공학박사학위논문

Financial Risk Assessment Automation: Hot Topic Detection in Speeches, Sentiment Analysis of News Articles, and Spam Filtering on Twitter

금융위험 평가 자동화: 연설문 핫토픽 탐지, 뉴스 감성 분석 및 트위터 스팸 필터링

2023 년 8 월

서울대학교 대학원

산업공학과

박 지 혜

Financial Risk Assessment Automation: Hot Topic Detection in Speeches, Sentiment Analysis of News Articles, and Spam Filtering on Twitter

금융위험 평가 자동화: 연설문 핫토픽 탐지, 뉴스 감성 분석 및 트위터 스팸 필터링

지도교수 조 성 준

이 논문을 공학박사 학위논문으로 제출함

2023 년 7 월

서울대학교 대학원

산업공학과

박 지 혜

박지혜의 공학박사 학위논문을 인준함

2023 년 7 월

위 원 장 ______윤 명 환______ (인)

부위원장 ______조 성 준______ (인)

위 원 ______이 재 욱______ (인)

위 원 ______이 영 훈______ (인)

위 원 ______김 은 지______ (인)

Abstract

Financial Risk Assessment Automation: Hot Topic Detection in Speeches, Sentiment Analysis of News Articles, and Spam Filtering on Twitter

Jihye Park
Department of Industrial Engineering
The Graduate School
Seoul National University

Text mining refers to the process of extracting interesting and significant information from textual data. It encompasses the process of performing various tasks such as hot topic detection, sentiment analysis, and spam filtering using a variety of text analysis tools including purpose-built frameworks, sentiment lexicons, and pretrained language models. Due to its broad applicability, text mining has been widely used to support decision making in various domains including politics, economics, and society. Especially, numerous researchers have attempted to assess financial risk by applying text mining techniques to financial texts. Since text mining-based approaches are less expensive in terms of time, human labor, and domain expertise than manual approaches, text mining enables real-time risk assessment that requires prompt detection of rapid changes in the financial domain. Most previous studies on financial text mining have directly applied general-purpose text analysis tools to financial texts. However, financial texts exhibit several linguistic characteristics that are distinct from those of general domain texts. Although several researchers have attempted to incorporate domain specificity of financial texts into text analysis tools, the detection of lexical items that play a crucial role in automated financial risk assessment has not been discussed sufficiently.

In this dissertation, financial domain-specific text analysis tools that can detect hot topics, sentiment words, and spam messages, respectively, are proposed. The proposed tools would contribute to the automation of financial risk assessment by supporting early warning, explainable market sentiment analysis of news articles, and spam filtering on real-time data feeds, respectively. First, a hot topic detection framework that incorporates the temporal importance of keywords is proposed. The framework is applied to speeches made by the chairs of a central bank, showing the possibility of text mining-based early warning. Second, an automatically constructed sentiment lexicon addressing the financial ontology that the sentiment of a word may change depending on the presence of directional expressions is proposed. The lexicon is applied to benchmark datasets regarding economic news headlines, demonstrating the explainability of the market sentiment analysis process. Third, company-related knowledge-enhanced language models are proposed to detect spam messages that promote non-blue-chip stocks as if they are blue-chip stocks. Specifically, a framework that uses corporate reports as a textual knowledge base is proposed to enhance factual knowledge of the model. The framework employs a novel company name masking method, which masks tokens associated with company names, allowing the model to learn company-related factual information in a sentence. The spam filtering performance of language models built through the proposed framework is validated using Twitter benchmark datasets to demonstrate the viability of automatic spam filtering for real-time data feeds to automated systems.

Keywords: Text mining, Financial text, Financial risk assessment, Automated system, Hot topic detection, Sentiment analysis, Spam filtering, Central banker's speech, Economic news article, Corporate report, Twitter

Student Number: 2018-28881

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Text mining is defined as the process of extracting interesting and significant information from textual data (Talib et al., 2016). As illustrated in Fig. 1.1, text mining encompasses the process of performing various tasks such as hot topic detection, sentiment analysis, and spam filtering using a variety of text analysis tools including purpose-built frameworks, sentiment lexicons, and pretrained language models (De Fortuny et al., 2012; Gupta et al., 2009; Johnsi et al., 2022; Macêdo et al., 2022; Zhao et al., 2021). Due to its broad applicability, text mining has been widely used to support decision making in various domains including politics, economics, and society (Hassani et al., 2020).

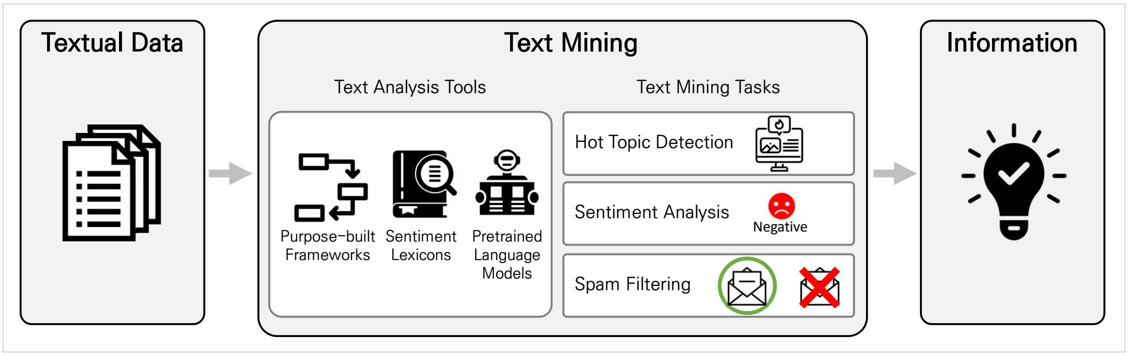


Figure 1.1: Illustration of text mining procedures

The financial domain is one of the domains where text mining is actively applied. Extensive studies have applied text mining techniques to financial texts for the purpose of solving financial problems. Financial texts broadly refer to texts that affect financial markets, primarily those issued by central banks, news organizations, companies, and investors (Daudert, 2021; Gupta et al., 2020; Man et al., 2019). Examples of financial texts include central bankers' speeches, economic news articles, corporate reports, and stock-related microblogs. Owing to the proliferation of posts on the Internet and heightened demand for market transparency, financial texts have become more readily available (Chan & Chong, 2017). As the number of financial texts increases, text mining techniques that can extract useful information hidden in plain sight from numerous pages have garnered significant attention from both academia and industry (Chan & Chong, 2017). To date, extensive research has employed text mining techniques to predict stock prices (Jaggi et al., 2021; Picasso et al., 2019; Zhang et al., 2022a); detect fraudulent activity by management (Craja et al., 2020; Krishnan et al., 2022; Maurya et al., 2022); and conduct other financial applications (Kumar & Ravi, 2016; Malandri et al., 2018; Xing et al., 2019).

Especially, numerous researchers have attempted to assess financial risk by applying text mining techniques to financial texts. Financial risk broadly refers to uncertainties associated with capital losses of stakeholders such as banks, companies, and investors (Horcher, 2011; Li et al., 2022; Peng et al., 2009). Since text mining-based approaches are less expensive in terms of time, human labor, and domain expertise than manual approaches, text mining enables real-time risk assessment that requires prompt detection of rapid changes in the financial domain. Fig. 1.2 illustrates text mining-based financial risk assessment procedures. An automated financial risk assessment system that applies text mining techniques to financial texts issued by central banks, media, corporations, and investors would enable a better understanding of rapidly evolving aspects of the financial domain. Automatically detected hot topics would assist in early warning of financial crises, thus helping financial experts' decision making process; explainable market sentiment analysis of news articles would expand understanding of the current collective assessment of financial markets; and automatic spam filtering would help produce more accurate analysis results by improving the quality of data generated in real time on numerous social media platforms.

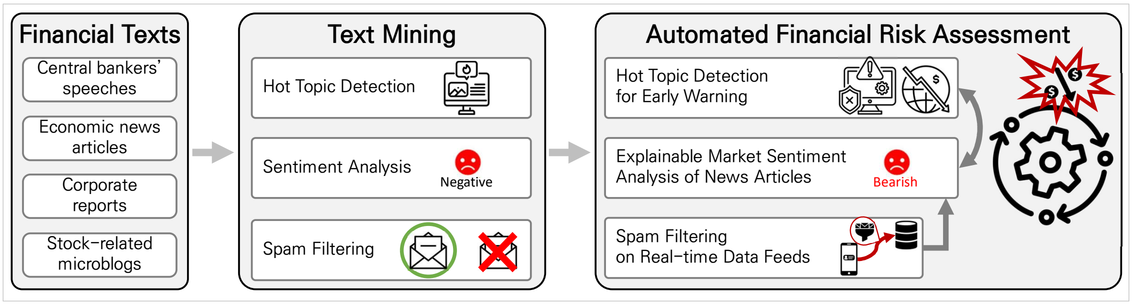

Financial Texts
Central bankers' speeches
Economic news articles
Corporate reports
Stock–related microblogs
Text Mining
Hot Topic Detection
Sentiment Analysis
Negative
Spam Filtering
Automated Financial Risk Assessment
Hot Topic Detection for Early Warning
Explainable Market Sentiment Analysis of News Articles
Bearish
Spam Filtering on Real–time Data Feeds

Figure 1.2: Illustration of text mining-based financial risk assessment procedures

Most previous studies on financial text mining have directly applied general-purpose text analysis tools to financial texts. However, financial texts exhibit several linguistic characteristics that are distinct from those of general domain texts (Montariol et al., 2020). First, financial texts, along with specialized vocabulary, feature a word distribution that is different from that of the general domain. The term "EBIT," which is an abbreviation for "earnings before interest and taxes," is a typically used term in the financial domain yet rarely used in the general domain. As a similar example, words that indicate directions, such as "increase" or "decrease," are more frequently used in financial texts than in general domain texts. Second, in some cases, the meanings and/or sentiments of words in the financial context are different from those in the general context. A classic example is the terms "bear" and "bull." Generally, these two words refer to animals and imply neutral sentiments. In the financial context, however, "bear" is an extremely negative word indicating decreasing prices (as in a "bear market"), whereas "bull" is a highly positive word indicating increasing prices (as in a "bull market"). As another example, the term "liability" is typically associated with a negative connotation in general discourse; however, when it is used in the financial domain, a neutral sentiment is conveyed (Cortis et al., 2017).

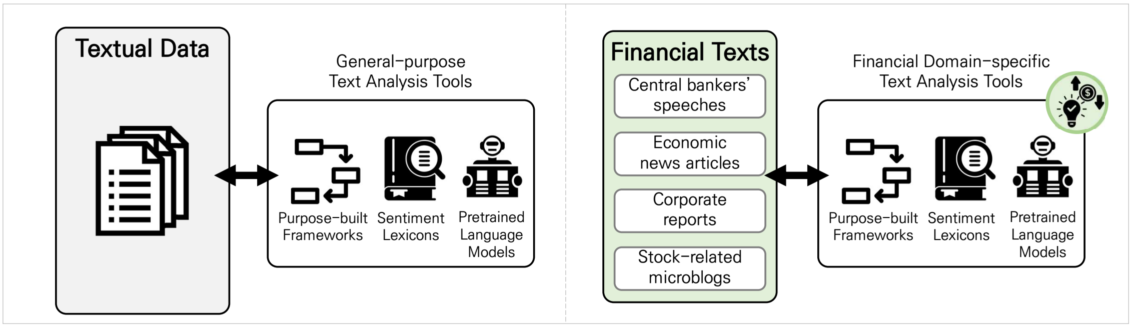


Figure 1.3: Comparison between general-purpose and financial domain-specific text analysis tools

To address the issues associated with these linguistic characteristics, a few studies have been conducted to develop financial domain-specific pretrained language models (Araci, 2019; Loukas et al., 2022; Yang et al., 2020), sentiment lexicons (Loughran & McDonald, 2011; Moreno-Ortiz et al., 2020; Yekrangi & Abdolvand, 2021), and other text analysis tools that reflect the domain specificity of financial texts (Ko et al., 2020; Li et al., 2020b; Masawi et al., 2018). Extensive research has shown that these financial domain-specific text analysis tools can improve the performance of financial applications over general-purpose tools (Liu et al., 2021). Fig. 1.3 compares general-purpose text analysis tools with financial domain-specific text analysis tools. Financial domain-specific text analysis tools are trained on financial texts, whereas most general-purpose text analysis tools are trained on general domain corpora such as Wikipedia. The text analysis tools that reflect the domain specificity of financial texts would offer more insightful analysis results than general-purpose text analysis tools. To this end, it is necessary to collect financial texts and analyze the domain-specific characteristics.

However, the detection of lexical items that play a crucial role in automated financial risk assessment has not been discussed sufficiently. In this dissertation, financial domain-specific text analysis tools that can detect hot topics, sentiment words, and spam messages, respectively, are proposed. The proposed tools are anticipated to contribute to the automation of financial risk assessment by supporting early warning, explainable market sentiment analysis, and stock-related spam filtering for real-time data feeds, respectively. The text analysis tools, detection targets, and validations/applications covered in this dissertation are summarized in Table 1.1.

Table 1.1: Text analysis tools, detection targets, and validations/applications covered in this dissertation

Chapter	Text analysis tools	Detection targets	Validations/applications
Chapter 3	Hot topic detection framework	Hot topics	Early warning using central bankers' speeches
Chapter 4	Sentiment lexicon	Sentiment words	Explainable market sentiment analysis of news articles
Chapter 5	Pretrained language model	Spam messages	Stock-related spam tweet filtering on real-time data feeds

First, a hot topic detection framework that incorporates the temporal importance of keywords is proposed. Identifying probable risks is an essential part of financial risk assessment (Ott, 2020), and hot topic detection helps to preemptively identify potential risk factors that may lead to a financial crisis. While it is almost impossible to predict the exact circumstances that will trigger a crisis (Bezemer, 2010), it is speculated that lexical items frequently appearing in textual data issued by financial officials, such as central bankers, would possess the potential to be risk factors. The sentence shown below (Bernanke, 2007b), which was announced by Ben Shalom Bernanke in February 2007, contains the term "subprime mortgages," which is one of the major risk factors for the 2008 global financial crisis.

> "The exception is **subprime mortgages** with variable interest rates, for which delinquency rates have increased appreciably."

To proactively detect and consistently track these risk factors that are challenging to discover using quantitative models (Gaytán et al., 2002), qualitative monitoring is performed by a small number of domain experts who read and summarize documents published by global financial institutions (Koyuncugil & Ozgulbas, 2012). However, manual analysis of documents is very expensive as the process requires a significant amount of time, human labor, and background knowledge. To solve the issues arising from manual inspection, a text mining framework that automatically detects potential risk factors is proposed. The framework uses a novel light-weight unsupervised keyword-scoring method, which treats bigrams as keywords and incorporates the temporal importance of keywords. This is done by estimating the growth rate of the term frequency. The framework is applied to speeches made by the chairs of the Federal Reserve System, showing the viability of text mining-based early warning.

Second, an automatically constructed financial ontology-aware sentiment lexicon is proposed. Explainability—the degree to which that an interested stakeholder can comprehend the main drivers of a model-driven decision (Bracke et al., 2019; Bussmann et al., 2021)—is an essential consideration for financial risk assessment (Mashrur et al., 2020);

and a sentiment lexicon is regarded as one of the tools that can provide a clear explanation of the market sentiment analysis process (Brazdil et al., 2022). One of the challenges in the construction of a financial sentiment lexicon is the existence of the domain-specific ontology that the sentiment orientation of a word may change significantly depending on the presence of directional expressions (Krishnamoorthy, 2018; Malo et al., 2014; Moreno-Ortiz et al., 2020). The sentence shown below (Sheetz, 2019), which was issued by a news organization in July 2019, contains the phrase "earnings dropped," which indicates negative sentiment.

"Boeing's **earnings dropped** nearly 275% from the same quarter last year."

The term "earnings" typically conveys a positive sentiment; however, when the word is juxtaposed with "dropped" to form the phrase "earnings dropped," the associated sentiment is negative. These direction-dependent words, such as "earnings," "profits," or "operating losses," are widely used in corporate memorandums, analyst reports, and news articles analyzing the financial market. They are important contextual words that affect sentiment analysis. In this context, a data-driven method to directly extract direction-dependent words is proposed. The proposed method addresses the financial ontology regarding direction-dependent words by estimating the degree of association between given words and their direction-dependency types. The sentiment lexicon constructed through the proposed method is applied to benchmark datasets regarding economic news headlines. Experimental results demonstrate that the proposed sentiment lexicon can achieve both explainability and reasonable performance.

Third, company-related knowledge-enhanced language models are proposed to detect stock-related spam messages. *Cashtag piggybacking*, which refers to malicious practices that attempt to promote low-value stocks by exploiting the popularity of high-value ones, is one of the most typically observed types of stock-related spam messages. Understandably, *cashtag piggybacking* practices must be preemptively filtered when supplying real-time data to systems for automated financial risk assessment. The tweet shown below, which was issued in January 2022, is an example of *cashtag piggybacking.*

> "The best investments over the last 30 years have had multiple drawdowns of 60-90% along the way to producing huge profits. **$AAPL $AMZN**. Why would **$LUNA** be any different?"

In this tweet, $AAPL—which indicates the stock of the company "Apple Inc." and is regarded as a blue-chip stock—and $LUNA—which indicates the "Terra Luna" coin—are mentioned as if they share similar characteristics. However, the Terra Luna coin is not a blue-chip stock, and its exchange value exhausted most of its $60 billion worth of investments in May 2022 (Briola et al., 2022; Kyosev et al., 2022). These spam messages would be accurately detected through manual spam filtering, but manual review cannot be performed in real time due to the cost involved in the investigation and the huge amount of data. In this context, a framework for constructing a company-related knowledge-enhanced language model is proposed to automatically filter spam messages. In the framework, corporate reports are used as a textual knowledge base. Additionally, the framework applies a novel masking method that masks tokens corresponding to company names and forces the model to predict the masked tokens using representations of other unmasked tokens in the sentence. This allows the model to learn the context in which a company name appears, thereby enhancing company-related knowledge. The language model built through the proposed framework is applied to benchmark datasets for stock-related microblogs. Experimental results show that the proposed model better detects spam messages than other financial domain-specific pretrained language models. Hence, the proposed method is anticipated to contribute to the real-time delivery of data to systems for automated financial risk assessment.

The remainder of this dissertation is organized as follows. Chapter 2 presents literature reviews on text mining techniques covered in this dissertation. Related studies regarding financial risk assessment automation are also provided. In Chapter 3, a hot topic detection framework for early warning is proposed. In Chapter 4, a financial ontology-aware sentiment lexicon for explainable market sentiment analysis is proposed. In Chapter 5, company-related factual knowledge-enhanced language models for stock-related spam filtering are proposed. Finally, the conclusions and directions for future research are discussed in Chapter 6.

Chapter 2

Literature Review

2.1 Text Mining Techniques

2.1.1 Hot topic detection methods

A “hot topic” is defined as a topic that appears frequently over a period of time (Bun & Ishizuka, 2002). To date, numerous researchers have used keyword extraction and topic modeling methods to detect hot topics.

Term frequency-inverse document frequency (TF-IDF) (Salton & Buckley, 1988) is one of the most common choices for keyword extraction. TF-IDF score of a word w in the document i is computed as follows:

$$\text{TF-IDF}(w, i) = \frac{Frequency(w, i)}{log(N/n_w)} \tag{2.1}$$

where $Frequency(w, i)$ is the frequency of the word w in document d_i, N is total number of documents, and n_w is the number of documents containing the word w.

TextRank (Mihalcea & Tarau, 2004) uses the relationship of vocabulary to sort the subsequent keywords based on the co-occurrence window. Fig. 2.1 shows a sample graph for keyword extraction by TextRank.

Yet Another Keyword Extractor (YAKE) (Campos et al., 2020) relies on statistical

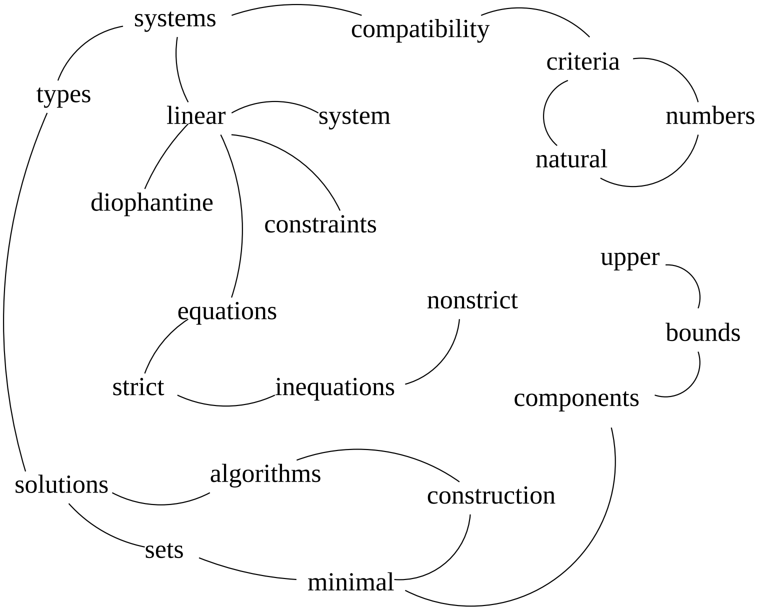


Figure 2.1: Sample graph for keyword extraction by TextRank (Mihalcea & Tarau, 2004)

features extracted from a given document to select the most representative keywords of the document. It uses statistical features regarding structure, term frequencies, and co-occurrence.

KeyBERT (Grootendorst, 2020), which is an abbreviation for "Keyword extraction with Bidirectional Encoder Representations from Transformers (BERT)," is a deep learning-based approach to extract keywords. KeyBERT uses contextual features from bidirectional transformers to automatically extract keywords from a given corpus. Specifically, BERT is used to extract word embeddings and document embeddings. To determine which words or phrases are most similar to the document, KeyBERT calculates the cosine similarity between them. Then, the words that best describe the entire document are determined by their similarity score.

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is one of the most commonly used topic modeling methods. LDA assumes that each document can be expressed as a probabilistic distribution of latent topics. That is, the document is treated as a mixture of latent topics, and the latent topic is treated as a probability distribution of a set of words.

Meanwhile, dynamic topic models (Blei & Lafferty, 2006; Lee & Seung, 1999) have been used to investigate the evolution of topics. For example, Greene & Cross (2017) used a dynamic topic model to analyze how the policy agenda on climate change in political speeches has evolved over time. However, dynamic topic models are not suitable for preemptively detecting rapidly emerging words.

2.1.2 Sentiment lexicons

A sentiment lexicon is a list of words or phrases mapped to positive or negative sentiment labels. Multi-perspective question answering opinion corpus (MPQA) (Wilson et al., 2005) contains 2,718 positive words and 4,911 negative words. SentiWordNet (SWN) (Baccianella et al., 2010) uses 117,659 terms with scores for positivity, negativity, and objectivity ranging from -1 to 1. Semantic orientation calculator (SO-CAL) (Taboada et al., 2011) uses 6,395 terms with discrete sentiment-scores ranging from -5 to 5. A lexicon developed by Finn Arup Nielsen (AFINN) (Nielsen, 2011) uses a lexicon containing 2,477 words with scores between -5 and 5. SWN, SO-CAL, and AFINN return polarity scores for individual sentences by subtracting the sums of negative scores from the positive scores of the words. Sentiment140 (Mohammad et al., 2013) uses 43,431 terms with scores ranging from -5 to 5 and returns the sum of the scores for the words in a given sentence. Valence aware dictionary for sentiment reasoning (VADER) (Hutto & Gilbert, 2014) uses a curated lexicon of 7,517 words and returns a normalized score ranging from -1 to 1. The terms used in Sentiment140 and VADER have sentiment scores with floating-point values. SentiStrength

(Thelwall et al., 2010) and TextBlob[1] are rule-based sentiment analyzers using lexicons of 2,918 and over 2,800 terms, respectively. Both analyzers return a polarity score ranging from -1 to 1.

While general-domain sentiment lexicons contain terms such as “happy” or “sad” that express human feelings, financial domain-specific sentiment lexicons contain terms such as “profitable” or “unprofitable” that can assess a corporation’s financial performance. The most popular lexicon in the financial domain is the Loughran-McDonald Word List created by Loughran & McDonald (2011). They claimed that approximately three-quarters of the negative words found in the Harvard General Inquirer word lists (Stone et al., 1962) were associated with non-negative sentiments when viewed from the perspective of business applications. To appropriately determine sentiment words in the financial domain, Loughran & McDonald (2011) created an accurate and reliable lexicon containing 354 positive words and 2,355 negative words by examining 2.5 billion words in the Form 10-K filings, which comprehensively summarize individual companies’ financial performance. However, manually constructed lexicons face limitations in terms of time, human labor, and background knowledge (Li et al., 2014).

Corpus-based approaches using statistical features have been proposed to automate the construction process. Yekrangi & Abdolvand (2021) used pointwise mutual information (PMI) to estimate the polarities of individual words. The authors first analyzed 554,915 textual documents published on Bloomberg and Reuters between 2006 and 2013 to identify the words frequently used in the financial domain, and then investigated the sentiment orientation of each word. Brazdil et al. (2022) proposed scoring methods that use word frequencies to estimate the distribution of word occurrence probabilities. The authors

[1] `https://github.com/sloria/TextBlob` [Accessed: December 15, 2022]

developed a manually labeled dataset comprising sentences from Portuguese news articles and analyzed the distribution of word occurrences across various sentiment scores.

As indicated in Chapter 1, one of the challenges in conducting sentiment analysis is the financial ontology that the sentiment orientation of a word may change significantly depending on the presence of directional expressions. To address the financial ontology when identifying sentiment words in financial texts, several studies have adopted manual approaches. Malo et al. (2014) extracted a list of terms from the Investopedia website and found 177 financial entities that could affect the sentiment of a sentence when used with motion verbs. Krishnamoorthy (2018) manually defined words that indicated the results of firm activities—such as improvement or decline in sales, market share, operating profit, operating cost, orders, and inventory turns—as lagging indicators and defined words that indicated future events—such as the number of new stores and employees—as leading indicators. Moreno-Ortiz et al. (2020) carefully analyzed business news articles to identify financial terms that conveyed a sentiment when combined with directional lexical elements. Defining the pairing of a term and a directional element as a multi-word expression, they constructed a lexicon containing 6,470 entries, including both single- and multi-word expressions. These studies on manual curation demonstrated improvement in terms of sentiment analysis, proving the importance of directionality in the financial domain. However, as has been repeatedly indicated, these manual approaches face limitations in terms of scale and cost.

To automate the extraction process, Oliveira et al. (2016) attempted to measure the relationship between words and modifiers. A modifier is an optional element that modifies the meaning of another element in a phrase or clause structure. For each word, the authors measured the degrees of association with sentiment labels and modifiers, respectively.

Two types of modifiers were used in their experiment: intensifiers (e.g., "more" and "increase") and diminishers (e.g., "less" and "decrease"). However, as the study relied on the aggregated estimates of the degree of word association, the given words and modifiers still had indirect relationships.

2.1.3 Pretrained models

Pretrained models have become state-of-the-art algorithms for general language-understanding tasks. Previous studies have shown that pretrained models can learn universal language representations, thereby achieving outstanding performance in various tasks for general language understanding (Qiu et al., 2020; Zhao et al., 2021).

Word2vec (Mikolov et al., 2013b) is one of the most well-known pretrained models. Word2vec uses two-layer neural networks to learn low-dimensional word embeddings. Two different methods exist to implement the Word2Vec model: continuous bag-of-words (CBOW) and skip-gram. In CBOW architecture, the model predicts the center word based on the neighboring words. Conversely, skip-gram architecture forces the model to predict all surrounding words ("context") based on the center word. Skip-gram is known to be more effective for a big corpus than the CBOW. This is because the skip-gram architecture treats every context-center pair as a new observation (Yilmaz & Toklu, 2020). Fig. 2.2 shows architectures of CBOW and skip-gram.

BERT (Devlin et al., 2019) is a pretrained language model that uses masked language modeling, which is a self-supervised pretraining objective that allows a Transformer (Vaswani et al., 2017) encoder to attend to bi-directional contexts during pretraining. Specifically, for an input sequence $S = w_1, ..., w_{N_w}$ of N_w tokens, BERT first randomly masks 15% of the tokens. A special symbol [MASK] is used to represent the masked tokens in the input sequence, and they are fed into a multi-layer transformer encoder. Following

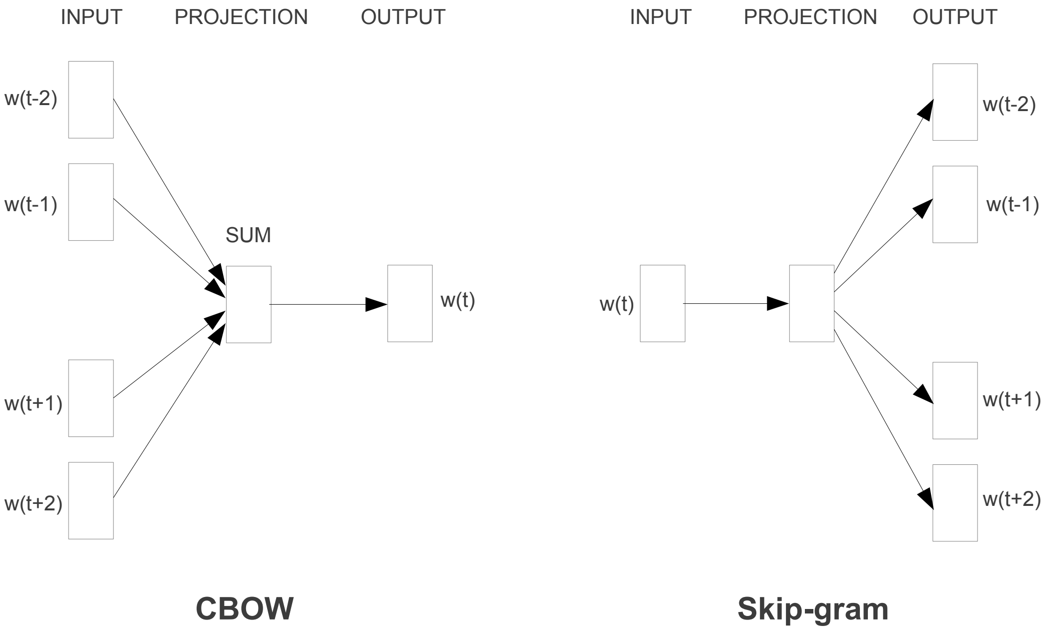


Figure 2.2: Comparison between CBOW and skip-gram architectures (Mikolov et al., 2013a)

this, the masked tokens are independently predicted using representations of the unmasked tokens in the top layer (Gu et al., 2020). Next sentence prediction is another pretraining task that BERT uses. In the next sentence prediction task, the model determines whether the second half of the input follows the first half of the input in the corpus or is a random paragraph (Clark et al., 2019).

A robustly optimized BERT (RoBERTa) (Liu et al., 2019) improves BERT with longer training, larger batches, and removal of the next sentence prediction task for pretraining (Ostendorff et al., 2021).

A few studies have been conducted to develop financial domain-specific pretrained language models. Whereas general-purpose models, including BERT (Devlin et al., 2019), are pretrained using general domain corpora, such as Wikipedia and Book Corpus, financial domain-specific pretrained language models are trained using financial texts. Araci's FinBERT (Araci, 2019), which indicates BERT for financial natural language processing tasks, is a model involving post-trained BERT with conventional subword masking based on financial news articles.

Yang et al.'s FinBERT (Yang et al., 2020), which indicates a finance domain-specific BERT, is a model pretrained using Form 10-K/10-Q filings, earnings conference call transcripts, and analyst reports. SEC-BERT (Loukas et al., 2022) is a pretrained model using Form 10-K filings submitted by companies listed on the U.S. Securities and Exchange Commission (SEC). Table 2.1 presents the characteristics of these financial domain-specific pretrained language models. Researchers have demonstrated that pretrained language models with financial text-driven linguistic knowledge can provide better performances for financial applications than general-purpose pretrained language models (Liu et al., 2021). These financial domain-specific pretrained language models contain enriched linguistic knowledge that can capture the semantic roles of words and the syntactic structures of sentences (Chiang et al., 2020).

Table 2.1: Characteristics of several financial domain-specific pretrained language models

Model	Method	Dataset
BERT (Devlin et al., 2019)	Pretraining from scratch	Wikipedia and Book Corpus
Araci's FinBERT (Araci, 2019)	Post-training on BERT	46,143 financial news articles that were published by Reuters from 2008 to 2010
Yang et al.'s FinBERT (Yang et al., 2020)	Pretraining from scratch	60,490 Form 10-K filings and 142,622 Form 10-Q filings of Russell 3000 firms from 1994 to 2019; 136,578 earnings conference call transcripts of 7,740 public firms from 2004 to 2019; 488,494 analyst reports issued for S&P firms from 1995 to 2008
SEC-BERT (Loukas et al., 2022)	Pretraining from scratch	260,773 Form 10-K filings from 1993 to 2019

2.1.4 Knowledge enhancement methods for pretrained language models

To enrich entity-level knowledge of general-purpose pretrained language models, various types of knowledge sources have been used. Several researchers have considered incorporating pretrained entity embeddings, which exist in static form, into pretrained language models. Zhang et al. (2019) identified named entities in a sentence using an entity linker and extracted the corresponding entity embeddings in Wikidata (Vrandečić & Krötzsch, 2014). Peters et al. (2019) proposed a context-sensitive entity linker that computes the weighted average of entity embeddings. Their proposed pretrained language models, Enhanced Language Representation with Informative Entities (ERNIE) (Zhang et al., 2019) and Knowledge Enhanced Contextual Word Representations (KnowBERT) (Peters et al., 2019), demonstrated promising performance on entity-related downstream tasks such as relationship extraction and entity typing. However, models that use static entity embeddings as inputs do not jointly learn entity and token representations, thus resulting in inherent gaps in the embedding space. Researchers have attempted to align these two types of representations into the same semantic space by simultaneously training both entity and token representations. Knowledge graphs are among the most well-known resources used to train entity representations. Wang et al. (2021b) encoded textual entity descriptions as entity representations and jointly optimized knowledge graph embedding (Sun et al., 2019) and masked language modeling losses. Their proposed language model exhibited superior performance over ERNIE and KnowBERT, thus demonstrating the effectiveness of knowledge graphs as an external knowledge base. However, both pretrained entity embeddings and knowledge graphs are disadvantageous in that they are static resources instead of dynamic resources that change in real time. Once information is updated, a certain amount of time and effort is required to incorporate the updated information into the designated structure. Particularly in the financial domain, the frequency of updates increases significantly since market- and business-related information changes more rapidly than that in other domains.

A possible solution for avoiding the inconvenience associated with converting information into a specified structure is to acquire factual knowledge from textual sources that directly reflect the updated information. Researchers have extensively employed Wikipedia as a textual knowledge base, which provides significant amounts of encyclopedic knowledge (Gabrilovich & Markovitch, 2009), such as human biographies and event backgrounds. Xiong et al. (2020) proposed a new training objective that allowed a model to incorporate information regarding real-world entities, where all titles on Wikipedia pages were regarded as entities. The entity name was replaced with another entity name of the same type, and the model was trained to determine whether the entity was replaced. The models trained with the proposed objective achieved significant improvements in fact completion, thus demonstrating their enhanced factual knowledge. Yamada et al. (2020) proposed a more straightforward task of randomly replacing named entities with `[MASK]` and training the model to predict the original tokens of these masked entities. Their proposed model achieved excellent performances on various entity-related tasks, such as question answering and named entity recognition. These studies demonstrated the importance of entity-level information for general language understanding, as well as the effectiveness of Wikipedia as a textual knowledge base.

2.2 Text Mining-based Early Warning

2.2.1 Macroeconomic indicator prediction

As listed in Table 2.2, text data gathered from various sources have been used to predict macroeconomic indicators to detect uncertainties in the financial market.

News articles, which contain data on various issues such as politics, economy, society, and culture from a journalist's perspective, convey the society's collective assessment of the market (Ko et al., 2020). Baker et al. (2016) proposed a new indicator based on the frequency of words related to economic uncertainty in news articles. The authors reported that their indicators were significantly similar to the market volatility index. Ko et al. (2020) applied Latent Dirichlet Allocation (Blei et al., 2003) to news articles containing economic outlook-related words. Following this, the authors constructed a regression model using text-driven features. The authors reported that the proposed regression model exhibited a high explanatory power of about 73% in predicting the price of the Korea Stock Price Index 200. Altogether, these results revealed the prospect of early warning indicators based on text mining techniques. However, the aforementioned studies focused only on a

Table 2.2: Past literature on text mining-based macroeconomic indicator prediction

Study	Data	Methodology	Purpose	Keyword extraction technique
Baker et al. (2016)	News articles	Regression model	Developing an early warning indicator	Frequency-based keyword scoring
Ko et al. (2020)	News articles	Regression model	Developing an early warning indicator	Latent Dirichlet Allocation
Kim (2018)	Monetary policy communications of the Bank of Korea	Random forest	Predicting decision result of Monetary Policy Committee	-
Nyman et al. (2021)	Reports published by the Bank of England	Vector autoregressive model	Developing an early warning indicator	-
Bennani & Neuenkirch (2017)	Speeches published by the European Central Bank	Sentiment analysis-based regression model	Predicting economic indicators	-
Masawi et al. (2018)	Speeches published by the Bank of Canada and the Reserve Bank of Australia	Event study methodology	Forecasting the exchange rate between Canada and Australia	Leximancer software-based keyword extraction

predetermined set of specific words. Predetermination of relevant words by researchers requires excellent background knowledge, and is expensive in terms of human labor and domain expertise.

Furthermore, news articles tend to be generated over a wide range of topics related to current issues. Therefore, for the purpose of economic analysis, an extra stage of manual or automatic data processing to categorize documents into finance-related categories is necessary. However, few category labels explicitly indicate the financial relevance of an article. Thus, manual and machine-based methods would be expensive in terms of human labor and computational costs, respectively.

One possible solution to this issue is to focus on documents that specifically address financial issues. Examples of such documents include central bank communications. Multiple studies have attempted to predict future trends in monetary policy by analyzing central bank communications. Kim (2018) trained a random forest-based classifier to predict future base rates using monetary policy communications published by the Bank of Korea. The proposed classifier achieved an accuracy of 84.6%, demonstrating the viability of text mining-based future base rate prediction. To develop an early warning indicator, Nyman et al. (2021) built a vector autoregressive model using the reports published by the Bank of England. The authors reported that the proposed indicator was useful in gauging risks against financial stability.

Several studies have focused on analyzing central bankers' speeches. Bennani & Neuenkirch (2017) conducted sentiment analysis of speeches issued by the European Central Bank to determine the correlation between inflation expectations and sentiment scores. Their results indicated that inflation and growth expectations had a significant, positive impact on the aggressive nature of speeches. Masawi et al. (2018) analyzed the impact of speeches made by central bankers of Canada and Australia on exchange rates. The authors quantified the usefulness of information obtained from the speeches by using information-theoretic measures and then analyzed the correlation between the calculated values and mean exchange rate returns. Their results indicated that speeches published by the Bank of Canada reduced the mean CAD/USD exchange rate returns, while those published by the Reserve Bank of Australia had no effect on exchange rate returns.

Despite extensive efforts to employ text mining techniques to analyze central bank communications, research on preemptive detection of risk factors directly related to financial crises has been scarce. Although Masawi et al. (2018) performed Leximancer software-based keyword extraction on central bankers' speeches, their work focused on the forecast of exchange rates, rather than the identification of risk factors leading to financial crises. Furthermore, most of the previous studies have extracted unigrams, which are words containing a single token. However, unigrams may deliver rather vague meanings, hindering proactive risk management.

2.2.2 Risk factor identification

The most typically used data source for extracting potential risk factors is Form 10-K filings. Form 10-K filings are the annual reports that public companies submit to the SEC. The filings provide detailed information about corporate financial conditions and potential risks (Zhu et al., 2016). While numerous studies have adopted topic modeling methods to detect risk factors, Sentence Latent Dirichlet Allocation (Sent-LDA) (Bao & Datta, 2014) has been the most popularly used one. Sent-LDA inherits the basic concept of Latent Dirichlet Allocation (Blei et al., 2003) and further adds the rule that each sentence discusses only one topic. Using Sent-LDA, Wei et al. (2019) created a hierarchical system for identifying corporate risk factors in the energy sector. The authors identified 66 risk factors for energy corporations using the proposed model. Li et al. (2020a) applied Sent-LDA to Form 10-K filings of tourism companies and identified 30 risk exposures in the tourism industry. Wei et al. (2022) applied Sent-LDA to Form 10-K filings of 34 fintech companies from 2015 to 2019 and identified 20 fintech risk factors. Zhu et al. (2022) analyzed 11,921 annual reports released by 1,570 companies from 2006 to 2019 and identified a total of 13 reputational risk drivers and their dynamic evolutions.

However, the overall risk factors extracted from these studies tended to be ambiguous, thereby seriously hindering proactive risk management (Zhu et al., 2022). The meanings of the topics were very open-ended, causing challenges in interpreting the topics (Jallan & Ashuri, 2020).

Furthermore, such approaches analyzing Form 10-K filings have limitations because they can only identify corporate-level risk factors, not macroeconomic risk factors that affect the entire financial market. Form 10-K filings, which comprehensively summarize individual companies' financial performance, are not suitable for identifying risk factors that may lead to a financial crisis.

One of the documents covering overall financial issues across the companies or industries is official financial documents published by trusted international organizations. These documents often contain information on the stance of international organizations on future monetary policies (Bennani et al., 2020). Especially, central bank communications, including minutes, statements, and speeches published by a central bank, comprise one of the most popular sets of official financial documents.

2.3 Explainable Market Sentiment Analysis

2.3.1 Knowledge graph-based approaches

Knowledge graphs are graph-structured knowledge bases, where information is represented in the form of entities (e.g., nodes) and their relationships (e.g., edges) (Dettmers et al., 2018). Since knowledge graph-based explanations have been considered understandable from a human perspective, numerous researchers have adopted knowledge graphs to support explainable systems (Tiddi & Schlobach, 2022).

SenticNet (Cambria & Hussain, 2015b) is a commonsense knowledge graph for opinion mining and sentiment analysis; and sentic computing framework (Cambria & Hussain, 2015a), which uses SenticNet, has been extensively used in sentiment-based financial forecasting. Fig. 2.3 describes sentiment analysis procedures of the sentic computing framework[2].

[2] `https://sentic.net/computing/` [Accessed: December 15, 2022]

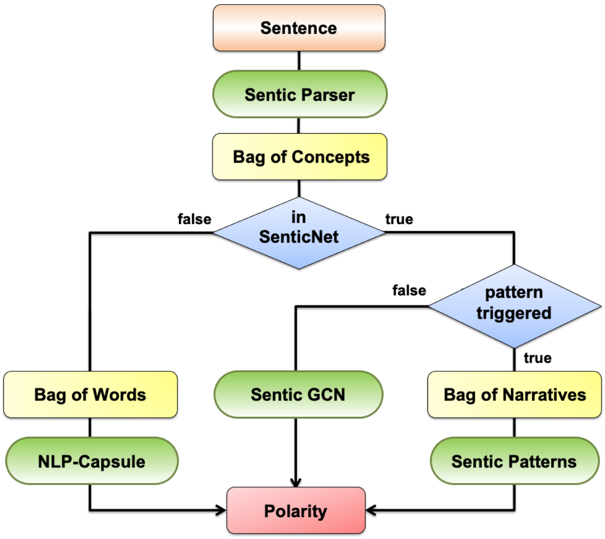


Figure 2.3: Sentiment analysis procedures of the sentic computing framework (Cambria & Hussain, 2015a)

Xing et al. (2018) obtained the sentic computing-based sentiment scores for microblogs and proposed an asset allocation framework incorporating market sentiment. Experimental results demonstrated that their proposed framework outperformed other successful forecasting techniques. Xing et al. (2019) calculated the sentic computing-based sentiment scores for posts on StockTwits. StockTwits[3] is a social media platform for investors, traders, and entrepreneurs to share their investment ideas. The authors then proposed a sentiment-aware volatility forecasting model that reflects market sentiment for predicting stock returns. Experimental results showed that their model outperformed both statistical models and deep learning-based models. Picasso et al. (2019) applied the sentic computing framework to news articles and used the sentiment scores to develop classification models for market trend forecasting. The authors conducted experiments on a portfolio composed of the most capitalized 20 companies listed in the Nasdaq 100 index, and the experi-

[3] `https://stocktwits.com/` [Accessed: December 15, 2022]

mental results indicated that their proposed model could predict trends of the portfolio. Altogether, these studies demonstrated the effectiveness of SenticNet as an explainable knowledge base for sentiment analysis.

2.3.2 Lexicon-based approaches

Lexicons are another tool that can explain the process of sentiment analysis. Lexicons are intuitive to interpret and easy to implement (Bandhakavi et al., 2017; Razova et al., 2022). Once a lexicon is compiled, a researcher can easily measure the text's sentiment value without additional training data and a long learning process. Additionally, the constructed lexicon can be used as a resource to train deep learning models (Choi et al., 2020) or conduct other tasks.

Lexicons that can achieve reasonable performance have constantly drawn the attention of the community (Cheng et al., 2022). Oliveira et al. (2017) used lexicons to assess the impact of tweets on the stock market. Through their experiments, the sentiment on Twitter and the volume of posts were found to be important for predicting the returns of the S&P 500 index. Song & Shin (2019) adopted a lexicon-based approach to conduct sentiment analysis of news articles and showed that an effective source for creating an economic indicator in Korea could be news articles. Gakhar & Kundlia (2021) developed a regression-based predictive model using features derived by lexicon-based sentiment analysis. The authors demonstrated that negative sentiment scores were crucial variables to predict the volatility and liquidity of stock returns. Erçen et al. (2022) used a hybrid strategy that combined lexicons and machine learning techniques to investigate the effect of news regarding macroeconomic policies on exchange rate fluctuations. According to the findings of their experiments, a significant relationship existed between exchange rate volatility and sentiment scores derived from news regarding inflation rates, interest rates, and credit ratings.

2.4 Stock-related Spam Filtering

2.4.1 Manual approaches

Spam filtering aims to determine whether a document is generated by a spam bot or a human. Most previous studies on stock-related spam filtering have focused on unveiling bots' activities by using a systematically designed analysis framework requiring manual inspection.

Cresci et al. (2018) performed the first large-scale systematic analysis of the existence and impact of spam and bot activities in stock-related microblogs. The authors focused on uncovering a malicious practice intended to promote low-value stocks by exploiting the popularity of high-value ones. Finally, the authors emphasized the importance of developing intelligent financial-spam filtering techniques and adopting them in all systems that use stock microblogs.

To understand the characteristics of stock-related spam bots, a few studies manually examined several accounts that produced spam messages (Cresci et al., 2019; Tardelli et al., 2020). These studies revealed that a number of cases existed where stocks that had low market capitalization and were mainly traded in over-the-counter markets were mentioned together with a few high-capitalization stocks traded in the Nasdaq and New York Stock Exchange.

2.4.2 Data-driven approaches

To address the issues regarding the manual inspection of bots' activities, Tardelli et al. (2022) developed a classifier using automatically calculated hundreds of features. The authors demonstrated that account information-based features, such as the number of days since account creation, had a much greater impact on the spam filtering performance than textual content-based features. However, it seems that the naive bag-of-words encoding method, which was used in their experiment, would have hindered extracting useful information from texts.

Zhang et al. (2022b) developed a machine learning system to detect financial disinformation spreading on social media platforms. To this end, they compiled a dataset that comprises financial news articles published on Seeking Alpha and the microblogs that disseminate the news on other social media platforms such as Twitter. Using this dataset, they created an extensive set of computable metrics to quantify the context and motive of financial news, sender demeanor, responses from third parties, content correspondence, and content coherence. According to their thorough analysis and performance evaluation, their system consistently achieved superior performance in identifying financial disinformation than baseline methods.

Chapter 3

Hot Topic Detection for Early Warning

3.1 Background

A financial crisis is academically defined to be an economic scenario where certain financial assets suddenly lose a significant portion of their nominal value (Frankel & Saravelos, 2012). While extensive efforts have been expended to explain or predict financial crises from various perspectives (Acharya & Richardson, 2009; Adrian & Shin, 2010; Foster & Magdoff, 2009; Thakor, 2015), several researchers have attempted to develop early warning systems to reduce crisis risk by identifying probable risks (Koyuncugil & Ozgulbas, 2010, 2012).

Quantitative models and qualitative monitoring and evaluation (QME) are regarded as the two pillars of early warning systems (Gaytán et al., 2002). In particular, QME is essential for the detection of risk factors that are difficult to identify using quantitative models (Lee, 2015). As indicated in Chapter 1, QME is performed by a small number of domain experts who read and summarize documents published by global financial institutions (Koyuncugil & Ozgulbas, 2012). However, manual QME, which requires domain experts to review each sentence to reveal the underlying meaning of the sentence, is not suitable for real-time analysis that requires timely detection and response to financial market changes. Furthermore, the pool of domain experts dedicated to QME is not constant.

Frequent changes occur owing to intra- or inter-company transfers, promotions, demotions, resignations, or layoffs. In such cases, manual screening processes may produce inconsistent results owing to a variation in the level of expertise and standards.

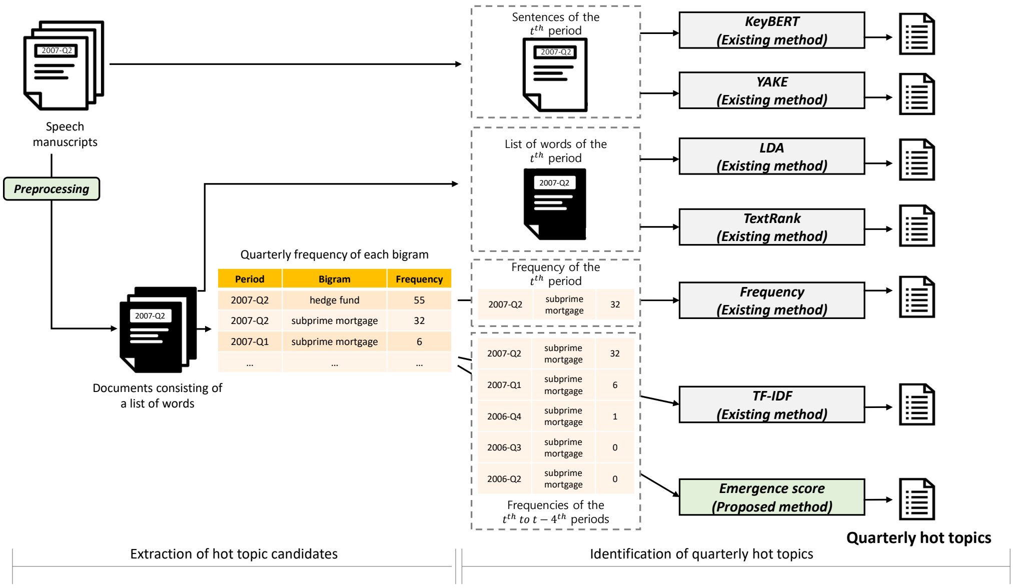
To address the issues associated with manual QME and satisfy the growing demand for early warning systems capable of analyzing vast amount of data in real time, text mining-based monitoring approaches have been proposed. Researchers have attempted to automate the analysis of finance-related documents and assess the uncertainty of the market in an objective manner. In some previous studies, news articles were analyzed to understand the behavior of market participants (Baker et al., 2016; Ko et al., 2020), or Form 10-K filings were analyzed to identify the risk factors of companies (Li et al., 2020a; Wei et al., 2022, 2019; Zhu et al., 2022). A few studies have focused on estimating future trends in monetary policies of central banks by analyzing their communications such as minutes, statements, or speeches published by central banks (Bennani & Neuenkirch, 2017; Kim, 2018; Masawi et al., 2018; Nyman et al., 2021). Public announcements of central banks are known to influence price fluctuations in financial instruments; thus, analyzing such announcements can facilitate market participants' deeper understanding of the market.

A central banker's speech is an important data source among its communications, as it reveals where speakers stand on certain monetary policies or the general economic outlook at the time of the speech (Bennani & Neuenkirch, 2017). Therefore, text mining-based approaches involving the scrutiny of a large number of central bankers' speeches are expected to broaden the understanding of market risk factors, while satisfying a fixed set of objective and data-driven standards. Although several studies (Bennani & Neuenkirch, 2017; Masawi et al., 2018) have undertaken text mining-based analysis of central bankers' speeches, there is room for improvement in that most studies have focused on predicting macroeconomic indicators rather than identifying specific words that have a significant impact on the market.

In this study, a hot topic detection framework is proposed to automate the identification of risk factors that are likely to appear before a financial crisis. The proposed framework detects quarterly hot topics by computing "emergence scores," which are designed to detect key phrases emerging in the given quarter.

3.2 Proposed Method

As shown in Fig. 3.1, the proposed hot topic detection framework comprises two stages: (1) the extraction of hot topic candidates and (2) the identification of quarterly hot topics.


2007-Q2
Speech manuscripts
Preprocessing
2007-Q2
Documents consisting of a list of words
Sentences of the t^{th} period
2007-Q2
List of words of the t^{th} period
2007-Q2
Frequency of the t^{th} period
Frequencies of the t^{th} to $t-4^{th}$ periods
KeyBERT (Existing method)
YAKE (Existing method)
LDA (Existing method)
TextRank (Existing method)
Frequency (Existing method)
TF-IDF (Existing method)
Emergence score (Proposed method)
Quarterly hot topics
Extraction of hot topic candidates
Identification of quarterly hot topics

Quarterly frequency of each bigram

Period	Bigram	Frequency
2007-Q2	hedge fund	55
2007-Q2	subprime mortgage	32
2007-Q1	subprime mortgage	6
...	...	...

2007-Q2	subprime mortgage	32

2007-Q2	subprime mortgage	32
2007-Q1	subprime mortgage	6
2006-Q4	subprime mortgage	1
2006-Q3	subprime mortgage	0
2006-Q2	subprime mortgage	0

Figure 3.1: Overview of the proposed bigram-based just-in-time hot topic detection framework. Six baseline methods—KeyBERT, Yet Another Keyword Extractor (YAKE), Latent Dirichlet allocation (LDA), TextRank, frequency-based method, and term frequency-inverse document frequency (TF-IDF)-based method—are also presented for comparison with the proposed emergence score-based method.

3.2.1 Extraction of hot topic candidates

Given n documents sorted chronologically, hot topic candidates are extracted by identifying bigrams in each document representing a single quarter. An n-gram is defined to be a sequence of n adjacent elements from a list of words; n-grams having 1 and 2 as the values of n are referred to as unigrams and bigrams, respectively. Bigrams, which provide more specific meaning than unigrams, are considered to be suitable as event representative keywords (Parikh & Karlapalem, 2013). To extract bigrams in this experiment, a list of lemmatized words composed of nouns, adjectives, adverbs, and verbs extracted from each sentence is used. Details on how to convert a sentence into a word list are presented in Section 3.3.1. Table 3.1 demonstrates the extraction of bigrams in a sentence. Then, the quarterly frequency of each candidate keyword, that is, its total number of occurrences per quarter, is calculated.

Table 3.1: Demonstration of the bigram extraction for the sentence "The exception is subprime mortgages with variable interest rates, for which delinquency rates have increased appreciably," which is a part of the speech addressed by Ben Shalom Bernanke in February 2007 (Bernanke, 2007a)

Sentence	"The exception is subprime mortgages with variable interest rates, for which delinquency rates have increased appreciably."
List of words	["exception", "is", "subprime", "mortgage", "variable", "interest", "rate", "delinquency", "rate", "have", "increased", "appreciably"]
List of bigrams	[("exception", "is"), ("is", "subprime"), ("subprime", "mortgage"), ("mortgage", "variable"), ("variable", "interest"), ("interest", "rate"), ("rate", "delinquency"), ("delinquency", "rate"), ("rate", "have"), ("have", "increased"), ("increased", "appreciably")]

3.2.2 Identification of quarterly hot topics

In the second stage, quarterly hot topics are identified using the emergence score-based keyword scoring method. It incorporates the temporal importance of keywords in a simple and effective manner by estimating the growth rate of the term frequency. The emergence score is a novel keyword-scoring metric proposed in this study. It measures the degree of emergence of a word compared to the past four quarters. It treats words that appear suddenly when compared to the past as risk factors. Specifically, the emergence score of a word w in the t^{th} document is computed as follows:

$$EmergenceScore(w,t) = \frac{Frequency(w,t)}{(\sum_{i=t-4}^{t-1} Frequency(w,i) + 1)/4} \tag{3.1}$$

where $Frequency(w,i)$ denotes the frequency of the word w in document d_i, which is the i^{th} document in chronological order. Words with a high frequency in a particular quarter compared to those over the previous four quarters are assigned high emergence scores. Finally, the top-n keywords in terms of scores are selected as hot topics.

The proposed emergence score can be compared to TF-IDF (Salton & Buckley, 1988). As indicated in Section 2.1.1, TF-IDF is the product of two statistics, term frequency, which is the frequency of the word in the subject document, and inverse document frequency. The inverse document frequency is obtained by dividing the total number of documents by the number of documents containing the term and then taking the logarithm of that quotient. If a document contains a particular word, it counts to 1 and is reflected in the scoring, regardless of the number of times the word appears. That is, documents with a lower frequency for the word and documents with a higher frequency for the word are treated the same. As a result, keywords tend to be extracted mainly based on words with a very high term frequency. On the other hand, the proposed emergence score reflects the number of times a word appears in the document when calculating the denominator; thus, words with higher growth rates than in the past are given higher scores, even if the term frequency of the word is relatively low. Table 3.2 lists a comparison of the top five hot topics extracted using TF-IDF and emergence score-based keyword scoring methods. This result shows that the proposed emergence score can complement the TF-IDF. The keywords “cash flow” and “job loss,” which explain the 2008 global financial crisis well, recorded high emergence scores despite relatively insignificant term frequencies; thus, they were extracted as hot topics by the emergence score-based method. A more detailed analysis of the results is described in Section 3.3.3.

Table 3.2: Comparison of the top five hot topics detected in speeches addressed in the second quarter of 2007

TF-IDF-based method			Emergence Score-based method		
Hot Topic	Term Frequency	Emergence Score	Hot Topic	Term Frequency	Emergence Score
hedge fund	55	4.0	subprime mortgage	32	86.2
u s	52	3.0	finance premium	20	60.2
subprime mortgage	32	86.2	external finance	20	60.2
finance premium	20	60.2	cash flow	16	48.2
external finance	20	60.2	job loss	14	41.2

3.3 Experiments

3.3.1 Data description

Manuscripts of speeches made by the chairs of the Federal Reserve System between January 1997 and September 2019 were collected from the official website of the Bank for International Settlements (BIS, 2019). Since speeches made by officials during their regime as a chair are expected to have a greater impact on the financial market than speeches made by them outside of their term, only speeches of the former type were collected. Finally, a dataset comprising 491 speeches was constructed. A summary of the dataset is presented in Table 3.3.

The collected documents were in PDF format; they were converted to text file format using the PDFMiner.six library (Shinyama et al., 2019). Then, each sentence was transformed into a list of words consisting of nouns, adjectives, adverbs, and verbs via

Table 3.3: Dataset characteristics

Chair of the Federal Reserve System	Start of term	End of term	# Speeches
Alan Greenspan	1987-08-11	2006-01-31	192
Ben Shalom Bernanke	2006-02-01	2014-01-31	218
Janet Yellen	2014-02-03	2018-02-03	57
Jerome Powell	2018-02-05	Incumbent as of September 2019	24
Total			491

tokenization and part-of-speech tagging using the Natural Language Toolkit library (Bird et al., 2009). All extracted words were lemmatized using the Natural Language Toolkit library. In addition, words appearing in more than 90% of speeches were considered to be domain-specific stopwords and were removed. During this process, care was taken to exclude words related to inflation and employment since they serve as the primary concerns of the Federal Reserve System (Thorbecke, 2002). The full list of 20 stopwords satisfying the aforementioned criteria is listed in Table 3.4. Finally, sentences containing less than three or more than 180 tokens were discarded as they tend to be either significantly short to be analytically informative or significantly long owing to errors regarding the pdf-to-text document conversion process. Subsequently, all sentences addressed during the same quarter were aggregated as a document. Data ranges from the first quarter of 1997 to the third quarter of 2019; hence, the aggregation process resulted in 91 documents. As the proposed framework detected the emergence of words relative to the previous four quarters, a list of hot topics from the first quarter of 1998 to the third quarter of 2019 was obtained.

Table 3.4: Full list of the selected domain-specific stopwords

alan	bank	banking	ben	bernanke	bi	board	central	federal	financial
greenspan	ha	ii	janet	jerome	market	mr	percent	policy	powell

3.3.2 Experimental settings

KeyBERT (Grootendorst, 2020), YAKE (Campos et al., 2020), LDA (Blei et al., 2003), TextRank (Mihalcea & Tarau, 2004), frequency-based keyword scoring, and TF-IDF-based keyword scoring (Salton & Buckley, 1988) were used as baseline methods for comparison. In an experiment using KeyBERT and YAKE, the entire text of the document was used as input for each quarter. LDA and TextRank were implemented through the gensim (Rehurek & Sojka, 2011) library; and a list of preprocessed words for each quarter was used as input. LDA and TextRank are algorithms originally used for unigrams; thus, they were implemented to extract unigrams. In frequency-based keyword scoring, the quarterly frequency of each word was assigned as its score. In the case of using the TF-IDF-based keyword scoring, documents published from four quarters ago to one quarter ago from that year were used as a comparative group to extract keywords for a particular quarter. TF-IDF-based keyword scoring was implemented through the scikit-learn (Kramer, 2016) library. The frequency- and TF-IDF-based keyword scoring were designed to extract bigrams as keywords.

Dynamic topic models (Blei & Lafferty, 2006; Wang & McCallum, 2006), which investigate the gradual changes in topic meanings, were excluded from baseline methods.

3.3.3 Experimental results

The top five keywords are presented in Table 3.5 and Table 3.6 in order of their scores, as detected in speeches addressed in the second quarter of 1998, the first quarter of 2000, and the second quarter of 2007. For LDA, five word-sets representing five latent topics for each quarter are presented. The results show that keywords extracted as unigrams through YAKE, LDA, and TextRank tend to contain insignificant meanings; thus, they appear to be insufficient to be treated as risk factors for financial crises. However, bigram keywords, such as “east asia,” “information technology,” and “subprime mortgage,” provide more detailed information about financial crises. This shows that bigrams, by construction, are more suitable than unigrams in detecting hot topics for potential risk factors if the topic words are compound nouns.

Meanwhile, keywords extracted by KeyBERT and YAKE seem far from financial risk factors, except for the word “East Asia.” This shows that KeyBERT and YAKE are inherently unsuitable to detect financial risk factors because they ignore the temporal importance of words. Although interesting results may be obtained by applying the sliding window concept to the existing keyword extraction techniques, it would require very troublesome and complicated work in that the structure of the models should be modified. In contrast, the proposed method can extract keywords in a very simple and effective way without such hassle. These extensive experimental results show that the proposed approach is simple yet effective in identifying potential financial risk factors in central bankers’ speeches.

Table 3.5: Top five hot topics detected in the speeches. For LDA, five sets consisting of several words were extracted to represent five latent topics for each quarter.

	KeyBERT	YAKE	LDA
1998-Q2	consumer markets	economic	have, is, imf, not, crisis, been, economy, system, asian, be
	consumers shifting	United States	have, is, be, not, are, system, been, risk, economy, more
	retailing supermarkets	market capitalism	is, economy, not, system, are, have, be, consumer, state, planning
	forces consumers	East Asia	have, been, growth, be, increase, economy, price, year, is, not
	consumers evolving	economies	system, risk, be, is, have, not, capital, more, are, crisis
2000-Q1	rising inflationary	business	is, have, be, are, year, budget, not, economy, price, been
	inflation years	productivity growth	is, be, have, are, technology, capital, not, growth, information, business
	growth inflationary	productivity	is, are, have, be, not, technology, business, capital, growth, increase
	continued economic	market	be, is, are, year, have, economy, economic, capital, growth, not
	decades economic	demand	have, worker, are, is, business, community, be, not, economic, s
2007-Q2	financial regulation	financial institutions	have, s, mortgage, are, is, trade, be, risk, borrower, subprime
	financial regulatory	credit	risk, fund, have, are, investor, hedge, be, s, is, regulation
	banking regulation	financial system	s, i, is, wa, not, today, washington, school, robinson, education
	bank regulatory	Review Federal Reserve	trade, s, job, is, u, are, more, have, firm, service
	bank regulation	banks	credit, have, mortgage, be, borrower, lending, are, s, cra, loan

Table 3.6: Top five hot topics detected in the speeches

	TextRank	Frequency	TF-IDF	Emergence Score (proposed)
1998-Q2	economy	have been	have been	east asia
	capital	safety net	safety net	new system
	increase	standard living	east asia	imf s
	risk	interest rate	international system	nonperforming loan
	risking	not have	new system	high tech
2000-Q1	business	have been	have been	information technology
	product	good service	good service	demand supply
	productive	recent year	information technology	excess demand
	increase	information technology	recent year	net import
	increasing	wealth effect	excess demand	working group
2007-Q2	trading	hedge fund	hedge fund	subprime mortgage
	trade	u s	u s	finance premium
	traded	united state	subprime mortgage	external finance
	risk	subprime mortgage	finance premium	cash flow
	regulated	reserve system	external finance	job loss

The Asian financial crisis

By the mid-1990s, East Asian countries such as Thailand, Indonesia, and South Korea had accrued large private current account deficits. This was because the maintenance of fixed exchange rates encouraged external borrowing and led to excessive exposure to foreign exchange risk (Chowdhry & Goyal, 2000). Fig. 3.2 depicts the variation in the USD/KRW exchange rate. This began to surge in December 1997 and fluctuated sharply until the third quarter of 1998.

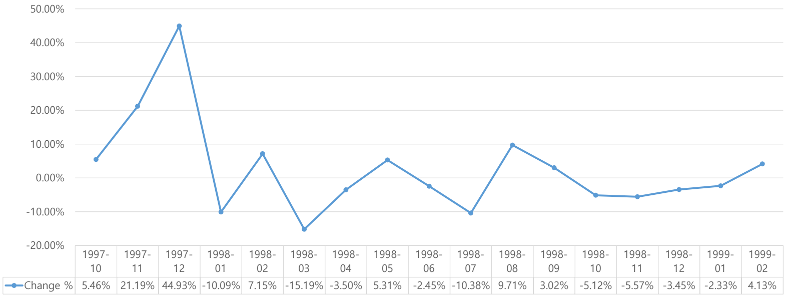


Figure 3.2: Variation in the USD/KRW exchange rate during the Asian financial crisis

As discussed in Section 3.3.1, the proposed framework was used to identify lists of hot topics from 1998 onwards, and not in 1997. As shown in Table 3.5 and Table 3.6, only the emergence score-based method successfully detected the terms “imf s” and “nonperforming loan,” which are keywords that narrow down the scope of the context of the crisis. Although the term “imf s” is a bigram extracted owing to errors in the tokenization process of recognizing “imf” and “s” as nouns in “IMF’s,” it serves as a hot topic that explains the financial crisis well. Detection of the term “imf s” is expected to help market participants carefully monitor future policy directions of the International Monetary Fund (IMF); the IMF provides financial assistance to countries facing balance-of-payments deficits (Joyce, 2000). Furthermore, the term “nonperforming loan” is one of the terms considered as the cause of the East Asian crisis; Yang (2003) reported that the steady rise in rates of nonperforming loans may have led to the Asian financial crisis.

The dot-com bubble

In the early 2000s, stock prices of information technology-related companies rapidly soared and then plummeted shortly after, leading to the bankruptcy of several small business owners (Min et al., 2008). Fig. 3.3 reveals that, in April 2000, the Nasdaq Composite index decreased by more than 15%, indicating the severity of the dot-com bubble crisis.

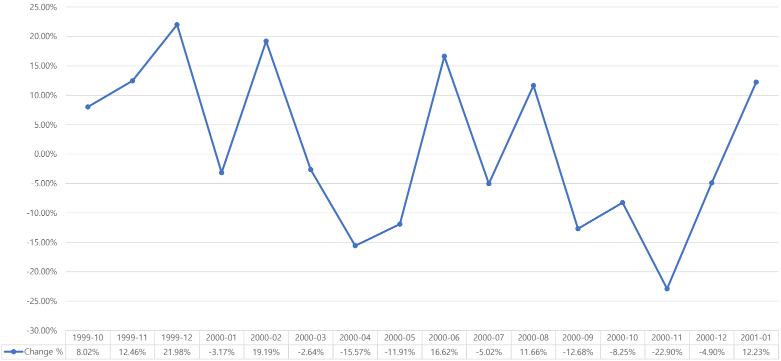


Figure 3.3: Variation in the Nasdaq Composite index during the dot-com bubble crisis

The keyword "information technology," which describes the dot-com bubble crisis well, was successfully detected as a hot topic in speeches published in the first quarter of 2000 by the frequency-, TF-IDF-, and emergence score-based methods. In Table 3.6, it can be clearly seen that the emergence score-based method assigned the highest score in the first quarter of 2000 to the keyword "information technology"; meanwhile, the frequency- and TF-IDF-based methods assigned the highest score to the meaningless term "have been." This is due to the fact that, as described in Section 3.2.2, both the frequency- and TF-IDF-based methods tend to detect keywords with high absolute term frequency; the proposed emergence score-based method, however, tends to reflect term frequency growth rates more than absolute term frequency.

The global financial crisis

In addition to the Asian financial crisis and the dot-com bubble, another case study on the 2008 global financial crisis was conducted. In 2008, the increase in delinquency rates of subprime mortgages, coupled with massive losses of investment companies, led to one of the most infamous global financial crises (Helleiner, 2011). Furthermore, the predicament worsened when hedge funds that traded subprime-backed products imploded owing to huge investment losses (Helleiner, 2011). Fig. 3.4 illustrates the downward trend of the S&P 500 index between the late 2007 and early 2009, indicating the severity of the crisis.

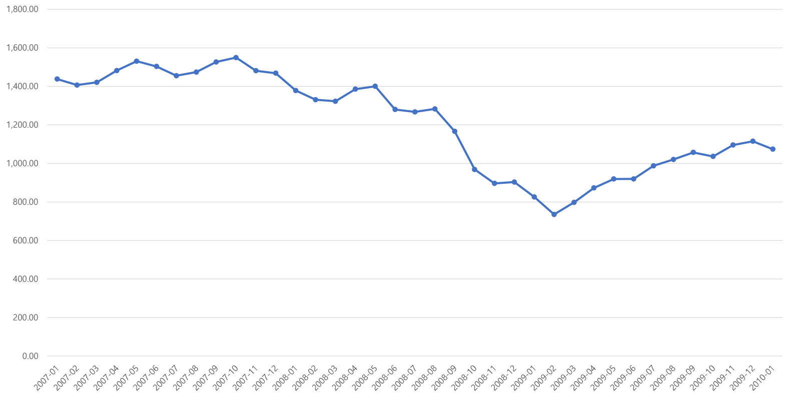


Figure 3.4: Variation in the S&P 500 index values during the global financial crisis

All of the bigram-based methods—frequency-, TF-IDF-, and the emergence score-based methods—successfully captured the circumstances by detecting the phrase "subprime mortgage" as an emerging keyword in speeches released during the second quarter of 2007, prior to the actual onset of the global financial crisis. This reflects the emerging concern of central bankers. Furthermore, the phrases "cash flow" and "job loss" were detected by the proposed emergence score-based scoring method but missed by the frequency- and TF-IDF-based scoring methods. These phrases can be considered as the most important keywords describing the global financial crisis; the keyword “cash flow” is related to the rise of subprime mortgage loans as risk factors (Erolu et al., 2012); the keyword “job loss” can be referred to be one of the consequential phenomena of the crisis (McDonnell & Burgess, 2013).

However, the keyword “hedge fund,” which is another key phrase describing the crisis, was detected only by the frequency- and TF-IDF-based methods. This suggests that the existing methods and the proposed emergence score-based method complement each other and together facilitate a better understanding of detected hot topics. Given that the global financial crisis had not begun in earnest in the second quarter of 2007, the aforementioned results suggest that the risk factors associated to a financial crisis may be preemptively identified by analyzing central bankers’ speeches.

The main financial crises that occurred between the late 1990s and the late 2000s—the Asian financial crisis, dot-com bubble, and global financial crisis—are closely linked to each other (Aliber & Kindleberger, 2015). Following the Asian financial crisis, funds from East Asia flowed into the United States of America (USA) (Aliber & Kindleberger, 2015), which influenced the outbreak of the dot-com bubble in the USA. As a result, housing loan interest rates were lowered and real estate prices increased, which raised the transaction volume of subprime mortgage loans by a significant volume and laid the foundation for the global financial crisis (Pavlov & Wachter, 2011). The fact that the keywords “East Asia,” “information technology,” and “subprime mortgage” that represent the internal sequence of these financial crises were successfully detected from the speeches suggests that speeches made by the chairs of the Federal Reserve System can be used as a useful data source for the proactive identification of early signs of a financial crisis.

Despite these encouraging results, the proposed approach exhibited poor performance in some cases. For example, a meaningless term, such as “http www,” was extracted as a hot topic for the second quarter of 2014, as recorded in Table 3.8. Through error analysis, it was found that most of the issues regarding the detection of contextually insignificant words could be attributed to pdf-to-text document conversion and preprocessing (Kim et al., 2020). For example, the term “http www” was extracted owing to the influence of the web links in the “References” section. This section ought to have been excluded from the analysis as it did not pertain to the contents of speeches; however, this was not possible during the conversion of the document from the PDF to the text format. Notwithstanding the limitations of the experiment, the aforementioned preliminary results demonstrate the viability of automatic detection of potential risk factors using the proposed framework.

Fig. 3.5 shows the sums of the emergence scores of the top five hot topics for each quarter. It should be noted that a high score does not indicate an increased possibility of a financial crisis in that quarter or the following quarters. Although words describing financial crises were observed to be frequent in the second quarter of 1998, the first quarter of 2000, and the second quarter of 2007, the sum of the scores of the top five hot topics detected in each of those quarters was not high.

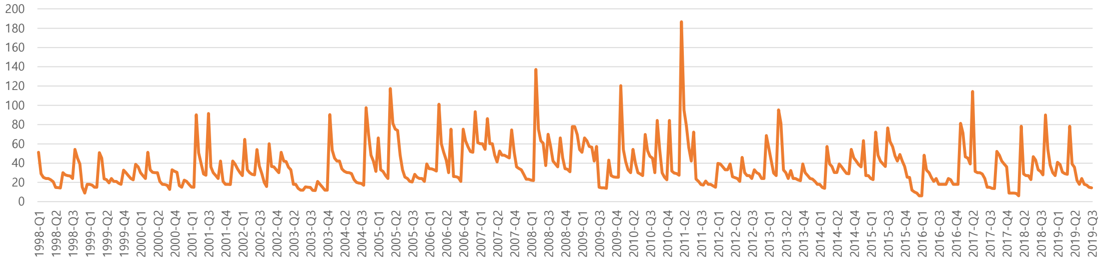


Figure 3.5: Sum of the emergence scores of the top five hot topics in each quarter

The primary objective of this study is to aid financial experts in identifying potentially influential keywords that may have contributed to the financial crisis. Financial experts will be able to manage financial risks more efficiently and effectively under a human-in-the-loop system (Zanzotto, 2019) that identifies the keywords discovered through the proposed method. The final risk assessment should be performed by financial experts.

Quantitative Evaluation

To investigate the timeliness of reporting potential risk factors of a financial crisis, this study followed the evaluation method of previous works (Petrovic et al., 2013) and compared the *ex ante* detection power of the proposed framework on speeches and news articles. In particular, the frequencies of the term "subprime mortgage" in central bankers' speeches and the New York Times news articles were compared over the first two quarters of 2007. As presented in Table 3.7, the number of the New York Times news articles mentioning the word "subprime mortgage" exhibited an approximate 0% increase in the second quarter of 2007 as compared to the first quarter of 2007, whereas the frequency of the term "subprime mortgage" increased by more than 400% in the speeches. These results not only suggest that it may be difficult to obtain hints regarding impending financial crises based on the New York Times news articles, but also indicate that central bankers' speeches often contain more useful information for the preemptive identification of risk factors than news articles.

Table 3.7: Frequencies of the term "subprime mortgage" in central bankers' speeches and the New York Times news articles during the first two quarters of 2007

	2007-Q1	2007-Q2	Growth rate
Number of the New York Times articles mentioning the word "subprime mortgage"	528	530	0.4%
Number of occurrences of the word "subprime mortgage" in the speeches	6	32	433.3%
Emergence scores of the word "subprime mortgage" in the speeches	18	86	377.8%

The fundamental aim of the proposed hot topic detection framework is the *ex ante* identification of the maximal possible number of indicators of potential economic threats at the cost of relatively low precision. This would allow users to identify factors that may lead to crises preemptively and prepare countermeasures as quickly as possible. Although the identified hot topics, which are listed in Table 3.8, may not correspond to the exact risk factors that would definitely develop into financial crises, it is expected that capturing the current economic trends and narrowing them down to specific topics to be closely monitored would help users prepare for impending crises proactively. Thus, the proposed hot topic detection framework should be used as a screening tool, and the final judgment for crisis response should be made by an expert.

Table 3.8: Top five hot topics detected using the emergence score-based keyword scoring

Period	Hot Topics
1998-Q1	regulatory capital, high tech, capital standard, capital regulation, vicious cycle
1998-Q2	east asia, new system, imf s, nonperforming loan, high tech
1998-Q3	distribution wealth, past year, equity premium, new economy, ownership rate
1998-Q4	fire sale, hedge fund, ltcm s, reserve new, exposure ltcm
1999-Q1	counterparty credit, notional value, saving rate, value derivative, potential future

1999-Q2	foreign exchange, exchange reserve, emerging economy, exchange rate, labor productivity
1999-Q3	acceleration productivity, last fall, pre emptive, s economy, reported earnings
1999-Q4	equity premium, supervision regulation, risk management, risk manager, internal risk
2000-Q1	information technology, demand supply, excess demand, net import, working group
2000-Q2	payment system, electronic payment, management system, holding company, private equity
2000-Q3	living standard, consumer spending, high level, labor resource, energy price
2000-Q4	old economy, oil price, th century, new economy, great grandparent
2001-Q1	energy cost, core deposit, price index, treasury security, medical service
2001-Q2	safety net, natural gas, gasoline price, electric power, u s
2001-Q3	capital gain, personal saving, equity extraction, saving rate, gain home
2001-Q4	international currency, currency is, foreign currency, vehicle currency, debt capacity

2002-Q1	stock price, capital investment, inventory liquidation, mortgage rate, corporate governance
2002-Q2	option grant, stock option, reserve ratio, stock price, reported earnings
2002-Q3	equity premium, stock price, corporate governance, earnings growth, new york
2002-Q4	labour social, risk management, south african, credit risk, african reserve
2003-Q1	thrift institution, home bias, mortgage debt, capital flow, population is
2003-Q2	benefit derivative, day to, counterparty credit, management failure, associated derivative
2003-Q3	balance sheet, firm have, structure economy, cid s, household have
2003-Q4	low wage, job loss, china s, not have, new job
2004-Q1	fannie freddie, intellectual property, s freddie, dollar s, service ratio
2004-Q2	current account, regulatory capital, high school, future price, capital standard
2004-Q3	aging population, fertility rate, baby boom, profit margin, population is

2004-Q4	payment system, electronic payment, member state, new member, oil price
2005-Q1	community development, tax system, not have, tax code, safety net
2005-Q2	fannie freddie, gse portfolio, gse debt, hedge fund, mortgage securitization
2005-Q3	intended saving, risk premium, intended investment, term premium, home equity
2005-Q4	surplus deficit, foreign saving, employment act, economic entity, balance dispersion
2006-Q1	low stable, price stability, inflation expectation, inflation is, term premium
2006-Q2	basel i, risk measurement, regulatory capital, working group, minimum regulatory
2006-Q3	economic integration, productivity gain, intangible capital, new technology, core periphery
2006-Q4	population aging, future generation, china s, lower income, monetary aggregate
2007-Q1	affordable housing, social security, gse portfolio, import price, domestic inflation
2007-Q2	subprime mortgage, finance premium, external finance, cash flow, job loss

2007-Q3	current account, desired saving, account balance, real interest, inflation expectation
2007-Q4	north carolina, acci n, inflation targeting, john s, structure economy
2008-Q1	higher priced, risk growth, loss mitigation, subprime arm, food price
2008-Q2	health care, primary dealer, bear stearns, commodity price, credit rating
2008-Q3	bear stearns, primary dealer, systemic risk, settlement system, price oil
2008-Q4	covered bond, mortgage securitization, deposit insurance, reserve treasury, systemic risk
2009-Q1	mutual fund, money mutual, s balance, lending program, press release
2009-Q2	community development, assessment program, capital assessment, capital buffer, loss rate
2009-Q3	merrill lynch, minority owned, america s, last fall, corporate bond
2009-Q4	asian economy, held reserve, liquidity risk, supervisory capital, category asset
2010-Q1	taylor rule, price appreciation, house price, housing bubble, output gap

2010-Q2	life satisfaction, well being, stress assessment, estate loan, community college
2010-Q3	small business, business owner, state government, participant noted, business reported
2010-Q4	fiscal rule, inflation rate, national income, budget deficit, sustainable rate
2011-Q1	banker speech, dodd frank, primary budget, business spending, state locality
2011-Q2	r d, banker speech, commodity price, community affair, d spending
2011-Q3	washington consensus, banker speech, economy scale, term prospect, price oil
2011-Q4	small business, flexible inflation, inflation targeting, business owner, framework monetary
2012-Q1	term unemployment, s law, okun s, see note, beveridge curve
2012-Q2	shadow system, commercial paper, lending standard, subprime mortgage, intraday credit
2012-Q3	well being, economic measurement, early childhood, agency mb, benefit cost
2012-Q4	capital flow, economy s, advanced economy, monetary fiscal, home purchase

2013-Q1	term premium, term rate, gold standard, expected inflation, real short
2013-Q2	stress testing, community development, low income, loss revenue, capital level
2013-Q3	early s, real bill, bill doctrine, great moderation, gold standard
2013-Q4	community banker, forward rate, rate guidance, mexico s, recent crisis
2014-Q1	woman s, productivity growth, l yellen, transparency accountability, fed s
2014-Q2	http www, slack labor, l yellen, small business, i believe
2014-Q3	labor slack, tighter monetary, stability risk, part time, macroprudential tool
2014-Q4	income wealth, wage growth, economic opportunity, bottom half, wealth distribution
2015-Q1	large firm, safety soundness, regulatory capture, supervision large, large institution
2015-Q2	equilibrium real, rhode island, real fund, real rate, economic mobility
2015-Q3	core inflation, run inflation, energy price, food energy, import price

2015-Q4	economic outlook, liscc firm, real gdp, have also, natural rate
2016-Q1	ioer rate, appreciation dollar, rrp operation, overnight rrp, price measured
2016-Q2	return text, neutral rate, labor improvement, economic development, baseline outlook
2016-Q3	regulatory capital, aggressive rule, be subject, future recession, interest reserve
2016-Q4	higher education, kansa city, aggregate demand, university baltimore, supply side
2017-Q1	workforce development, taylor rule, high school, http www, income community
2017-Q2	woman s, labor force, female labor, many woman, work family
2017-Q3	reform have, final rule, system is, market based, loss absorbing
2017-Q4	security holding, price inflation, food energy, unconventional tool, term premium
2018-Q1	chairman governor, chair yellen, are doing, s export, express appreciation
2018-Q2	capital flow, international monetary, s monetary, transparency accountability, monetary fund

2018-Q3	natural rate, real time, great inflation, http www, budget office
2018-Q4	phillips curve, community development, retrieved reserve, inflation rate, higher inflation
2019-Q1	rural community, labor force, force participation, economic development, new york
2019-Q2	business debt, leveraged loan, middle class, short term, risk stability
2019-Q3	great recession, figure panel, symmetric objective, second era, challenge monetary

3.4 Chapter Summary

In this chapter, the viability of text mining-based early detection of risk factors describing a financial crisis was investigated by applying the proposed hot topic detection framework to speeches made by the chairs of the Federal Reserve System. The primary contributions of this study are as follows. First, a hot topic detection framework capable of performing real-time analysis was proposed. Second, a new keyword-scoring method incorporating the temporal importance of keywords was proposed. Experimental results showed that the proposed method better filtered out meaningless terms such as “have been” than the TF-IDF-based method. Third, experimental results demonstrated that the proposed framework using central bankers’ speeches would greatly assist in identifying financial risk factors.

Chapter 4

Explainable Market Sentiment Analysis of News Articles

4.1 Background

The goal of sentiment analysis is the detection of the sentiment polarities of sentences, paragraphs, or documents based on textual content (Malo et al., 2014). Beyond academia, sentiment analysis has attracted significant attention in a number of industries owing to its applicability to a wide range of target populations including consumers, companies, banks, and the general public (Feng et al., 2022; Ruiz-Martínez et al., 2012; Vidanagama et al., 2022). Especially, numerous researchers have conducted sentiment analysis of news articles to estimate market sentiment, which refers to investors' overall attitude toward the financial market (Li et al., 2020c). News media has been described as the fundamental propagator of speculative price movements (Shiller, 2016), and extensive studies have suggested that the media would affect market sentiment (Campbell et al., 2012; Dougal et al., 2012; Engelberg & Parsons, 2011; Garcia, 2013; Hanna et al., 2020; Tetlock, 2007).

As indicated in Chapter 1, explainability—the degree to which an interested stakeholder can understand the key factors that led to a data-driven model's decision (Bracke et al., 2019; Bussmann et al., 2021)—has been considered an essential consideration in the financial domain (Mashrur et al., 2020). Hence, knowledge graphs (Cambria & Hussain, 2015a; Picasso et al., 2019; Xing et al., 2018, 2019) and lexicons, which are regarded as explainable text analysis tools that can provide evidence for the model's decision, have been widely used for sentiment analysis in the financial domain. Especially, lexicon-based sentiment analysis has been commonly conducted (Erçen et al., 2022; Gakhar & Kundlia, 2021; Song & Shin, 2019) because lexicon-based methods can achieve reasonable performance and provide a clear explanation to users (Brazdil et al., 2022).

The construction of a domain-specific lexicon is particularly important because the sentiment orientations of words can vary by domain (Wu et al., 2019). As described in Chapter 1, "liability," a term with a generally negative connotation, is neutral when used in the financial domain (Cortis et al., 2017). One of the most widely used financial domain-specific sentiment lexicons that treat the term "liability" as a neutral sentiment word is the Loughran-McDonald Word List (Loughran & McDonald, 2011). It comprises unigrams, which are words containing a single token. A unigram lexicon works well with words having relatively straightforward associated sentiments (e.g., "profitable," which implies a positive sentiment, or "unprofitable," which implies a negative sentiment).

However, a unigram-based lexicon might not be sufficient to capture the financial domain-specific ontology that the true sentiment of a given word can change significantly depending on the presence of directional expressions. For example, the word "cost" is typically associated with a negative sentiment; when it is juxtaposed with "decrease" to form the phrase "cost decrease," however, a positive sentiment is conveyed. The fact that even the Hugging Face pipeline (Jain, 2022) contextual representation-based method, one of the most powerful sentiment analysis tools, misclassifies the phrase "cost decrease" as strong negative starkly illustrates the importance of directional expressions. Given their significance, the presence of directional expressions should be carefully considered when conducting sentiment analysis of financial documents.

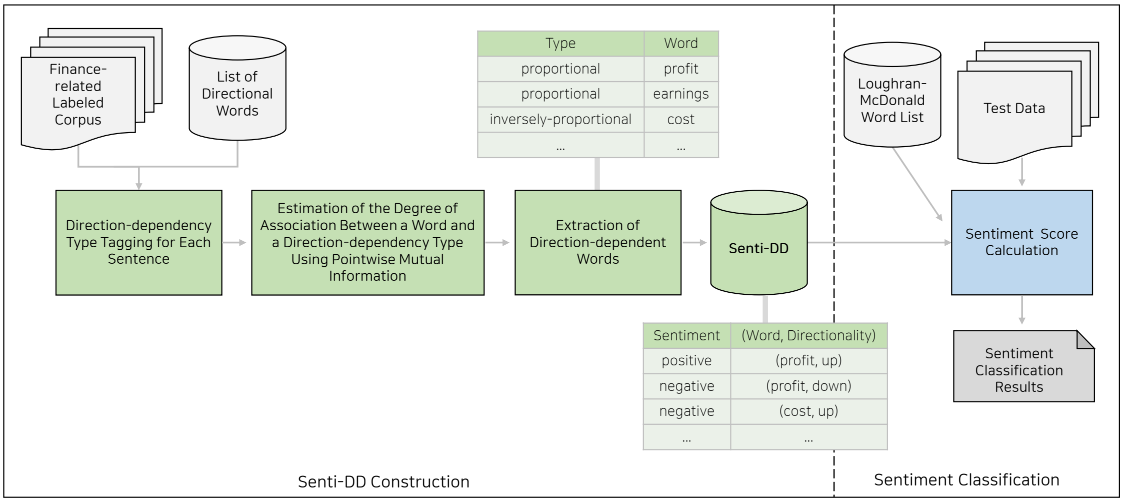
In this study, a sentiment lexicon named "sentiment lexicon composed of direction-dependent words" (Senti-DD) is proposed. Each element in Senti-DD is a pair comprising a directional word and a direction-dependent word. Senti-DD is constructed by directly extracting direction-dependent words, the sentiment labels of which change when combined with directional words, based on the measure of association between a word and its direction-dependency type. The proposed lexicon is then built by adding pairs comprising directional and direction-dependent words. Table 4.1 compares Senti-DD's lexical items with those proposed in previous studies.

Table 4.1: Examples of words in the lexicons proposed in previous studies

Study	Construction Approach	Positive	Negative
Harvard General Inquirer word lists (Stone et al., 1962)	Manual	("profit")	("tax") , ("liability")
Loughran-McDonald Word List (Loughran & McDonald, 2011)	Manual	("profitable")	("unprofitable")
Malo et al. (2014)	Manual	("profit" & "up")	("profit" & "down")
Oliveira et al. (2016)	Automatic	("more profit")	("less profit")
Senti-DD (proposed)	Automatic	("profit" & "up")	("profit" & "down")

4.2 Proposed Method

As shown in Fig. 4.1, the proposed data-driven sentiment analysis framework comprises two stages: (1) Senti-DD construction and (2) sentiment classification. In the first stage, Senti-DD is constructed by computing the PMI score as an estimate of a given word's direction-dependency type. In the second stage, the sentiment classification problem is solved using Senti-DD.



Finance-related Labeled Corpus
List of Directional Words
Type | Word
proportional | profit
proportional | earnings
inversely-proportional | cost
... | ...
Direction-dependency Type Tagging for Each Sentence
Estimation of the Degree of Association Between a Word and a Direction-dependency Type Using Pointwise Mutual Information
Extraction of Direction-dependent Words
Senti-DD
Sentiment | (Word, Directionality)
positive | (profit, up)
negative | (profit, down)
negative | (cost, up)
... | ...
Senti-DD Construction
Loughran-McDonald Word List
Test Data
Sentiment Score Calculation
Sentiment Classification Results
Sentiment Classification

Figure 4.1: Overview of the proposed sentiment analysis framework

4.2.1 Senti-DD construction

As a first step in constructing Senti-DD, polar sentences representing either positive or negative sentiment are gathered from a finance-related labeled corpus. Given a set of polar sentences and a word list with each word assigned the directional label “up” or “down,” the $UpScore$ and $DownScore$ of a subject sentence are defined as the number of “up” and “down” words, respectively, found in the sentence. Finally, each sentence, s, is given a direction score, $DirectionScore(s) = UpScore(s) - DownScore(s)$, that reflects the degree of direction conveyed by s.

Based on the relationship between its direction score and sentiment label, each sentence is assigned a tag representing a direction-dependency type. Two direction-dependency tags are used: “proportional” and “inversely proportional.” A sentence is tagged “proportional” if its sentiment is either positive with a direction score greater than zero or negative with a direction score less than zero. Similarly, a sentence is tagged “inversely proportional” if its sentiment is either positive with a direction score less than zero or negative with a direction score greater than zero. The proposed framework uses only “proportional” and “inversely proportional”-type sentences.

Then, each sentence is transformed into a list of nouns via tokenization and part-of-speech tagging using the Natural Language Toolkit library (Bird et al., 2009). All extracted words are lemmatized using the Natural Language Toolkit library, and words appearing more than five times within the entire corpus are retained.

The association between each word, w, and its direction-dependency type, which is either “proportional,” t_p, or “inversely proportional,” t_i, is measured using the following definition of the PMI score—$PMI(w,t) = log_2 \frac{p(w,t)}{p(w)p(t)}$—where $p(w,t)$ is the probability that a sentence of direction-dependency type t containing the word w is found in the subject corpus, $p(w)$ is the probability that w is found in the subject corpus, and $p(t)$ is the probability that a sentence of direction-dependency type t is found in the subject corpus. PMI is a popular lexical statistic that computes the intensity of coexistence between two variables. Previous studies have used PMI to generate sentiment words by calculating the degree of association between words and sentiments (Oliveira et al., 2016) or expand seed words by calculating the degree of association between given words (Yekrangi & Abdolvand, 2021; Yu et al., 2013).

To simplify the calculation, the dependency score of a given word w is defined as follows:

$$DependencyScore(w) = \begin{cases} |PMI(w,t_p)| & \text{if } PMI(w,t_p) - PMI(w,t_i) > \delta \\ 0 & \text{if } |PMI(w,t_p) - PMI(w,t_i)| \leq \delta \\ -|PMI(w,t_i)| & \text{if } PMI(w,t_p) - PMI(w,t_i) < -\delta \end{cases} \tag{4.1}$$

where $\delta \geq 0$ is a parameter that adjusts the number of direction-dependent entities; setting a larger value of δ decreases the number of entities, and vice versa.

Table 4.2: Demonstration of calculation of the dependency score for the word "profit"

Steps for score calculations	Score
# sentences containing "profit"	217
# "proportional" type sentences containing "profit"	216
# "inversely proportional" type sentences containing "profit"	1
# total sentences	719
# "proportional" type sentences	691
# "inversely proportional" type sentences	28
$p($**"profit"**$)$	$p(w) = 217/719 = 0.30$
$p($**"profit"**, **"proportional"**$)$	$p(w, t_p) = 216/719 = 0.30$
$p($**"profit"**, **"inversely proportional"**$)$	$p(w, t_i) = 1/719 = 0.0014$
$p($**"proportional"**$)$	$p(t_p) = 691/719 = 0.96$
$p($**"inversely proportional"**$)$	$p(t_i) = 28/719 = 0.04$
$PMI($**"profit"**, **"proportional"**$)$	$PMI(w, t_p) = log_2(0.30/0.30 * 0.96) = 0.05$
$PMI($**"profit"**, **"inversely proportional"**$)$	$PMI(w, t_p) = log_2(0.0014/0.30 * 0.04) = -3.0$
$DependencyScore($**"profit"**$)$	0.05

Table 4.2 demonstrates the calculation of the dependency score for the word "profit." A word with a positive dependency score is regarded as a candidate word of the "proportional" type, which represents a positive sentiment when used with "up" words and a negative sentiment when used with "down" words. Similarly, a word with a negative dependency score is regarded as a candidate word of the "inversely proportional" type, which represents a positive sentiment when used with "down" words and a negative sentiment when used with "up" words.high

Based on the relationship between the direction-dependency tag of a sentence and the dependency score of a word, a single representative word from each sentence is extracted according to the following rules: if a sentence is "proportional," the word with the highest dependency score among the candidate "proportional" words is extracted as a "proportional" type direction-dependent word. Conversely, if a sentence is "inversely proportional," the word with the lowest dependency score among the candidate "inversely proportional" words is extracted as an "inversely proportional" type direction-dependent word. For post-processing, words containing non-alphabet characters and words with less than three letters are treated as noise and filtered out.

To construct Senti-DD, pairs of words are created from the lists of directional and direction-dependent words, respectively. A pair comprising an "up" and a "proportional" word or a pair comprising a "down" and an "inversely proportional" word is labeled as a positive-context pair; similarly, a pair comprising an "up" and an "inversely proportional" word or a pair comprising a "down" and a "proportional" word is labeled as a negative-context pair.

4.2.2 Sentiment classification

Sentiment classification is performed based on an augmented lexicon combining the Loughran-McDonald Word List (Loughran & McDonald, 2011) and Senti-DD. Using the Loughran-McDonald Word List, the overall polarity of a sentence is determined; then, the score is refined using Senti-DD to capture the co-occurrence of direction-dependent and directional words. Finally, based on the refined score, the sentence is classified as a positive, negative, or neutral class.

For a given sentence, s, its $PosScore$ and $NegScore$ are defined as the number of positive and negative words in the Loughran-McDonald Word List, respectively, that it contains. The sentiment score of s is then computed as $SentimentScore(s) = PosScore(s) - NegScore(s)$.

To refine the sentiment score, it can be recalculated for a given context. Using Senti-DD, $ContPosScore$ and $ContNegScore$ are defined as the number of positive- and negative-context pairs found, respectively, in the subject sentence and the context score is computed

Table 4.3: Demonstration of calculation of refined score for the sentence “Profit for the period was EUR 10.9 mn, down from EUR 14.3 mn in 2009” (Malo et al., 2014)

Steps for score calculations	Score	Sentiment word
$SentimentScore$	$0 - 0 = 0$	
$ContextScore$	$0 - 1 = -1$	(**“profit”**, **“down”**)
$RefinedScore$	-1	

as follows:

$$ContextScore(s) = ContPosScore(s) - ContNegScore(s). \tag{4.2}$$

Finally, the sentiment score is refined based on the $ContextScore$ to reflect the extra positivity/negativity driven by the context of the sentence by adding one point to or subtracting one point from the sentiment score for $ContextScore > 0$ and $ContextScore < 0$, respectively. Table 4.3 demonstrates the calculation of a refined score.

Sentences with refined scores greater than, equal to, or less than zero are classified as positive, neutral, or negative, respectively.

4.3 Experiments

4.3.1 Data description

Three dataset containing labeled financial news headlines were used: the Financial Phrase Bank (FPB) (Malo et al., 2014), the dataset created for subtask 2 of Task 5 in SemEval 2017 (SemEval) (Cortis et al., 2017), and the dataset created for Task 1 of the financial opinion mining and question answering (FiQA) challenge (Maia et al., 2018). The FPB comprises 4,835 English sentences annotated by 16 experts in finance and business. The annotators were instructed to give a positive, negative, or neutral label according to how they thought the information in a sentence might affect the stock price of the mentioned company. Based on the level of agreement (50, 66, 75, and 100%) among the annotators, the FPB was divided into four subsets: DS50, DS66, DS75, and DS100, respectively. Each message in the SemEval database was annotated with a floating-point value between -1 (negative) and 1 (positive) denoting the sentiment expressed towards the mentioned company; a value of 0 denoted neutral sentiment. A total of 960 annotated sentences produced in the challenge were released to the public. Each of the 436 publicly available FiQA sentences was annotated with a target aspect sentiment score ranging from -1 (negative) to 1 (positive). In the experiment, the sentiment score of the aspect in a given sentence was treated as the sentiment score of the sentence; for sentences with multiple aspects, one was selected randomly and the rest were removed. As the original labels of the sentences in SemEval and FiQA have continuous sentiment scores, these sentences were categorized into positive, neutral, and negative classes if their scores were greater than, equal to, or less than zero, respectively. The characteristics of each dataset are listed in Table 4.4.

Table 4.4: Dataset characteristics

Dataset	% positive	% neutral	% negative	# sentences
DS50	28.2	59.3	12.5	4,835
DS66	27.8	60.0	12.2	4,209
DS75	25.7	62.1	12.2	3,447
DS100	25.2	61.4	13.4	2,259
SemEval	55.3	3.3	41.4	960
FiQA	64.4	2.8	32.8	436

To ensure the robustness of the results, a stratified five-fold cross-validation was conducted. The process was repeated five times, with each of the five folds used exactly once as test data, and the average was obtained.

4.3.2 Experimental settings

Directional words were defined following the experimental settings used in previous works (Krishnamoorthy, 2018; Malo et al., 2014). “Up” and “down” terms were formed by using the Harvard General Inquirer word lists (Stone et al., 1962) as seed lists, with words defined under the “increase” and “rise” categories classified as “up” terms and those under the “decrease” and “fall” categories classified as “down” terms. Following manual review, 20 terms were classified under the “up” category and 11 were classified under the “down” category. Table 4.5 presents the full list of carefully selected directional words. The words in each sentence as well as the directional words were then stemmed and compared for matches using the Natural Language Toolkit library (Bird et al., 2009).

Table 4.5: List of directional words

Directionality type	Words
Up	accelerate, advance, award, better, climb, double, faster, gain, grow, higher, increase, jump, quicken, rebound, recover, rise, rose, step-up, surge, up
Down	constrain, decelerate, decline, decrease, down, drop, fall, fell, slower, weaken, weaker

From the 4,835 sentences in the DS50 dataset, 691 sentences were tagged as *proportional* and 28 were tagged as *inversely proportional.* Table 4.6 lists examples of sentences with tags. To obtain as many direction-dependent entities as possible, the value of δ in Equation 4.1 was set to 0.

Seniment140 (Mohammad et al., 2013), SWN (Baccianella et al., 2010), SO-CAL (Taboada et al., 2011), MPQA (Wilson et al., 2005), TextBlob, VADER (Hutto & Gilbert, 2014), SentiStrength (Thelwall et al., 2010), AFINN (Nielsen, 2011), and Loughran-McDonald Word List (Loughran & McDonald, 2011) were used as baseline lexicons for comparison. In an experiment using Seniment140 and VADER, scores greater than or equal to 0.05,

Table 4.6: Examples of sentences tagged as “proportional” or “inversely proportional”

Direction-dependency	Sentences
Proportional	- “Orion’s net profit went up by 33.8% year-on-year to EUR 33 million.” - “Agricultural newspaper Maaseudun Tulevaisuus had 318,000 readers, representing a decrease of 6%.”
Inversely proportional	- “Unit costs for flight operations fell by 6.4%.” - “Operating loss increased to EUR 17 mn from a loss of EUR 10.8 million in 2005.”

less than or equal to -0.05, and between -0.05 and 0.05 were classified as positive, negative, and neutral, respectively. For Loughran-McDonald Word List and MPQA, the sentiment score was defined by subtracting the number of negative words from the number of positive words. In an experiment using SWN, SO-CAL, AFINN, TextBlob, SentiStrength, Loughran-McDonald Word List, and MPQA, a sentence was classified as positive, negative, or neutral if its polarity score was greater than, less than, or equal to zero, respectively.

The proposed method, which achieves reasonable performance using lexicon-based intuitive inference, was also compared with several pretrained models—Word2Vec model (Mikolov et al., 2013b) with logistic regression (Word2Vec), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019)—that achieve high performance but are treated as black boxes. Word2Vec obtained sentence embedding by averaging the embeddings of words within a sentence; using these sentence embeddings as feature vectors, logistic regression was then conducted. BERT and RoBERTa adopted fine-tuning approaches. For fine-tuning, an embedding layer was added on top of the existing hidden layers; following this, classification was conducted using the embedding vectors obtained for the given sentences as feature vectors.

In this experiment, word2vec-google-news-300[1], bert-base-uncased[2], and roberta-base[3] were used for Word2Vec, BERT, and RoBERTa models, respectively. For BERT and RoBERTa, model parameters were optimized using the Adam optimizer (Kingma & Ba, 2015) with a weight decay of 0.01. The batch size was set to eight, the learning rates were based on the warm-up schedule strategy proposed by Vaswani et al. (2017) with warm-up occurring over the first 500 steps, and the maximum number of training epochs was set to three.

4.3.3 Experimental results

Table 4.7 shows the classification performance of the respective methods on the sentiment classification task. All values are weighted average values, with the best value among lexicon-based methods per measure marked in bold. LM indicates Loughran-McDonald Word List.

The results on the four subsets of the FPB indicate that the proposed LM+Senti-DD consistently outperforms other baseline lexicons by achieving higher F1 scores. Considering that low levels of agreement imply low-quality labels, the results prove the robustness of the proposed method against variations in labeling quality. This indicates that the Senti-DD score refinement process, which reflects context by capturing the co-occurrence of directional and direction-dependent words, is effective when applied to documents with both high and low levels of consent for sentiment (high- and low-quality labels, respectively).

[1] `https://huggingface.co/fse/word2vec-google-news-300` [Accessed: December 15, 2022]
[2] `https://huggingface.co/bert-base-uncased` [Accessed: December 15, 2022]
[3] `https://huggingface.co/roberta-base` [Accessed: December 15, 2022]

Table 4.7: Experimental results for the classification task on the SemEval, FiQA, and four subsets of the FPB. LM indicates Loughran-McDonald Word List.

		Lexicons										Pretrained models		
Dataset	Measure	Sentiment 140	SWN	SO-CAL	MPQA	TextBlob	VADER	Senti Strength	AFINN	LM	LM+ Senti-DD	Word2Vec	BERT	RoBERTa
DS50	Precision	0.5345	0.4778	0.5161	0.5143	0.5155	0.6028	0.5641	0.6332	0.6147	**0.7090**	0.7392	0.8487	0.8578
	Recall	0.2196	0.3969	0.4546	0.4567	0.4852	0.5396	0.5713	0.5897	0.6232	**0.7055**	0.7466	0.8474	0.8523
	F1-score	0.1540	0.4107	0.4575	0.4630	0.4953	0.5452	0.5644	0.5960	0.5914	**0.7001**	0.7328	0.8457	0.8523
DS66	Precision	0.5514	0.4851	0.5248	0.5215	0.5275	0.6194	0.5794	0.6504	0.6337	**0.7389**	0.7650	0.8827	0.8948
	Recall	0.2112	0.4044	0.4588	0.4631	0.4968	0.5534	0.5849	0.6054	0.6363	**0.7315**	0.7741	0.8750	0.8826
	F1-score	0.1462	0.4194	0.4633	0.4708	0.5070	0.5599	0.5795	0.6130	0.6023	**0.7271**	0.7622	0.8757	0.8844
DS75	Precision	0.5500	0.4916	0.5444	0.5284	0.5426	0.6409	0.6063	0.6713	0.6507	**0.7796**	0.8017	0.9222	0.9366
	Recall	0.1984	0.4009	0.4616	0.4601	0.5039	0.5590	0.6069	0.6159	0.6556	**0.7702**	0.8103	0.9188	0.9341
	F1-score	0.1344	0.4199	0.4725	0.4726	0.5169	0.5702	0.6049	0.6265	0.6174	**0.7673**	0.7986	0.9185	0.9344
DS100	Precision	0.5581	0.4624	0.5431	0.5186	0.5476	0.6405	0.6137	0.6868	0.6377	**0.8238**	0.8125	0.8733	0.9642
	Recall	0.1948	0.3873	0.4723	0.4604	0.5228	0.5688	0.6171	0.6392	0.6476	**0.8128**	0.8185	0.8596	0.9615
	F1-score	0.1257	0.4058	0.4821	0.4733	0.5317	0.5770	0.6144	0.6477	0.5982	**0.8105**	0.8050	0.8462	0.9620
SemEval	Precision	0.5880	0.5879	0.6414	0.7154	0.6428	0.6855	0.6496	0.7055	**0.7929**	0.7858	0.7453	0.6427	0.8003
	Recall	**0.5271**	0.3500	0.3646	0.3708	0.2292	0.4688	0.2927	0.4469	0.2875	0.3344	0.7552	0.6615	0.8167
	F1-score	0.5175	0.4218	0.4470	0.4701	0.3046	**0.5447**	0.3635	0.5331	0.3632	0.4201	0.7458	0.5970	0.8029
FiQA	Precision	0.6081	0.6498	0.6421	0.7691	0.7329	0.7153	0.6935	0.7377	**0.8334**	0.8127	0.7610	0.5356	0.4182
	Recall	**0.4863**	0.3624	0.3716	0.3900	0.2546	0.4679	0.2730	0.4471	0.2890	0.3325	0.7683	0.5502	0.6444
	F1-score	0.4924	0.4418	0.4526	0.4980	0.3459	**0.5513**	0.3327	0.5403	0.3635	0.4130	0.7413	0.5186	0.5064

By contrast, for SemEval and FiQA, in which the content of sentences is expressed more implicitly, most of the lexicons produce low F1 scores of less than 0.5. This is because lexicon-based methods can degrade sentiment analysis performance if sentiment words are not explicitly expressed within a sentence. For example, the sentence "BT finance director Tony Chanmugam to step down," (Cortis et al., 2017) which is labeled as a negative class in SemEval, is classified as a neutral class by LM+Senti-DD because the sentence contains no sentiment words. If the sentence is re-expressed as "BT finance director Tony Chanmugam, who made a big investment, to step down," LM+Senti-DD would correctly classify the sentence as a negative class based on the words "investment" and "down."

Pretrained models demonstrate an improved performance relative to lexicon-based methods on nearly all measures. This is because pretrained models trained on various sources of data such as Wikipedia can detect contexts that are not expressed in direction-dependent words. For example, the sentence "Cuts equivalent to the costs of about 35-45 employees are the target, the company said.," (Malo et al., 2014) which is labeled as a negative class in DS50, is incorrectly classified as a neutral class by LM+Senti-DD but correctly classified as a negative class by pretrained models such as BERT and RoBERTa. It is presumed that these pretrained models detect that the sentence refers to layoffs; then, they classify the sentence as negative based on the Wikipedia-driven knowledge that layoffs occur in crises. However, pretrained models that perform artificial neural network-based operations have a disadvantage arising from the fact that their inference bases are considered black boxes. Thus, this study focused on building a lexicon that provides an intuitive reasoning basis, albeit with a performance worse than that of pretrained models.

Table 4.8: Experimental results for the classification task on the DS100 dataset. LM indicates Loughran-McDonald Word List.

		Lexicons										Pretrained models		
Class	Measure	Sentiment 140	SWN	SO-CAL	MPQA	TextBlob	VADER	Senti Strength	AFINN	LM	LM+ Senti-DD	Word2Vec	BERT	RoBERTa
Positive	Precision	0.1797	0.2284	0.3158	0.3026	0.3439	0.3889	0.4465	0.4574	0.6371	**0.8528**	0.6886	0.7025	0.9422
	Recall	0.1864	0.3132	0.6121	0.5208	0.3931	**0.7384**	0.3922	0.7298	0.1632	0.6101	0.6959	0.8243	0.9506
	F1-score	0.1827	0.2637	0.4163	0.3827	0.3668	0.5085	0.4172	0.5615	0.2596	**0.7106**	0.6918	0.7562	0.9458
Negative	Precision	0.1782	0.2562	0.2282	0.3711	0.3689	0.3479	0.3321	0.4196	0.4184	**0.6264**	0.7565	0.9029	0.8928
	Recall	**0.9466**	0.4690	0.1475	0.3317	0.4768	0.2255	0.3652	0.3369	0.4004	0.8340	0.3767	0.4210	0.9607
	F1-score	0.2997	0.3310	0.1782	0.3497	0.4143	0.2720	0.3471	0.3694	0.4087	**0.7145**	0.5008	0.5250	0.9234
Neutral	Precision	0.7959	0.6017	0.7041	0.6385	0.6686	0.8058	0.7435	0.8392	0.6852	**0.8544**	0.8739	0.9372	0.9877
	Recall	0.0332	0.3988	0.4857	0.4629	0.5850	0.5743	0.7650	0.6686	**0.9005**	0.8918	0.9653	0.9705	0.9663
	F1-score	0.0633	0.4790	0.5748	0.5366	0.6238	0.6705	0.7539	0.7439	0.7782	**0.8725**	0.9173	0.9532	0.9766

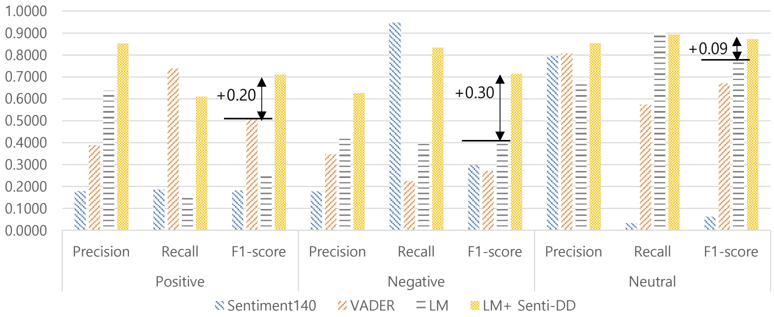


Figure 4.2: Graphical comparison of the results for the DS100 dataset

Table 4.8 and Fig. 4.2 show the experimental results on the DS100 dataset in detail. Interestingly, the F1 score for the negative class achieved by LM+Senti-DD is nearly twice the score achieved by LM. This indicates that LM+Senti-DD generally outperforms other lexicons, particularly in classifying sentences into positive and negative classes. Although VADER achieves a high recall for the positive class, its precision is significantly lower than that of LM+Senti-DD; consequently, LM+Senti-DD achieves a higher F1 score. It appears that the rules for calculating sentiment scores in VADER tend to be biased toward predicting a large number of positive sentiments. Similarly, Sentiment140 records a high recall and relatively lower precision for the negative class, which can be attributed to Sentiment140's bias toward predicting a substantial number of negative sentiments.

The performance improvement achieved by LM+Senti-DD can be attributed to its ability to reflect context by incorporating the effects of directional words when classifying sentences. For example, LM misclassifies the sentence “Profit for the period was EUR 10.9 million, down from EUR 14.3 million in 2009” (Malo et al., 2014) as neutral because the sentence contains neither positive nor negative words in the LM. The proposed method, however, correctly classifies the sentence as negative because it contains both “down” and “profit” as “down-” and “proportional-” type words. Fig. 4.3 illustrates this example of sentiment analysis using Senti-DD. It should be noted that the key idea underlying the proposed framework is to develop financial domain-specific clues that achieve extremely high performance in predicting positive or negative sentiments at the cost of a relatively low recall for the neutral class.

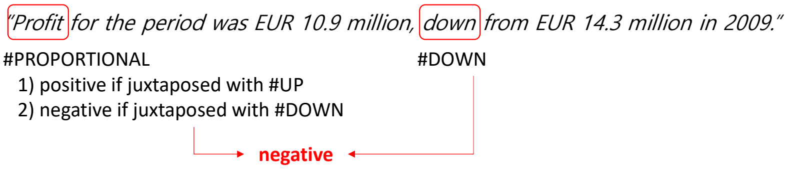

"Profit for the period was EUR 10.9 million, down from EUR 14.3 million in 2009."
#PROPORTIONAL
1) positive if juxtaposed with #UP
2) negative if juxtaposed with #DOWN
#DOWN
negative

Figure 4.3: Example of sentiment analysis using Senti-DD

Table 4.9 lists all of the direction-dependent words extracted from the DS50 dataset. The table lists 73 “proportional” and 7 “inversely proportional”-type words in alphabetical order. As described in Section 4.2, “proportional”-type words can lead to a positive or negative sentiment when combined with an “up” or a “down” type word, respectively, and the opposite applies to “inversely proportional”-type words. A majority of the words appear to be appropriately identified. Intuitively, the terms “capital,” “demand,” “investment,” “profit,” and “revenue” are correctly listed as “proportional” words; and the term “cost” is correctly listed as “inversely proportional” word.

As indicated in Section 4.3.2, the number of sentences in the DS50 dataset tagged as “inversely proportional” is relatively small, leading to a small number of “inversely proportional”-type words. Furthermore, the imbalance between the number of “proportional” and “inversely proportional”-type sentences appears to produce noisy words such as “beer” and “day” that are not intuitively interpreted as direction-dependent words. This is due to the possibility that some words that should be frequently used regardless of direction-dependency types might appear only in certain types of sentences and not in other types

Table 4.9: Direction-dependent words extracted from the entire DS50 dataset

Direction-dependency	Words
Proportional	acquisition, agreement, area, beer, brewery, business, capital, cargotec, cash, cent, communication, contract, currency, customer, demand, division, ebit, efficiency, electronics, end, eur, finnair, food, group, growth, income, interest, investment, item, june, konecranes, liter, maker, management, manufacturer, march, margin, medium, metal, mln, month, net, news, order, orion, oyj, paper, passenger, percent, period, phone, product, profit, property, pulp, quarter, report, revenue, sale, september, share, solution, system, teleste, time, tonne, trade, turnover, use, value, volume, world, year
Inversely proportional	company, construction, cost, day, plant, result, traffic

of sentences. Notwithstanding these limitations, these preliminary results demonstrate the possibility of automatically acquiring direction-dependent words using the proposed PMI-based method.

4.4 Chapter Summary

In this chapter, a financial domain-specific sentiment lexicon, Senti-DD, was proposed. Senti-DD identified “acquisition,” “agreement,” and “communication” as “proportional” type words that create a positive/negative sentiment when they are combined with an up/down-type word.

The main contributions of this study are as follows. First, a data-driven method for automatically extracting direction-dependent words was proposed. Second, a framework integrating Senti-DD as a plug-in lexicon to an existing lexicon was proposed to achieve enhanced sentiment classification performance. Third, in-depth experiments were carried out to compare the proposed lexicon with other conventional lexicons. Experimental results showed that the proposed lexicon could perform reasonably well even when compared to pretrained models.

Chapter 5

Spam Filtering on Real-time Data Feeds

5.1 Background

Filtering spam messages is one of the essential tasks in financial text analysis because disinformation proliferating on social media threatens investors and automated trading systems that rely on social information for predicting stock prices (Tardelli et al., 2022). For example, a recent investigation revealed that fake online discussions resulted in a rapid increase in the stock price of Cynk Technology (Ferrara et al., 2016). However, the stock price plummeted shortly after, resulting in significant losses to investors (Cresci et al., 2019). To prevent such incidents and safeguard investors' investments, disinformation distributed in financial microblogs should be preemptively filtered.

Researchers have previously discovered that *cashtag piggybacking* is one of the most typically observed types of stock-related spam messages (Cresci et al., 2018, 2019; Orabi et al., 2020; Tardelli et al., 2022). As described in Chapter 1, *cashtag piggybacking* refers to malicious practices that attempt to promote low-value stocks by exploiting the popularity of high-value ones (Cresci et al., 2019).

While accurate spam filtering can be done by manually checking spam messages, manual inspection has the fatal disadvantage of being very costly as the process requires significant time and human labor. On the other hand, data-driven approaches are inexpensive in terms of time and human labor, thereby supporting real-time spam filtering. This, therefore, enables real-time delivery of data to systems for automated financial risk assessment.

One of the most promising approaches of the automatic inspection of spam messages would be the use of pretrained language models. Although existing financial domain-specific pretrained language models contain enriched linguistic knowledge that can capture the semantic roles of words and the syntactic structures of sentences, they can be further improved by enhancing company-related factual knowledge.

Acquiring company-related factual knowledge is particularly important for understanding the context of financial texts (Elhammadi et al., 2020; Wang et al., 2021a). In most documents that refer to a company name, providing all relevant information regarding the company is impossible; thus, the readers of these documents are assumed to possess background knowledge related to the company. For instance, many investors would tag only the company name on their microblogs to evaluate the company's stock price but rarely provide additional information such as the company's market capitalization. Understandably, a language model with company-related information, which contains information regarding the company's market capitalization, flagship products or services, and the industry to which it belongs, can better interpret the context of financial texts than general-purpose language models. It is speculated that a company-related factual knowledge-enhanced language model can effectively filter spam messages that promote non-blue-chip stocks as if they are blue-chip stocks.

Most previous studies regarding factual knowledge enhancement methods of pretrained language models have used Wikipedia as a knowledge base. However, to extract company-level information from texts, Wikipedia—which includes a wide range of topics—presents several limitations, i.e., it requires additional procedures in manual or automatic data processing to classify documents into company-related categories. Furthermore, when evaluating the performance of a model, only data issued in a certain period must be used to avoid look-ahead bias. However, owing to the characteristics of Wikipedia, which reflects all updates on a single webpage without storing historical data, newly generated articles in Wikipedia are difficult to distinguish.

In this study, an easily implementable knowledge incorporation framework using Form 10-K filings as a textual knowledge base is proposed to enhance factual knowledge of the model. This framework applies a novel company name masking method that masks tokens corresponding to company names, thus allowing the model to learn company-related factual information in a sentence. The proposed framework adopts a post-training approach instead of forcing the model to be pretrained from scratch. Post-training refers to the process of using the pretrained weights of the model for initialization and further training the model to perform new tasks using unlabeled data (Du et al., 2020). The first advantage of this approach is that the token representations are updated efficiently because the model is not trained from scratch and only performs necessary updates (Xu et al., 2019). Second, the model can achieve gradual transfer learning by reflecting on the knowledge obtained from pre and post-training processes (Whang et al., 2020). Post-training approaches have been used extensively to obtain linguistic knowledge regarding new domains that has never been revealed in the pretraining process (Du et al., 2020; Luo et al., 2021; Whang et al., 2020; Xu et al., 2019).

5.2 Proposed Method

The proposed framework incorporates company-related factual knowledge, which is essential for solving real-world problems in the financial domain, into pretrained language models. As illustrated in Fig. 5.1, the framework comprises two steps: (1) *identification and normalization of company names* for preprocessing and (2) *company name masking* for post-training.

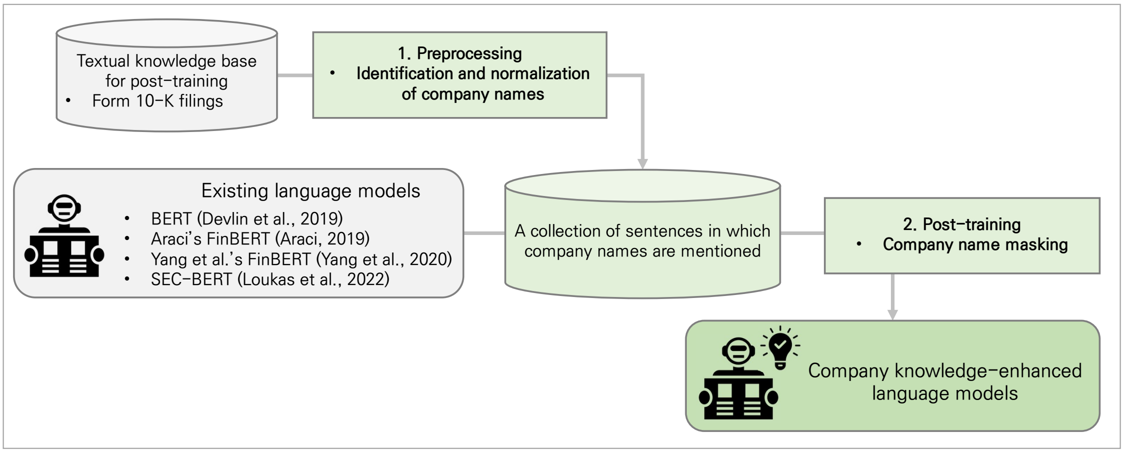

Textual knowledge base
for post–training
• Form 10–K filings
1. Preprocessing
• Identification and normalization of company names
Existing language models
• BERT (Devlin et al., 2019)
• Araci's FinBERT (Araci, 2019)
• Yang et al.'s FinBERT (Yang et al., 2020)
• SEC–BERT (Loukas et al., 2022)
A collection of sentences in which company names are mentioned
2. Post–training
• Company name masking
Company knowledge–enhanced language models

Figure 5.1: Overview of the proposed company-related factual knowledge incorporation framework

5.2.1 Identification and normalization of company names

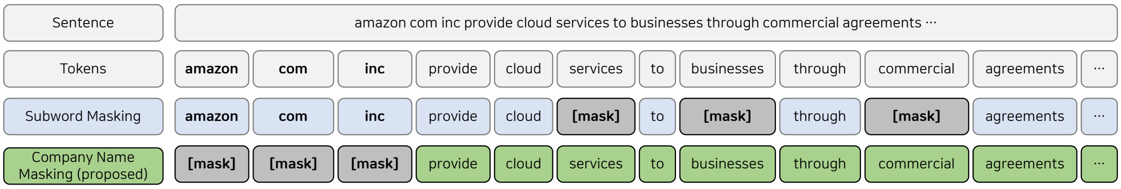
In Form 10-K filing, the company name appears in various forms; thus, the various forms representing one company must be identified and then normalized into one unified form. In this study, several heuristic rules were applied to identify company names in Form 10-K filings. First, special characters were removed, and all letters in a text were converted to lowercase. Subsequently, the terms “we” and “the company” were converted into the company name that published the corresponding filing. Furthermore, when the company name

Table 5.1: Example of preprocessing results

Before preprocessing	After preprocessing
We completed our initial public offering in May 1997 and our common stock is listed on the NASDAQ Global Select Market under the symbol AMZN. ***The company*** ...	***amazon com inc*** completed our initial public offering in may 1997 and our common stock is listed on the nasdaq global select market under the symbol amzn. ***amazon com inc*** ...
As used herein, ***Amazon.com***, and similar terms include ***Amazon.com, Inc.***	as used herein ***amazon com inc*** and similar terms include ***amazon com inc***

comprises three or more words separated by spaces or punctuation marks, the combination of words obtained by removing the last word was regarded as the company name. For example, the words "amazon com," and "amazon" in Form 10-K filing for Amazon.com, Inc. were converted into the word "amazon com inc." Table 5.1 presents an example of preprocessing results for a sentence in Form 10-K filing of Amazon.com, Inc.

5.2.2 Company name masking



Sentence	amazon com inc provide cloud services to businesses through commercial agreements ···										
Tokens	amazon	com	inc	provide	cloud	services	to	businesses	through	commercial	agreements	···
Subword Masking	amazon	com	inc	provide	cloud	[mask]	to	[mask]	through	[mask]	agreements	···
Company Name Masking (proposed)	[mask]	[mask]	[mask]	provide	cloud	services	to	businesses	through	commercial	agreements	···

Figure 5.2: Illustration of subword masking and the proposed company name masking methods

Among the preprocessed sentences, only the sentences in which the company names appear were used for post-training. For each sentence, a conventional masked language modeling objective (Devlin et al., 2019) was applied, which masked 15% of the tokens in a sentence and enforced a model to predict the masked tokens using representations of other unmasked tokens in the sentence. The most typically used masking method for masked language modeling is subword masking (Devlin et al., 2019), which randomly masks tokens. However, the proposed company name masking differs from the conventional subword masking in that it prioritizes masking tokens that correspond to company names, resulting in a nonrandom process. In the experiment using the proposed company name masking method, a whole-word masking (Cui et al., 2021) strategy, which masked all tokens for the company name, was adopted. If the number of tokens for all company names in the sentence does not exceed the total number of tokens to be masked, then other individual tokens may be masked. Conversely, if the number of tokens for the company name in the sentence exceeds the total number of tokens to be masked, then only other individual tokens become candidates to be masked. Fig. 5.2 shows the differences between the proposed company name masking and conventional subword masking methods. The proposed company name masking method allows the model to predict the masked tokens corresponding to the company name based on the representations of other unmasked tokens surrounding the company name. Hence, the model can learn the context in which a company name appears, thereby improving company-related knowledge.

5.3 Experiments

5.3.1 Evaluation framework

When financial "fill-in-the-blank" statements that can directly test a model's knowledge (e.g., "As of August 2022, the company with the largest market capitalization is _____.") are unavailable, evaluating the enhanced level of company-related factual knowledge in the model becomes extremely challenging. Thus, this study propose to assess the performance of three knowledge incorporation methods—no post-training, post-training with subword masking, and post-training with company name masking—by performing spam filtering as a downstream task. Fig. 5.3 presents an overview of the proposed evaluation framework. Spam filtering is done by fine-tuning existing language models applied with no

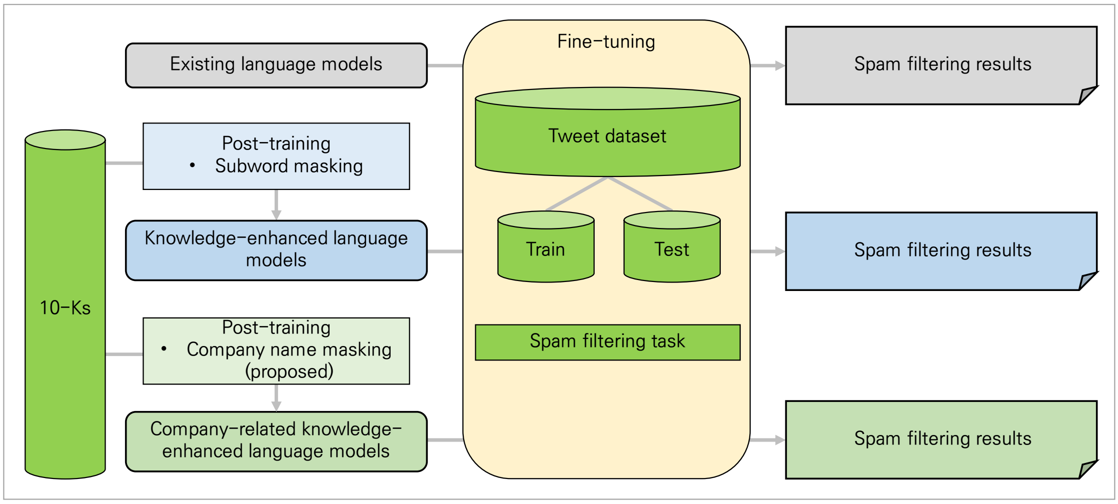


Figure 5.3: Overview of the proposed evaluation framework performing spam filtering as a downstream task

post-training, existing language models applied with post-training using subword masking, and existing language models with post-training. The tweet dataset was used as the data needed for the training phase of the fine-tuning task and the data needed to measure performance in the testing phase.

5.3.2 Data description

For post-training, Item 1 sections of Form 10-K filings were used as a textual knowledge base. First, information regarding 12,057 companies registered with the SEC as of August 2022 was acquired using the SEC CIK Mapper[1], which is a Python package that provides data regarding stocks and their Central Index Key information listed on the SEC. The companies' information was acquired based on the time at which the experiment was conducted, as companies registered in the SEC based on a specific period could not be identified. Subsequently, using the SEC-API.io[2] library, 3,999 pieces of Form 10-K filings

[1] `https://sec-cik-mapper.readthedocs.io` [Accessed: December 15, 2022]

[2] `https://sec-api.io` [Accessed: December 15, 2022]

Table 5.2: Characteristics of the post-training dataset comprising Item 1 section of Form 10-K filings

Total # of Item 1 sections	3,990
Total # of the sentences in Item 1 sections	1,048,470
Total # of the sentences containing company names	487,012

published by these companies in 2016 were obtained. Reports issued in 2017 or later were not included. Because the tweet dataset published in 2017 will be used in the evaluation framework, reports published later than those tweets may result in look-ahead bias. Finally, Item 1 sections from these reports were extracted. The section was extracted using the SEC-API.io library, where 3,990 documents were successfully acquired, excluding nine cases where technical problems occurred. As the documents contained unnecessary HyperText Markup Language (HTML) tags, the Python html[3] library was used to exclude HTML tags. A summary of the final dataset of the Item 1 sections used in the experiment is presented in Table 5.2.

To construct datasets for fine-tuning, tweets that satisfy the following conditions were extracted from a benchmark stock-related tweet dataset (Cresci et al., 2018): (1) written in English; (2) labeled as either "human" or "bot." To facilitate interpretation, labels "human" and "bot" were changed to "non-spam" and "spam," respectively. After removing tweets with duplicate content, the tweets were organized in chronological order. Tweets issued in May, June, July, and August 2017 were classified as training data, whereas tweets issued in September 2017 were classified as test data. While constructing one test dataset containing 10,000 tweets, multiple training datasets were constructed by changing the number of tweets from 400 to 2,000 to observe the effect of the amount of training data on the performance of a model. In the experiment, eight training datasets with 400, 500,

[3] `https://docs.python.org/3/library/html.html` [Accessed: December 15, 2022]

600, 700, 800, 900, 1,000, 2,000, 10,000, 20,000, and 40,000 tweets were used. Each dataset contained the same number of spam and non-spam messages.

Several preprocessing steps were performed on each tweet. The combination of the dollar character and a ticker (e.g., $APPL) was converted into the corresponding company name (e.g., Apple Inc.); HTML tags and special characters were removed; all letters in the text were converted to lowercase; usernames were removed; and URLs were replaced with a special token [URL].

5.3.3 Experimental settings

BERT (Devlin et al., 2019), Araci's FinBERT (Araci, 2019), Yang et al.'s FinBERT (Yang et al., 2020), and SEC-BERT (Loukas et al., 2022) were used as base models in the experiment. To perform spam filtering, the models were fine-tuned to classify a document as "spam" or "non-spam." For fine-tuning, an embedding layer was added on top of the existing hidden layers; subsequently, classification was conducted using the embedding vectors obtained for the sentences as feature vectors.

To perform post-training, fine-tuning, and test procedures, a computer with two GTX 1080 Ti GPUs was used. The batch size was set to eight; the learning rates were based on the warm-up schedule strategy proposed by Vaswani et al. (2017), with warm-up occurring over the first 500 steps; and the maximum number of training epochs was set to three.

5.3.4 Experimental results

Table 5.3 and Fig. 5.4 compare the performances of the following post-training methods on the spam filtering task: no post-training, post-training with subword masking, and post-training with company name masking.

Table 5.3: Accuracy scores of post-training methods using various sizes of training data for fine-tuning. For each method, four models—BERT (Devlin et al., 2019), Araci's FinBERT (Araci, 2019), Yang et al.'s FinBERT (Yang et al., 2020), and SEC-BERT (Loukas et al., 2022)—were used as base models, and the average scores for these models are reported. The best value among the methods is indicated in bold.

# of training data used for fine-tuning	no post-training	post-training with subword masking	post-training with company name masking (proposed)
400	53.29	58.28	**67.08**
500	61.51	71.11	**74.71**
600	64.48	66.44	**77.13**
700	66.47	73.00	**76.88**
800	67.20	73.02	**76.77**
900	70.50	77.52	**79.46**
1000	73.94	78.27	**79.70**
2000	80.31	80.84	**81.71**
10000	84.43	**84.67**	84.54
20000	**85.12**	85.12	84.87
40000	85.58	**85.75**	85.19

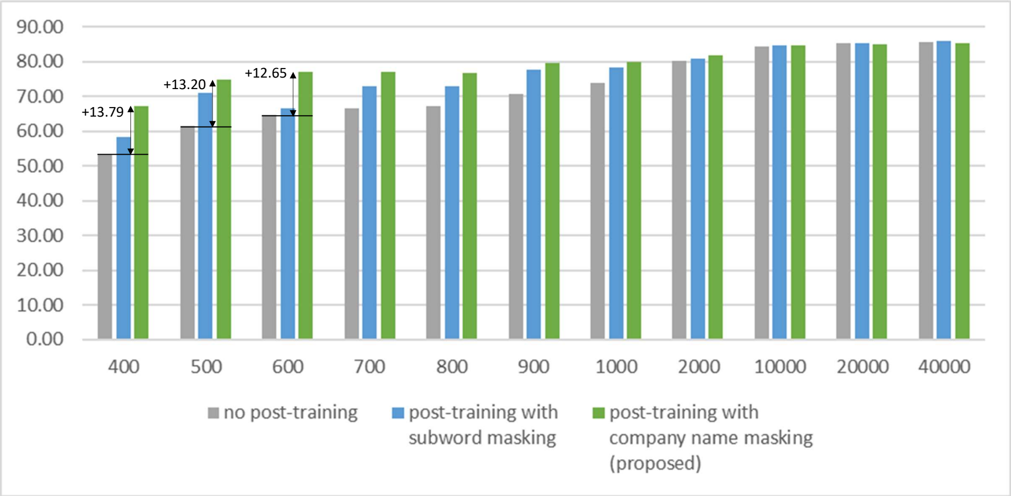


Figure 5.4: Graphical comparison of accuracy scores of post-training methods

For low-resource settings with fine-tuning data sizes of 400, 500, and 600, the proposed company name masking method significantly outperformed the baseline methods, as evidenced by its higher accuracy in classifying stock-related spam messages. Specifically, the proposed method recorded 13.79, 13.20, and 12.65 higher accuracy than the no post-training method when the number of fine-tuning data was 400, 500, and 600, respectively. This suggests that company-related factual knowledge, which was not obtained in the fine-tuning stage due to insufficient amount of training data, was incorporated into the model through the post-training process using the company name masking method; the incorporated knowledge appears to have improved the spam-filtering performance.

For low- and mid-resource settings with fine-tuning data sizes from 400 to 2000, the application of post-training—using company name masking or subword masking methods—showed superior performance compared with the performance without post-training. This implies that Form 10-K filings, which were used as post-training data, are an excellent textual knowledge base for the financial domain.

Meanwhile, for rich-resource settings with fine-tuning data sizes of 10000, 20000, and 40000, the performance differences between the methods were not noticeable. Although the baseline methods recorded better spam filtering performance than the proposed method, the differences between them were all below 0.6. The results are consistent with Mehrafarin et al. (2022)'s finding that models fine-tuned on large datasets would show similarly high performance regardless of their pretraining datasets.

Table 5.4, Table 5.5, and Table 5.6 present the accuracy scores of fine-tuned models in low-resource settings (e.g., training data with sizes of 400, 500, and 600), mid-resource settings (e.g., training data with sizes of 700, 800, 900, 1000, and 2000), and rich-resource settings (e.g., training data with sizes of 10000, 20000, and 40000), respectively. Each of the four models—BERT (Devlin et al., 2019), Araci's FinBERT (Araci, 2019), Yang et al.'s FinBERT (Yang et al., 2020), and SEC-BERT (Loukas et al., 2022)—was applied to three post-training methods—no post-training (*No PT*), post-training with subword masking (*PT with SM*), and post-training with company name masking (*PT with CM*). The term "performance improvement" in the tables indicates the subtraction of the score obtained through the *No PT* method from the score obtained through the *PT with CM* method.

As shown in Table 5.4 and Table 5.5, the proposed company name masking method generally outperformed other baseline methods at low- and mid-resource settings. Considering that the maximum value of performance improvement in low-resource settings is 23.66 and that in mid-resource settings is 10.62, the proposed method contributes to spam filtering more significantly in low-resource settings.

In particular, Araci's FinBERT (Araci, 2019), which was pretrained using financial news articles, achieved the most significant performance improvement in both low- and mid-resource settings; and SEC-BERT (Loukas et al., 2022), which was pretrained using Form 10-K filings, achieved the lowest performance improvement in both low- and mid-resource settings. These results suggest that complementary points existed between the knowledge acquired from financial news articles and that acquired from Form 10-K filings, thus allowing Araci's model to acquire additional knowledge that had not been learned previously. By contrast, SEC-BERT, which used one data source (e.g., Form 10-K filings) for both pre and post-training, appears to have failed to obtain additional knowledge during post-training.

Table 5.4: Accuracy scores of the models that were fine-tuned using training data with sizes of 400, 500, and 600. Each of the four models—BERT (Devlin et al., 2019), Araci's FinBERT (Araci, 2019), Yang et al.'s FinBERT (Yang et al., 2020), and SEC-BERT (Loukas et al., 2022)—was applied to three post-training methods—no post-training (*No PT*), post-training with subword masking (*PT with SM*), and post-training with company name masking (*PT with CM*). Performance improvement indicates the subtraction of the score obtained through the *No PT* method from the score obtained through the *PT with CM* method. The best value among the methods per model is indicated in bold.

# of training data for fine-tuning	BERT (Devlin et al., 2019)			Araci's FinBERT (Araci, 2019)			Yang et al.'s FinBERT (Yang et al., 2020)			SEC-BERT (Loukas et al., 2022)		
	No PT	PT with SM	PT with CM (proposed)	No PT	PT with SM	PT with CM (proposed)	No PT	PT with SM	PT with CM (proposed)	No PT	PT with SM	PT with CM (proposed)
400	62.26	50.74	**67.98**	41.49	**60.96**	60.21	48.54	58.82	**72.70**	60.86	62.61	**67.41**
500	66.64	71.24	**72.54**	50.45	70.23	**75.84**	60.76	73.45	**77.39**	68.20	69.50	**73.06**
600	66.39	62.08	**76.39**	52.91	61.99	**79.77**	68.82	72.67	**76.13**	69.79	69.02	**76.24**
Average	65.10	61.35	**72.30**	48.28	64.39	**71.94**	59.37	68.31	**75.41**	66.28	67.04	**72.24**
Performance improvement	+0		+7.21	+0		+23.66	+0		+16.03	+0		+5.95

Table 5.5: Accuracy scores of the models that were fine-tuned using training data with sizes of 700, 800, 900, 1000, and 2000. Each of the four models—BERT (Devlin et al., 2019), Araci's FinBERT (Araci, 2019), Yang et al.'s FinBERT (Yang et al., 2020), and SEC-BERT (Loukas et al., 2022)—was applied to three post-training methods—no post-training (*No PT*), post-training with subword masking (*PT with SM*), and post-training with company name masking (*PT with CM*). Performance improvement indicates the subtraction of the score obtained through the *No PT* method from the score obtained through the *PT with CM* method. The best value among the methods per model is indicated in bold.

# of training data for fine-tuning	BERT (Devlin et al., 2019)			Araci's FinBERT (Araci, 2019)			Yang et al.'s FinBERT (Yang et al., 2020)			SEC-BERT (Loukas et al., 2022)		
	No PT	PT with SM	PT with CM (proposed)	No PT	PT with SM	PT with CM (proposed)	No PT	PT with SM	PT with CM (proposed)	No PT	PT with SM	PT with CM (proposed)
700	64.16	70.64	**75.79**	62.61	74.94	**79.63**	68.51	74.44	**78.35**	70.58	71.96	**73.75**
800	64.76	**74.69**	70.76	64.65	68.07	**80.42**	69.96	76.29	**79.02**	69.41	73.04	**76.87**
900	71.68	**79.83**	78.54	70.11	77.94	**80.43**	71.13	79.17	**80.29**	69.08	73.14	**78.57**
1000	73.14	**80.26**	77.32	73.41	79.77	**80.47**	74.50	78.74	**80.63**	74.72	74.29	**80.37**
2000	80.81	80.93	**82.35**	78.89	80.69	**81.84**	81.09	80.70	**82.46**	80.44	**81.04**	80.17
Average	70.91	**77.27**	76.95	69.93	76.28	**80.56**	73.04	77.87	**80.15**	72.85	74.69	**77.95**
Performance improvement	+0		+6.04	+0		+10.62	+0		+7.11	+0		+5.1

Additionally, Yang et al.'s FinBERT with the company name masking method obtained lower performance improvement (e.g., 16.03 in low-resource settings and 7.11 in mid-resource settings) than that of Araci's model (e.g., 23.66 in low-resource settings and 10.62 in mid-resource settings). However, Yang et al.'s FinBERT showed excellent spam filtering performance (e.g., 75.41 in low-resource settings and 80.15 in mid-resource settings) comparable to Araci's model (e.g., 71.94 in low-resource settings and 80.56 in mid-resource settings), based on the average scores shown in the tables. It appears that various data sources that were used to pretrain Yang et al.'s FinBERT—such as analyst reports and earnings conference call transcripts—created a synergy effect owing to the consideration of knowledge obtained from both pre and post-training processes.

Meanwhile, BERT, which possessed the least financial knowledge compared with the other base models, achieved improved performance by only 7.21 in low-resource settings and 6.04 in mid-resource settings. It appears that BERT failed to achieve gradual domain adaptation—which refers to fitting a model that has been trained on particular datasets on new datasets (Patel et al., 2015)—because of the significant difference in knowledge obtained from pre and post-training. Hence, these results demonstrate that using various types of data related to the financial domain for both pre and post-training is an effective approach to incorporate financial knowledge.

Table 5.6 shows detailed results supporting the previously identified finding that the performance differences between the methods in rich-resource settings are negligible. While a few cases indicate that the baseline methods—*No PT* and *PT with SM* methods—outperform the proposed method in spam filtering, the differences between them are trivial.

Table 5.6: Accuracy scores of the models that were fine-tuned using training data with sizes of 10000, 20000, and 40000. Each of the four models—BERT (Devlin et al., 2019), Araci's FinBERT (Araci, 2019), Yang et al.'s FinBERT (Yang et al., 2020), and SEC-BERT (Loukas et al., 2022)—was applied to three post-training methods—no post-training (*No PT*), post-training with subword masking (*PT with SM*), and post-training with company name masking (*PT with CM*). Performance improvement indicates the subtraction of the score obtained through the *No PT* method from the score obtained through the *PT with CM* method. The best value among the methods per model is indicated in bold.

# of training data for fine-tuning	BERT (Devlin et al., 2019)			Araci's FinBERT (Araci, 2019)			Yang et al.'s FinBERT (Yang et al., 2020)			SEC-BERT (Loukas et al., 2022)		
	No PT	PT with SM	PT with CM (proposed)	No PT	PT with SM	PT with CM (proposed)	No PT	PT with SM	PT with CM (proposed)	No PT	PT with SM	PT with CM (proposed)
10000	84.48	**84.49**	84.29	84.46	84.46	**84.83**	84.83	**85.19**	84.63	83.93	**84.54**	84.40
20000	84.40	84.46	**84.75**	84.98	**85.00**	84.95	**85.54**	85.38	84.95	85.54	**85.62**	84.83
40000	85.53	**85.89**	85.02	84.91	85.60	**85.61**	**86.18**	85.41	85.57	85.70	**86.08**	84.55
Average	84.80	**84.95**	84.69	84.78	85.02	**85.13**	**85.52**	85.33	85.05	85.06	**85.41**	84.59
Performance improvement	+0		-0.12	+0		+0.35	+0		-0.47	+0		-0.46

5.4 Chapter Summary

In this chapter, a knowledge incorporation framework using a novel company name masking method was proposed to enhance the financial domain specificity of pretrained language models. The main contributions of this study are as follows. First, an easily implementable factual knowledge incorporation framework using Form 10-K filings as a textual knowledge base was proposed. Second, a company name masking method was proposed to incorporate company-related knowledge. Third, an evaluation framework involving stock-related spam tweet filtering was proposed to quantitatively assess the improved level of knowledge. Finally, extensive experimental results showed that the proposed company name masking method achieved superior spam filtering over the conventional subword masking method, particularly at low-resource settings with training data sizes of 600 or less.

Chapter 6

Conclusion

To date, numerous attempts have been made to assess financial risk by applying text mining techniques to financial texts. Text mining-based approaches are advantageous in that they are less expensive in terms of time, human labor, and domain expertise than manual approaches. Thus, text mining enables real-time risk assessment that requires prompt detection of rapid changes in the financial domain. Meanwhile, several studies have been conducted to develop financial domain-specific text analysis tools that can improve the performance of financial applications. However, the detection of lexical items that play a crucial role in automated financial risk assessment has received insufficient attention.

The main contributions of this dissertation are as follows. First, three text mining tasks essential for automated financial risk assessment were specified: hot topic detection, sentiment analysis, and spam filtering. Second, financial domain-specific text analysis tools that can detect hot topics, sentiment words, and spam messages, respectively, were proposed. Third, the proposed text analysis tools were validated to support early warning, explainable market sentiment analysis, and automatic spam filtering, respectively. Finally, extensive experimental results demonstrated the effectiveness of the proposed text analysis tools over the conventional general-purpose tools.

To detect hot topics, a text mining framework detecting emerging keywords in speeches made by the chairs of the Federal Reserve System was proposed. In the framework, a novel light-weight unsupervised keyword-scoring method was used, which treated bigrams as keywords and incorporated the temporal importance of keywords by estimating the growth rate of the term frequency. Application of the proposed framework to manuscripts of speeches made between 1997 and 2019 revealed that the recurrence of the terms "East Asia," "information technology," and "subprime mortgage," which describe the Asian financial crisis, dot-com bubble, and global financial crisis, respectively, could have been detected prior to the onset of the respective crises.

To detect sentiment words, an automatically constructed financial ontology-aware sentiment lexicon was proposed. The proposed lexicon reflected the financial ontology regarding direction-dependent words by estimating the degree of association between given words and their direction-dependency types. Experimental results demonstrated that the proposed lexicon outperformed existing general-purpose sentiment lexicons in solving sentiment classification tasks. Specifically, the proposed lexicon identified "acquisition," "agreement," and "communication" as "proportional" type words that create a positive/negative sentiment when they are combined with an up/down-type word.

To detect spam messages, company-related knowledge-enhanced language models were proposed. Existing pretrained language models were post-trained using the proposed knowledge incorporation framework using Form 10-K filings as a textual knowledge base. The framework used a novel company name masking method that masked tokens corresponding to company names, thus allowing the model to learn company-related information in a sentence. Extensive experimental results showed that the proposed company name masking method enriched the model's company-related factual knowledge, thereby improving spam filtering performance, particularly at low-resource settings with training data sizes of 600 or less.

In this dissertation, the effectiveness of the proposed hot topic detection framework, sentiment lexicon, and knowledge-enhanced language models was demonstrated through experiments using central bankers' speeches, economic news articles, and corporate reports, respectively. However, there is still room for improvement in that the scope of the analysis target was limited in terms of temporal range, document category, and language type.

First of all, the temporal range of the analysis was limited to a specific period in the past: speech manuscripts made between January 1997 and September 2019, economic news headlines produced before 2018, and Form 10-K filings published in 2016 were used to detect hot topics, sentiment words, and company-related information, respectively. However, to solve the current financial problems, it is necessary to integrate data generated from the past to the present and reflect the updated information in real time. Therefore, further research is needed to collect ever-increasing data and provide it to text analysis tools. Furthermore, it would be interesting to analyze the variation in hot topics, sentiment words, and company-related information over time by extending the temporal range of the analysis to include the previous 100 years.

Second, the document category of the analysis should be expanded. Further research is required to detect hot topics in other modes of central bank communications such as the statements or minutes of the Federal Open Market Committee. The Federal Open Market Committee meets eight times each year to decide on monetary policy (Cannon et al., 2015); during these meetings, the participants formulate their views on economic conditions and determine their stance on monetary policy (Cannon et al., 2015). Thus, automatic analysis of statements and minutes released at these meetings will broaden the understanding of current financial conditions and future monetary policy directions. Additionally, this would enable a comparison of the usefulness of each type of document in preemptively identifying risk factors of a financial crisis. It would be also interesting to extend the regional scope of analysis by analyzing speeches addressed by other countries' central bankers. To improve the quality of the proposed Senti-DD lexicon, it would be helpful to acquire analyst reports containing a large number of directional words and extract direction-dependent words from them. In addition, analyzing preferential trade agreements, which are crucial documents that have a huge impact on the foreign trade performance of related companies (Alschner et al., 2018; Hofmann et al., 2019), will help the language model learn trade-related knowledge.

Third, the language type addressed in this study was limited to English; however, the proposed text analysis tools can be applied to documents written in other languages such as Korean. To this end, continued efforts are required to collect finance-related documents written in Korean: speech manuscripts addressed by Korean central bankers, economic news headlines published by Korean news organizations, and corporate reports issued by Korean companies. An additional preprocessing step for Korean, which is a morphologically-rich language (Kim, 2019), is also required to analyze these documents.

Furthermore, the generalizability of the experimental results presented in this dissertation is subject to several limitations as follows. The proposed approach presented in Chapter 3 requires domain-specific stopwords, which should be carefully selected by the researchers, to filter meaningless terms when detecting hot topics. However, constructing a predefined set of stopwords is a labor-intensive and subjective process. To reduce the researchers' intervention in defining stopwords, automatically identifying domain-specific stopwords can be considered for future work. Further exploration of to what extent the past should be reflected in the scoring process can also be conducted as a future study. The proposed emergence score considers the temporal importance of keywords using the word frequency in the previous four quarters; however, it may be necessary to consider a period shorter or longer than the four quarters.

The proposed method presented in Chapter 4 classifies sentences as neutral when it contains equal numbers of positive and negative words; this is a passive approach to detecting neutrality. Valdivia et al. (2018) demonstrated that detecting neutrality first before classifying sentences as positive or negative can improve sentiment analysis performance. Thus, pre-detection of neutrality in financial documents can be considered in future work. Fine-grained sentiment analysis, which classifies sentiments into multi-classes (Van de Kauter et al., 2015) rather than binary classes, can also be explored. In general, classes can be subdivided into anxiety, sadness, anger, excitement, and happiness (Wang et al., 2020); in the case of financial documents, classes can be further segmented by identifying whether positivity/negativity is directed toward the entire market or a particular company. Meanwhile, the proposed method addressed the financial domain-specific ontology by constructing a sentiment lexicon, which constitutes a simple and intuitive format. It would be interesting to construct knowledge graphs that reflect this domain-specific ontology in future studies.

The evaluation framework proposed in Chapter 5 indirectly assesses the performance of knowledge incorporation settings by performing spam filtering as a downstream task. For future work, a dataset containing "fill-in-the-blank" statements (e.g., "As of August 2022, the company with the largest market capitalization is ______.") should be constructed. This dataset would enable a straightforward analysis of knowledge; further investigations could be conducted to categorize the types of company-related knowledge in the model by identifying whether the model is aware of a company's market capitalization or flagship products/services.

The proposed text analysis tools, which automatically detect certain types of entities in a sentence, are anticipated to be easily applied to other areas. For example, the proposed techniques can assist in evaluating environmental, social, and governance (ESG) performance of a company. Over the past few years, attempts to reflect ESG considerations into business decisions and investment strategies have increased (Raman et al., 2020). Among the three pillars, the *social* pillar refers to an organization's relationships with internal and external stakeholders (Paraschi et al., 2022). Examples of sub-dimensions of the *social* pillar include employee rights and compulsory labor (Kuo et al., 2021; Li et al., 2021). In this context, analyzing employee reviews posted on company evaluation sites such as Glassdoor (Pak, 2021) will reveal the company's activities on the *social* pillar. Filtering reviews that are too biased toward positive or negative sentiments, identifying lexical items that occur simultaneously with the terms indicating "rights" or "labor," and detecting emerging topics among those lexical items would broaden the understanding of the *social* performance of a company.

Another promising area for future research is digital marketing. Digital marketing broadly refers to the process of using digital technologies such as social media to promote brands, acquire customers, and increase sales (Kannan et al., 2017). In the digital environment, customers can post reviews on products, services, brands, and companies on websites or social media platforms; and these reviews reach a much wider audience (Kannan et al., 2017). Customers can share word-of-mouth, which refers to informal communications between private parties concerning evaluations of goods and services (Anderson, 1998; Fornell, 1992; Singh, 1988; Westbrook, 1987), with thousands of users around the world as well as a few close friends. Thus, marketers are interested in promoting positive word-of-mouth and preventing negative word-of-mouth that could damage the brand's image (Gildin, 2022).

The proposed methods—which detect hot topics, sentiment words, and spam messages for automated risk assessment—can assist preemptive detection of negative word-of-mouth spreading on social media. First, a hot topic detection framework that incorporates the temporal importance of keywords would assist in discovering emerging hashtags that customers use to share their negative experiences. Second, a sentiment lexicon that detects words co-occurring with certain types of words would assist accurate identification of sentiment words that lead to negative customer experiences. For example, in the case of online reviews of electronics such as monitors and televisions, the sentiment orientation of the term "definition" changes depending on the presence of the terms "high" and "low." The phrase "high-definition" leads to a positive sentiment, while "low-definition" leads to a negative sentiment. In this context, the proposed lexicon would accurately conduct sentiment analysis by treating the term "definition" as a "proportional"-type word that represents a positive sentiment when used with "high" and a negative sentiment when used with "low." Lastly, the language model which is post-trained using microblogs as a textual knowledge base would acquire knowledge regarding jargon that customers use when expressing their negative experiences. This model would facilitate the identification of key messages in customer reviews.

Bibliography

Acharya, V. V., & Richardson, M. (2009). Causes of the financial crisis. *Critical Review*, *21*, 195–210.

Adrian, T., & Shin, H. S. (2010). The changing nature of financial intermediation and the financial crisis of 2007–2009. *Annual Review of Economics*, *2*, 603–618.

Aliber, R. Z., & Kindleberger, C. P. (2015). *Manias, panics, and crashes: A history of financial crises*. Springer.

Alschner, W., Seiermann, J., & Skougarevskiy, D. (2018). Text of trade agreements (tota)—a structured corpus for the text-as-data analysis of preferential trade agreements. *Journal of Empirical Legal Studies*, *15*, 648–666.

Anderson, E. W. (1998). Customer satisfaction and word of mouth. *Journal of Service Research*, *1*, 5–17.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. `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 Conference on International Language Resources and Evaluation* (pp. 2200–2204).

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *131*, 1593–1636.

Bandhakavi, A., Wiratunga, N., Massie, S., & Padmanabhan, D. (2017). Lexicon generation for emotion detection from text. *IEEE Intelligent Systems*, *32*, 102–108.

Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, *60*, 1371–1391.

Bennani, H., Fanta, N., Gertler, P., & Horvath, R. (2020). Does central bank communication signal future monetary policy in a (post)-crisis era? the case of the ecb. *Journal of International Money and Finance*, *104*, 102167.

Bennani, H., & Neuenkirch, M. (2017). The (home) bias of european central bankers: new evidence based on speeches. *Applied Economics*, *49*, 1114–1131.

Bernanke, B. S. (2007a). Federal reserve board's semiannual monetary policy report to the congress. `https://www.bis.org/review/r070216a.pdf`. [Accessed: December 15, 2022].

Bernanke, B. S. (2007b). Federal reserve board's semiannual monetary policy report to the congress. `https://www.bis.org/review/r070216a.pdf`. [Accessed: December 15, 2022].

Bezemer, D. J. (2010). Understanding financial crisis through accounting models. *Accounting, Organizations and Society*, *35*, 676–688.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Beijing: O'Reilly Media, Inc. URL: `http://www.nltk.org/book`.

BIS (2019). Central bankers' speeches. `https://www.bis.org/cbspeeches/index.htm`. [Accessed: December 15, 2022].

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine learning* (pp. 113–120).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of Machine Learning Research*, *3*, 993–1022.

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance: an application to default risk analysis. Bank of England Working Paper, https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.

Brazdil, P., Silvano, M. d. P., Silva, M. d. F. H. d., Muhammad, S., Oliveira, F., Cordeiro, J., & Leal, A. (2022). Extending general sentiment lexicon to specific domains in (semi-) automatic manner. In *1st Workshop on Sentiment Analysis & Linguistic Linked Data: Proceedings of the Workshops and Tutorials held at LDK 2021 co-located with the 3rd Language, Data and Knowledge Conference (LDK 2021)*.

Briola, A., Vidal-Tomás, D., Wang, Y., & Aste, T. (2022). Anatomy of a stablecoin's failure: The terra-luna case. *Finance Research Letters*, (p. 103358).

Bun, K. K., & Ishizuka, M. (2002). Topic extraction from news archive using tf* pdf algorithm. In *Proceedings of the Third International Conference on Web Information Systems Engineering, 2002. WISE 2002.* (pp. 73–82). IEEE.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, *57*, 203–216.

Cambria, E., & Hussain, A. (2015a). Sentic computing. *Cognitive Computation*, *7*, 183–185.

Cambria, E., & Hussain, A. (2015b). Senticnet. In *Sentic Computing* (pp. 23–71). Springer.

Campbell, G., Turner, J. D., & Walker, C. B. (2012). The role of the media in a bubble. *Explorations in Economic History*, *49*, 461–481.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257–289.

Cannon, S. et al. (2015). Sentiment of the fomc: Unscripted. *Economic Review - Federal Reserve Bank of Kansas City*, *5*.

Chan, S. W., & Chong, M. W. (2017). Sentiment analysis in financial texts. *Decision Support Systems*, *94*, 53–64.

Cheng, W. K., Bea, K. T., Leow, S. M. H., Chan, J. Y.-L., Hong, Z.-W., & Chen, Y.-L. (2022). A review of sentiment, semantic and event-extraction-based approaches in stock forecasting. *Mathematics*, *10*, 2437.

Chiang, C.-H., Huang, S.-F., & Lee, H.-y. (2020). Pretrained language model embryology: The birth of ALBERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6813–6828). Online: Association for Computational Linguistics. URL: `https://aclanthology.org/2020.emnlp-main.553`. doi:`10.18653/v1/2020.emnlp-main.553`.

Choi, S., Park, H., Yeo, J., & Hwang, S.-w. (2020). Less is more: Attention supervision with counterfactuals for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6695–6704).

Chowdhry, B., & Goyal, A. (2000). Understanding the financial crisis in asia. *Pacific-Basin Finance Journal*, *8*, 135–152.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What Does BERT Look At? An Analysis of BERT's Attention. `arXiv:1906.04341`.

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. In *Association for Computational Linguistics (ACL)* (pp. 519–535).

Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. *Decision Support Systems*, *139*, 113421.

Cresci, S., Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2018). $FAKE: Evidence of spam and bot activity in stock microblogs on twitter. In *Twelfth International AAAI Conference on Web and Social Media*.

Cresci, S., Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2019). Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on twitter. *ACM Transactions on the Web (TWEB)*, *13*, 1–27.

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3504–3514.

Daudert, T. (2021). Exploiting textual and relationship information for fine-grained financial sentiment analysis. *Knowledge-Based Systems*, *230*, 107389.

De Fortuny, E. J., De Smedt, T., Martens, D., & Daelemans, W. (2012). Media coverage in times of political crisis: A text mining approach. *Expert Systems with Applications*, *39*, 11616–11622.

Dettmers, T., Minervini, P., Stenetorp, P., & Riedel, S. (2018). Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 1811–1818).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186).

Dougal, C., Engelberg, J., Garcia, D., & Parsons, C. A. (2012). Journalists and the stock market. *The Review of Financial Studies*, *25*, 639–679.

Du, C., Sun, H., Wang, J., Qi, Q., & Liao, J. (2020). Adversarial and domain-aware bert for cross-domain sentiment analysis. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 4019–4028).

Elhammadi, S., V.S. Lakshmanan, L., Ng, R., Simpson, M., Huai, B., Wang, Z., & Wang, L. (2020). A High Precision Pipeline for Financial Knowledge Graph Construction. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 967–977). International Committee on Computational Linguistics.

URL: https://aclanthology.org/2020.coling-main.84. doi:10.18653/v1/2020.coling-main.84.

Engelberg, J. E., & Parsons, C. A. (2011). The causal impact of media in financial markets. *the Journal of Finance*, *66*, 67–97.

Erçen, H. İ., Özdeşer, H., & Türsoy, T. (2022). The impact of macroeconomic sustainability on exchange rate: Hybrid machine-learning approach. *Sustainability*, *14*, 5357.

Erolu, N. et al. (2012). Monetary transmission channels and an assessment within the framework of the 2008 global financial crisis. *African Journal of Business Management*, *6*, 8554–8563.

Feng, Z., Zhou, H., Zhu, Z., & Mao, K. (2022). Tailored text augmentation for sentiment analysis. *Expert Systems with Applications*, (p. 117605).

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*, 96–104.

Fornell, C. (1992). A national customer satisfaction barometer: The swedish experience. *Journal of Marketing*, *56*, 6–21.

Foster, J. B., & Magdoff, F. (2009). *The great financial crisis: Causes and consequences*. NYU Press.

Frankel, J., & Saravelos, G. (2012). Can leading indicators assess country vulnerability? evidence from the 2008–09 global financial crisis. *Journal of International Economics*, *87*, 216–231.

Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, *34*, 443–498.

Gakhar, D. V., & Kundlia, S. (2021). Impact of sentiments on stock returns, volatility and liquidity. *International Journal of Economic Policy in Emerging Economies*, *14*, 536–565.

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

Gaytán, A., Johnson, C. A. et al. (2002). *A review of the literature on early warning systems for banking crises*. Central Bank of Chile.

Gildin, S. Z. (2022). Understanding the power of word-of-mouth. *Revista de Administração Mackenzie*, *4*, 92–106.

Greene, D., & Cross, J. P. (2017). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, *25*, 77–94.

Grootendorst, M. (2020). KeyBERT: Minimal keyword extraction with BERT. URL: `https://maartengr.github.io/KeyBERT/`. doi:`10.5281/zenodo.4461265`.

Gu, J., Kuen, J., Joty, S., Cai, J., Morariu, V., Zhao, H., & Sun, T. (2020). Self-supervised relationship probing. *Advances in Neural Information Processing Systems*, *33*, 1841–1853.

Gupta, A., Dengre, V., Kheruwala, H. A., & Shah, M. (2020). Comprehensive review of text-mining applications in finance. *Financial Innovation*, *6*, 1–25.

Gupta, V., Lehal, G. S. et al. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*, 60–76.

Hanna, A. J., Turner, J. D., & Walker, C. B. (2020). News media and investor sentiment during bull and bear markets. *The European Journal of Finance*, *26*, 1377–1395.

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing*, *4*, 1.

Helleiner, E. (2011). Understanding the 2007–2008 global financial crisis: Lessons for scholars of international political economy. *Annual Review of Political Science*, *14*, 67–87.

Hofmann, C., Osnago, A., & Ruta, M. (2019). The content of preferential trade agreements. *World Trade Review*, *18*, 365–398.

Horcher, K. A. (2011). *Essentials of financial risk management*. John Wiley & Sons.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media* (pp. 216–225).

Jaggi, M., Mandal, P., Narang, S., Naseem, U., & Khushi, M. (2021). Text mining of stocktwits data for predicting stock prices. *Applied System Innovation*, *4*, 13.

Jain, S. M. (2022). *Introduction to Transformers for NLP*. Berkeley, CA: Apress. URL: `https://link.springer.com/book/10.1007/978-1-4842-8844-3`. doi:`doi:10.1007/978-1-4842-8844-3`.

Jallan, Y., & Ashuri, B. (2020). Text mining of the securities and exchange commission financial filings of publicly traded construction firms using deep learning to identify and assess risk. *Journal of Construction Engineering and Management*, *146*, 04020137.

Johnsi, R., Kumar, G. B., & Sariki, T. P. (2022). A concise survey on datasets, tools and methods for biomedical text mining. *International Journal of Applied Engineering Research*, *17*, 200–217.

Joyce, J. (2000). The IMF and global financial crises. *Challenge*, *43*, 88–107.

Kannan, P. et al. (2017). Digital marketing: A framework, review and research agenda. *International Journal of Research in Marketing*, *34*, 22–45.

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*, 4999–5010.

Kim, H. (2019). *Unsupervised Korean Tokenizer and Extractive Document Summarization to Solve Out-of-Vocabulary and Dearth of Data*. Ph.D. dissertation, Seoul National University.

Kim, M. (2018). *Prediction of stock price, base rate, and interest rate spread with text data*. Ph.D. dissertation, Seoul National University.

Kim, S., Kim, R., Nam, H.-J., Kim, R.-G., Ko, E., Kim, H.-S., Shin, J., Cho, D., Jin, Y., Bae, S. et al. (2020). Organizing an in-class hackathon to correct pdf-to-text conversion errors of genomics & informatics 1.0. *Genomics & Informatics*, *18*.

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations*.

Ko, K., Oh, S., & Baek, J. (2020). Development of economic fluctuation topic indices and topic indices regression model for kospi200 index. *The Korean Data and Information Science Society*, *31*, 579–594.

Koyuncugil, A. S., & Ozgulbas, N. (2010). *Surveillance Technologies and Early Warning Systems: Data Mining Applications for Risk Detection: Data Mining Applications for Risk Detection*. Igi Global.

Koyuncugil, A. S., & Ozgulbas, N. (2012). Financial early warning system model and data mining application for risk detection. *Expert Systems with Applications*, *39*, 6238–6253.

Kramer, O. (2016). Scikit-learn. In *Machine learning for evolution strategies* (pp. 45–53). Springer.

Krishnamoorthy, S. (2018). Sentiment analysis of financial news articles using performance indicators. *Knowledge and Information Systems*, *56*, 373–394.

Krishnan, S., Shashidhar, N., Varol, C., & Islam, A. R. (2022). A novel text mining approach to securities and financial fraud detection of case suspects. *International Journal of Artificial Intelligence and Expert Systems*, *10*.

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, *114*, 128–147.

Kuo, T.-C., Chen, H.-M., & Meng, H.-M. (2021). Do corporate social responsibility practices improve financial performance? a case study of airline companies. *Journal of Cleaner Production*, *310*, 127380.

Kyosev, D., Anastasovski, D., Kikovic, M., & Voutyras, O. (2022). Terra luna and the future of internet investments: Towards a framework for investors' protections. doi:`10.20944/preprints202210.0027.v1`. `arXiv:10.20944`.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791.

Lee, H. (2015). Modularization of korea's development experience: early warning system for financial crisis. Available at KDI School of Public Policy and Management: `https://www.kdi.re.kr/kdi_eng/publications/publication_view.jsp?pub_no=14704`.

Li, J., Feng, Y., Li, G., & Sun, X. (2020a). Tourism companies' risk exposures on text disclosure. *Annals of tourism research*, *84*, 102986.

Li, J., Li, Y., & Xue, Z. (2020b). Keywords extraction algorithm of financial review based on dirichlet multinomial model. In *Chinese Intelligent Systems Conference* (pp. 107–116). Springer.

Li, J.-H., You, C.-F., Huang, C.-S. et al. (2020c). Do mutual fund managers time market sentiment? *International Journal of Financial Research*, *11*, 527–537.

Li, T.-T., Wang, K., Sueyoshi, T., & Wang, D. D. (2021). Esg: Research progress and future prospects. *Sustainability*, *13*, 11663.

Li, W., Paraschiv, F., & Sermpinis, G. (2022). A data-driven explainable case-based reasoning approach for financial risk detection. *Quantitative Finance*, (pp. 1–18).

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:1907.11692`.

Liu, Z., Huang, D., Huang, K., Li, Z., & Zhao, J. (2021). Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence* (pp. 4513–4519).

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*, 35–65.

Loukas, L., Fergadiotis, M., Chalkidis, I., Spyropoulou, E., Malakasiotis, P., Androutsopoulos, I., & Paliouras, G. (2022). FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1)* (pp. 4419–4431).

Luo, R., Huang, G., & Quan, X. (2021). Bi-granularity contrastive learning for post-training in few-shot scene. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP* (pp. 1733–1742).

Macêdo, J. B., das Chagas Moura, M., Aichele, D., & Lins, I. D. (2022). Identification of risk features using text mining and bert-based models: Application to an oil refinery. *Process Safety and Environmental Protection*, *158*, 382–399.

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* (pp. 1941–1942).

Malandri, L., Xing, F. Z., Orsenigo, C., Vercellis, C., & Cambria, E. (2018). Public mood–driven asset allocation: The importance of financial sentiment in portfolio management. *Cognitive Computation*, *10*, 1167–1176.

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*, 782–796.

Man, X., Luo, T., & Lin, J. (2019). Financial sentiment analysis (fsa): A survey. In *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)* (pp. 617–622). IEEE.

Masawi, B., Bhattacharya, S., & Boulter, T. (2018). Does the information content of central bank speeches impact on the level of exchange rate? A comparative study of Canadian and Australian Central Bank communications. *Review of Pacific Basin Financial Markets and Policies*, *21*, 1850005.

Mashrur, A., Luo, W., Zaidi, N. A., & Robles-Kelly, A. (2020). Machine learning for financial risk management: A survey. *IEEE Access*, *8*, 203203–203223.

Maurya, P., Singh, A., & Salim, M. (2022). The application of text mining in detecting financial fraud: A literature review. *Business Intelligence and Human Resource Management*, (pp. 243–260).

McDonnell, A., & Burgess, J. (2013). The impact of the global financial crisis on managing employees. *International Journal of Manpower*, *34*, 184–197.

Mehrafarin, H., Rajaee, S., & Pilehvar, M. T. (2022). On the importance of data size in probing fine-tuned models. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 228–238). Dublin, Ireland: Association for Computational Linguistics. URL: `https://aclanthology.org/2022.findings-acl.20`. doi:`10.18653/v1/2022.findings-acl.20`.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 404–411).

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (p. 3111–3119). volume 26.

Min, H., Caltagirone, J., & Serpico, A. (2008). Life after a dot-com bubble. *International Journal of Information Technology and Management*, *7*, 21–35.

Mohammad, S., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the Seventh International Workshop on Semantic Evaluation* (pp. 321–327).

Montariol, S., Allauzen, A., & Kitamoto, A. (2020). Variations in word usage for the financial domain. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing* (pp. 8–14).

Moreno-Ortiz, A., Fernández-Cruz, J., & Hernández, C. P. C. (2020). Design and evaluation of SentiEcon: A fine-grained economic/financial sentiment lexicon from a corpus of business news. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 5065–5072).

Nielsen, F. Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. `arXiv:1103.2903`.

Nyman, R., Kapadia, S., & Tuckett, D. (2021). News and narratives in financial systems: exploiting big data for systemic risk assessment. *Journal of Economic Dynamics and Control*, *127*, 104119.

Oliveira, N., Cortez, P., & Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, *85*, 62–73.

Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, *73*, 125–144.

Orabi, M., Mouheb, D., Al Aghbari, Z., & Kamel, I. (2020). Detection of bots in social media: a systematic review. *Information Processing & Management*, *57*, 102250.

Ostendorff, M., Ash, E., Ruas, T., Gipp, B., Moreno-Schneider, J., & Rehm, G. (2021). Evaluating document representations for content-based legal literature recommendations. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (pp. 109–118).

Ott, C. (2020). The risks of mergers and acquisitions—analyzing the incentives for risk reporting in item 1a of 10-k filings. *Journal of Business Research*, *106*, 158–181.

Pak, E. S. (2021). *Mining Intangible Internal Resources from Employee Voice with Deep Learning*. Master's thesis, Seoul National University.

Paraschi, E. P. et al. (2022). Why esg reporting is particularly important for the airlines during the covid-19 pandemic. *Journal of Business and Management Studies*, *4*, 63–67.

Parikh, R., & Karlapalem, K. (2013). ET: events from tweets. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 613–620).

Patel, V. M., Gopalan, R., Li, R., & Chellappa, R. (2015). Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, *32*, 53–69.

Pavlov, A., & Wachter, S. (2011). Subprime lending and real estate prices. *Real Estate Economics*, *39*, 1–17.

Peng, Y., Kou, G., & Shi, Y. (2009). Knowledge-rich data mining in financial risk detection. In *International Conference on Computational Science* (pp. 534–542). Springer.

Peters, M. E., Neumann, M., Logan, R., Schwartz, R., Joshi, V., Singh, S., & Smith, N. A. (2019). Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 43–54). Hong Kong, China: Association for Computational Linguistics.

Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I., & Shrimpton, L. (2013). Can Twitter Replace Newswire for Breaking News? In *Seventh International AAAI Conference on Weblogs and Social Media*.

Picasso, A., Merello, S., Ma, Y., Oneto, L., & Cambria, E. (2019). Technical analysis and sentiment embeddings for market trend prediction. *Expert Systems with Applications*, *135*, 60–70.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*, 1872–1897.

Raman, N., Bang, G., & Nourbakhsh, A. (2020). Mapping esg trends by distant supervision of neural language models. *Machine Learning and Knowledge Extraction*, *2*, 453–468.

Razova, E., Vychegzhanin, S., & Kotelnikov, E. (2022). Does BERT look at sentiment lexicon? In *International Conference on Analysis of Images, Social Networks and Texts* (pp. 55–67). Springer.

Rehurek, R., & Sojka, P. (2011). Gensim—statistical semantics in python. `https://radimrehurek.com/gensim/`. [Accessed: December 15, 2022].

Ruiz-Martínez, J. M., Valencia-García, R., García-Sánchez, F. et al. (2012). Semantic-based sentiment analysis in financial news. In *Proceedings of the 1st International Workshop on Finance and Economics on the Semantic Web* (pp. 38–51).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*, 513–523.

Sheetz, M. (2019). Boeing reports $2.9 billion quarterly loss—its worst ever—after taking 737 Max charge. `https://www.cnbc.com/2019/07/24/boeing-earnings-q2-2019.html`. [Accessed: December 15, 2022].

Shiller, R. J. (2016). *Irrational Exuberance*. Princeton: Princeton University Press. URL: `https://doi.org/10.1515/9781400865536`. doi:`doi:10.1515/9781400865536`.

Shinyama, Y., Guglielmetti, P., & Marsman, P. (2019). PDFMiner.six. `https://github.com/pdfminer/pdfminer.six`. [Accessed: December 15, 2022].

Singh, J. (1988). Consumer complaint intentions and behavior: definitional and taxonomical issues. *Journal of Marketing*, *52*, 93–107.

Song, M., & Shin, K.-s. (2019). Forecasting economic indicators using a consumer sentiment index: Survey-based versus text-based data. *Journal of Forecasting*, *38*, 504–518.

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, *7*, 484.

Sun, Z., Deng, Z.-H., Nie, J.-Y., & Tang, J. (2019). RotatE: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*.

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

Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text mining: techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, *7*.

Tardelli, S., Avvenuti, M., Tesconi, M., & Cresci, S. (2020). Characterizing social bots spreading financial disinformation. In *International Conference on Human Computer Interaction* (pp. 376–392). Springer.

Tardelli, S., Avvenuti, M., Tesconi, M., & Cresci, S. (2022). Detecting inorganic financial campaigns on twitter. *Information Systems*, *103*, 101769.

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

Thakor, A. V. (2015). The financial crisis of 2007–2009: Why did it happen and what did we learn? *The Review of Corporate Finance Studies*, *4*, 155–205.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology*, *61*, 2544–2558.

Thorbecke, W. (2002). A dual mandate for the federal reserve: The pursuit of price stability and full employment. *Eastern Economic Journal*, *28*, 255–268.

Tiddi, I., & Schlobach, S. (2022). Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, *302*, 103627.

Valdivia, A., Luzón, M. V., Cambria, E., & Herrera, F. (2018). Consensus vote models for detecting and filtering neutrality in sentiment analysis. *Information Fusion*, *44*, 126–135.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vidanagama, D., Silva, A., & Karunananda, A. (2022). Ontology based sentiment analysis for fake review detection. *Expert Systems with Applications*, *206*, 117869.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, *57*, 78–85.

Wang, W., Xu, Y., Du, C., Chen, Y., Wang, Y., & Wen, H. (2021a). Data set and evaluation of automated construction of financial knowledge graph. *Data Intelligence*, *3*, 418–443.

Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., & Tang, J. (2021b). Kepler:

A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, *9*, 176–194.

Wang, X., & McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 424–433).

Wang, Z., Ho, S.-B., & Cambria, E. (2020). Multi-level fine-scaled sentiment sensing with ambivalence handling. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *28*, 683–697.

Wei, L., Deng, Y., Huang, J., Han, C., & Jing, Z. (2022). Identification and analysis of financial technology risk factors based on textual risk disclosures. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*, 590–612.

Wei, L., Li, G., Zhu, X., Sun, X., & Li, J. (2019). Developing a hierarchical system for energy corporate risk factors based on textual risk disclosures. *Energy Economics*, *80*, 452–460.

Westbrook, R. A. (1987). Product/consumption-based affective responses and postpurchase processes. *Journal of Marketing Research*, *24*, 258–270.

Whang, T., Lee, D., Lee, C., Yang, K., Oh, D., & Lim, H. (2020). An effective domain adaptive post-training method for bert in response selection. In *Proc. Interspeech 2020*.

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

Wu, S., Wu, F., Chang, Y., Wu, C., & Huang, Y. (2019). Automatic construction of target-specific sentiment lexicon. *Expert Systems with Applications*, *116*, 285–298.

Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Intelligent asset allocation via market sentiment views. *IEEE Computational Intelligence Magazine*, *13*, 25–34.

Xing, F. Z., Cambria, E., & Zhang, Y. (2019). Sentiment-aware volatility forecasting. *Knowledge-Based Systems*, *176*, 68–76.

Xiong, W., Du, J., Wang, W. Y., & Stoyanov, V. (2020). Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *International Conference on Learning Representations*.

Xu, H., Liu, B., Shu, L., & Yu, P. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (pp. 2324–2335). Minneapolis, Minnesota: Association for Computational Linguistics.

Yamada, I., Asai, A., Shindo, H., Takeda, H., & Matsumoto, Y. (2020). LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yang, L. (2003). The Asian financial crisis and non-performing loans: evidence from commercial banks in Taiwan. *International Journal of Management*, *20*, 69.

Yang, Y., UY, M. C. S., & Huang, A. (2020). FinBERT: A Pretrained Language Model for Financial Communications. `arXiv:2006.08097`.

Yekrangi, M., & Abdolvand, N. (2021). Financial markets sentiment analysis: Developing a specialized lexicon. *Journal of Intelligent Information Systems*, *57*, 127–146.

Yilmaz, S., & Toklu, S. (2020). A deep learning analysis on question classification task using Word2vec representations. *Neural Computing and Applications*, *32*, 2909–2928.

Yu, L.-C., Wu, J.-L., Chang, P.-C., & Chu, H.-S. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems*, *41*, 89–97.

Zanzotto, F. M. (2019). Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, *64*, 243–252.

Zhang, M., Yang, J., Wan, M., Zhang, X., & Zhou, J. (2022a). Predicting long-term stock movements with fused textual features of chinese research reports. *Expert Systems with Applications*, *210*, 118312.

Zhang, X., Du, Q., & Zhang, Z. (2022b). A theory-driven machine learning system for financial disinformation detection. *Production and Operations Management*, *31*, 3160–3179.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1441–1451). Florence, Italy: Association for Computational Linguistics.

Zhao, L., Li, L., Zheng, X., & Zhang, J. (2021). A BERT based sentiment analysis and key entity detection approach for online financial texts. In *2021 IEEE 24th Inter-*

national Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp. 1233–1238). IEEE.

Zhu, X., Wang, Y., & Li, J. (2022). What drives reputational risk? Evidence from textual risk disclosures in financial statements. *Humanities and Social Sciences Communications*, *9*, 1–15.

Zhu, X., Yang, S. Y., & Moazeni, S. (2016). Firm risk identification through topic analysis of textual financial disclosures. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–8). IEEE.

국문초록

텍스트마이닝은 텍스트 데이터로부터 유용한 정보를 추출하는 과정을 일컫는 개념이다. 이는 특정 목적을 위해 설계된 프레임워크, 감성사전, 사전학습 언어모델 등의 텍스트 분석 도구를 활용하여 핫토픽 탐지, 감성 분석, 스팸 필터링 등의 과업을 수행하는 과정을 포괄한다. 텍스트마이닝이 그 광범위한 적용 가능성 덕분에 정치, 경제, 사회 등 다양한 도메인의 의사결정 과정을 돕는 데에 활용되어 온 한편, 많은 연구자들이 금융 텍스트에 텍스트마이닝을 적용하여 금융위험을 탐지하기 위한 시도를 수행해왔다. 텍스트마이닝 기반 방법은 수작업 기반 방법보다 시간, 노동력, 전문 지식 측면에서 비용 효율적인 방법으로서, 금융 도메인의 급변하는 양상을 신속하게 감지하는 실시간 위험 평가를 가능케 한다. 금융 텍스트마이닝과 관련한 대부분의 기존 연구들은 범도메인 텍스트 분석 도구를 금융 텍스트에 적용해왔다. 그러나 금융 텍스트는 일반 도메인의 텍스트와 구별되는 몇 가지 특징을 갖고 있다. 몇몇 연구자들이 텍스트 분석 도구에 금융 텍스트의 도메인 특수성을 반영하려는 시도를 수행해 왔지만, 자동화된 금융위험 평가에 중요한 역할을 하는 어휘 항목을 탐지하는 것에 대해서는 아직 충분한 논의가 이루어지지 않았다.

본 논문에서는 핫토픽, 감성어 및 스팸 메시지를 탐지할 수 있는 금융 도메인 특화 텍스트 분석 도구를 제안한다. 제안된 도구들은 각각 금융위기 조기경보, 뉴스 기사 대상 설명가능한 시장심리 추정, 실시간 발생하는 데이터 대상 주식 관련 스팸 필터링을 지원함으로써 금융위험 평가 자동화에 기여할 것으로 기대된다. 첫째, 키워드의 시간적 중요도를 반영하는 핫토픽 탐지 프레임워크를 제안한다. 이 프레임워크는 중앙은행 총재의 연설문에 적용되어 텍스트마이닝 기반 조기경보시스템에의 가능성을 시사한다. 둘째, 어떤 단어의 감성이 그 주변에서 등장하는 방향성 단어의 존재 유무에 따라 변화할 수 있다는 금융 도메인 온톨로지를 반영하여 자동으로 구축된 감성사전을 제안한다. 경제 뉴스 헤드라인 벤치마크 데이터셋에 이

감성사전을 적용하여 제안하는 감성사전이 시장심리 추정 과정에 대한 설명력을 갖추었음을 입증한다. 셋째, 비우량주를 우량주인 것처럼 홍보하는 스팸 메시지를 탐지하기 위해 기업 관련 지식을 강화한 언어 모델을 제안한다. 구체적으로, 모델에 사실적 지식을 주입하기 위해 기업 보고서를 지식 기반으로 사용하는 지식 통합 프레임워크를 제안한다. 이 프레임워크에서 사용하는 마스킹 방법은 기업명에 해당하는 토큰을 마스킹함으로써 모델이 어떤 문장에 표현된 기업 관련 사실적 정보를 학습하게 한다. 제안하는 프레임워크를 통해 학습된 언어 모델의 스팸 필터링 성능은 트위터 벤치마크 데이터셋을 대상으로 검증되어 이러한 자동 스팸 필터링이 자동화 시스템에 데이터를 실시간으로 공급하는 데에 기여할 수 있다는 가능성을 입증한다.

주요어: 텍스트마이닝, 금융 텍스트, 금융위험 평가, 자동화 시스템, 핫토픽 탐지, 감성 분석, 스팸 필터링, 중앙은행 총재 연설문, 경제 뉴스 기사, 기업 보고서, 트위터

학번: 2018-28881