). Evaluating machine learning classification for financial trading: An empirical approach. *Expert Systems with Applications*, 54:193–207.

[Guelman, 2012] Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, 39(3):3659–3667.

[Haddi et al., 2013] Haddi, E., Liu, X., and Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17:26–32.

[Hu and Liu, 2004] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177. ACM.

[Huffman, 1952] Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101.

[Hussain and Prieto, 2016] Hussain, K. and Prieto, E. (2016). *Big Data in the Finance and Insurance Sectors*, pages 209–223. Springer International Publishing, Cham.

[Kearns and Nevmyvaka, 2013] Kearns, M. and Nevmyvaka, Y. (2013). Machine learning for market microstructure and High Frequency Trading. *High Frequency Trading: New Realities for Traders, Markets, and Regulators*.

[Khandani et al., 2010] Khandani, A. E., Kim, A. J., and Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787.

[Klein and Manning, 2004] Klein, D. and Manning, C. D. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 478. Association for Computational Linguistics.

[Lee and Chen, 2009] Lee, R. P. and 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.

[Liu, 2012] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

[Loughran and McDonald, 2009] Loughran, T. and McDonald, B. (2009). When is a Liability not a Liability? *Journal of Finance*.

[Malo et al., 2014] Malo, P., Sinha, A., Korhonen, P., Wallenius, J., and 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 et al., 2013] Malo, P., Sinha, A., Takala, P., Ahlgren, O., and 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 (ICDMW)*, pages 945–954. IEEE.

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

[Mao et al., 2012] Mao, Y., Wei, W., Wang, B., and Liu, B. (2012). Correlating S&P 500 stocks with Twitter data. In *Proceedings of the first ACM International Workshop on hot topics on interdisciplinary Social Networks research*, pages 69–72. ACM.

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

[Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical methods in Natural Language Processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

[Park et al., 2013] Park, J., Konana, P., Gu, B., Kumar, A., and Raghunathan, R. (2013). Information valuation and confirmation bias in virtual communities: Evidence from stock message boards. *Information Systems Research*, 24(4):1050–1067.

[Pedregosa et al., 2011] 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., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

[Plakandaras et al., 2015] Plakandaras, V., Gupta, R., Gogas, P., and Papadimitriou, T. (2015). Forecasting the US real house price index. *Economic Modelling*, 45:259–267.

[Ramos et al., 2003] Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first Instructional Conference on Machine Learning*, volume 242, pages 133–142.

[Ranco et al., 2015] Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., and Mozetič, I. (2015). The effects of Twitter sentiment on stock price returns. *PLOS ONE*, 10(9):e0138441.

[Riloff and Wiebe, 2003] Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in Natural Language Processing*, pages 105–112. Association for Computational Linguistics.

[Salampasis et al., 2017] Salampasis, D., Mention, A.-L., and Kaiser, A. O. (2017). Wealth Management in Times of Robo: Towards Hybrid Human-Machine Interactions.

[Shah, 2007] Shah, V. H. (2007). Machine learning techniques for stock prediction. *Foundations of Machine Learning| Spring*, pages 1–19.

[Stone et al., 1966] Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

[Šubelj et al., 2011] Šubelj, L., Furlan, Š., and Bajec, M. (2011). An expert system for detecting automobile insurance fraud using social network analysis. *Expert Systems with Applications*, 38(1):1039–1052.

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

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

[Tsai and Chen, 2010] Tsai, C.-F. and Chen, M.-L. (2010). Credit rating by hybrid machine learning techniques. *Applied soft computing*, 10(2):374–380.

[Vapnik, 2013] Vapnik, V. (2013). *The nature of statistical learning theory.* Springer science & business media.

[Viaene et al., 2005] Viaene, S., Dedene, G., and Derrig, R. A. (2005). Auto claim fraud detection using Bayesian learning neural networks. *Expert Systems with Applications*, 29(3):653–666.

[von Beschwitz et al., 2015] von Beschwitz, B., Keim, D., and Massa, M. (2015). First to 'read' the news: News analytics and high frequency trading.

[Warner et al., 1988] Warner, J. B., Watts, R. L., and Wruck, K. H. (1988). Stock prices and top management changes. *Journal of financial Economics*, 20:461–492.

[Wiebe et al., 2005] Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

[Wilson et al., 2005] Wilson, T., Wiebe, J., and 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*, pages 347–354. Association for Computational Linguistics.

[Wysocki, 1998] Wysocki, P. (1998). Cheap talk on the web: The determinants of postings on stock message boards.

[Zhang et al., 2010] Zhang, W., Skiena, S., et al. (2010). Trading Strategies to Exploit Blog and News Sentiment. In *ICWSM*.

[Zhang et al., 2011] Zhang, X., Fuehres, H., and Gloor, P. A. (2011). Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences*, 26:55–62.