공학석사학위논문

Automatic Classification of News on CEO Turnover Cause for post-turnover Volatility prediction

대표이사 변경 이후 주가 변동성 예측을 위한 뉴스의 대표이사 변경 사유 분류 자동화

2023 년 2 월

서울대학교 대학원

산업공학과

함 영 석

Automatic Classification of News on CEO Turnover Cause for post-turnover Volatility prediction

대표이사 변경 이후 주가 변동성 예측을 위한 뉴스의 대표이사 변경 사유 분류 자동화

지도교수 조 성 준

이 논문을 공학석사 학위논문으로 제출함

2022 년 12 월

서울대학교 대학원

산업공학과

함 영 석

함영석의 공학석사 학위논문을 인준함

2022 년 12 월

위 원 장 이 재 욱 (인)

부위원장 조 성 준 (인)

위 원 장 우 진 (인)

Abstract

Automatic Classification of News on CEO Turnover Cause for post-turnover Volatility prediction

Youngseok Hahm
Department of Industrial Engineering
The Graduate School
Seoul National University

A CEO turnover event is an event significantly influencing the company. The role of CEO at a firm is to manage overall operations, and thus a change in CEO could affect not only the firm's strategic direction but also consumer perception, investment decision and eventually the share price. Thus, shareholders and investors keep an eye on the change of CEO, especially on the reason why the CEO has changed. CEO turnover causes can be inferred from the detailed information about the firm such as the firm performance and stock price prior to the event. However, in financial news related to CEO turnover specifically describe the motivation of the turnover. In this paper, a machine learning thecniques such as the TF-IDF method and the fine-tuned DistilBERT language model were utilized to classify the turnover causes from financial news related to CEO turnover. The main contribution of this paper is to automate the manual labeling process to aid shareholders and investors to capture the investment opportunity in a timely manner. A contextualized embedding

of news articles obtained from the language model is then further utilized as an additional feature for predicting the post-event stock volatility of a firm.

Keywords: Text Classification, Natural Language Processing, Stock Volatility, TF-IDF, DistilBERT

Student Number: 2020-23018

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

On top of the complex organizational structures of companies, there exists a group of individuals called "C-Suite" in charge of ensuring the companies' business plans. Common C-suite executives include Chief Executive Officer (CEO), Chief Financial Officer (CFO) and Chief Technology Officer (CTO). Each position's responsibility varies by one's area of expertise and each plays critical roles in a company. In particular, a CEO is the top-echelon executive who gets reported by other C-suite members and generally has greater responsibility for managing the overall operations and resources.

A CEO is an individual employee of a company who is responsible for the ultimate success of failure of the company. Although the role of CEOs varies by the size or organizational structure of companies, CEOs decisions ultimately determine the strategic direction of the firm and thus have influence on firm performance. The relationship between CEOs and firm performance has been examined by business scholars and practitioners for centuries [2, 4, 7, 9, 10, 25]. CEOs are expected to develop new strategic objectives and direction with the purpose of boosting firm performance, to engage in public relations for marketing purposes as being the face

of the company and to communicate with employees to ensure well-established work cultures.

1.2 Problem Description

Despite the fact that every CEO has an objective of ensuring the profitability and development of the company, the outcomes based on complex factors are not always as expected. Not only the decisions made by CEOs but also personal and social behaviors by CEOs are reflected in the firm performance and share prices [2, 7, 9]. For example, Elon Musk, the CEO of Tesla, smoking marijuana during a YouTube talk show resulted in a 7% drop in Tesla's share price. Such CEO-derived events have both short-term and long-term effects on the share price either positively or negatively. In extreme cases, the Board of Directors (BOD) removes potential threats to the company by terminating the incumbent CEO.

A CEO turnover, which is a change in executive leadership, is a significant event in the life of a company. The most apparent turnover causes would be the poor performance of the incumbent CEO. Quantitative measures of firm performance such as Cumulative Abnormal Returns (CARs) are used to evaluate the incumbent CEO's performance [22]. Besides the quantitative measures, sociopolitical factors are also considered as turnover motivating reasons, such as when the incumbent CEO's management skills do not meet the BOD's expectations or when the strategic objectives between the incumbent CEO and the BOD could not reach a consensus, the BOD is motivated to find a successor [8]. Along with these involuntary forms of turnover causes, there exist voluntary types such as retirement. Several studies have investigated the impact of CEO turnover on equity volatility using regression

analysis using turnover causes as predictor variables as well as other financial factors such as the size of firm and the performance level estimated with CARs [22, 8].

Volatility is a statistical measure of dispersion of return of a security for a given period of time. A higher volatility means there would be a large fluctuation in the price of security over a given time period and a lower volatility means the price of the security would not fluctuate dramatically. Historical Volatility v_t on day t has been computed using the standard deviation of the historical close price of the security given n as the amount of days to look back for rolling window method and P_t as the adjusted close price on day t as in (1.1, 1.2) . The computed historical volatility of the security explains the magnitude of fluctuation of closing prices as illustrated in the Figure 1.1. Shareholders and investors pay attention to the volatility of the security because there exist several investment strategies related to volatility especially related to Options Trading. For example, a strategy called Long Straddle could profit when the stock price moves largely in either direction. Therefore, the prediction of volatility in the stock market has therefore been an interesting area of research for both scholars and shareholders [28]. Based on the result of studies on the relationship between CEO turnover causes and volatility, this paper proposes an automated method for classification of CEO turnover causes using Natural Language Processing (NLP) techniques in order to provide valuable strategic guidance for shareholders and investors.

$$x_t = ln(\frac{P_t}{P_{t-1}}) \tag{1.1}$$

$$v_t = \sqrt{\frac{\Sigma_{t=1}^{n}(x_t - \overline{X})}{n}} \tag{1.2}$$

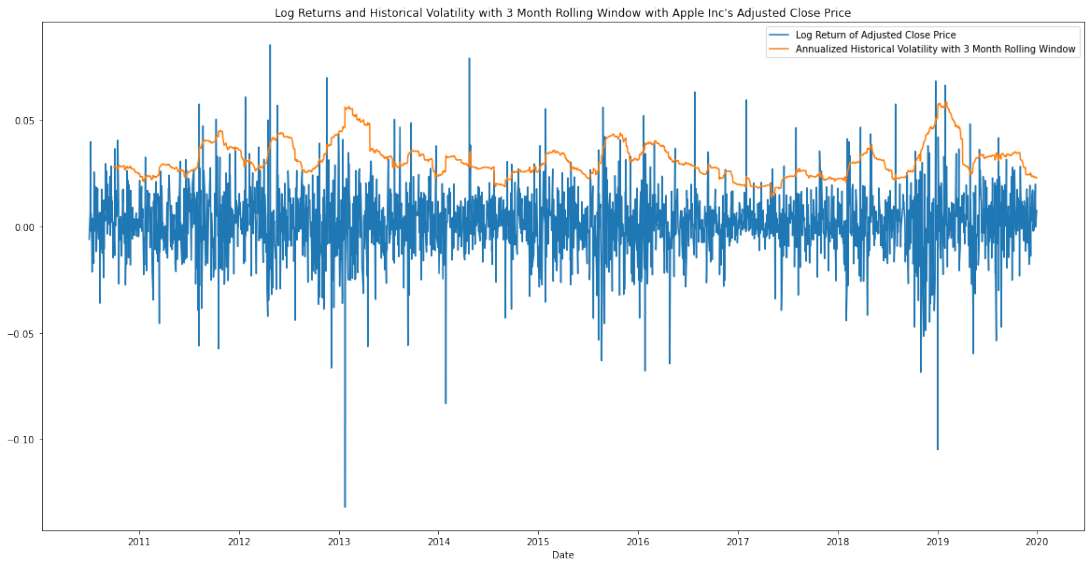


Figure 1.1: A change in volatility and log returns of Apple Inc

1.3 Research Motivation and Contribution

An estimated number of companies worldwide has continuously increased and exceeded 300 million in 2021 as shown in Figure 1.3 [32]. There could be a co-CEO, or CEO Jr., and hence multiple CEOs in a company. Assuming there exists at least one CEO in a company, there are more than 300 million CEOs in a company. According to the reports made by Challenger, Gray & Christmas, Inc., there are at least 1,000 cases of CEO turnover every year on average in the U.S.-based companies as shown

in Figure 1 1.2 [6]. Previous studies manually labeled the samples of historical CEO turnover events of size less than 1,000 based on articles related to the event while the dataset used in this experiment contains at leas 5,000 turnover events for firms in S&P 1500 only. This study is motivated by the fact that the manual classification of turnover events is inextensible and inefficient as the number of CEOs and the turnover rates have increased [6]. The key contribution of this paper is to apply NLP with machine learning on news articles related to CEO turnover to extract text embeddings to classify the turnover causes automatically. By validating and utilizing the result of previous studies on the relationship of turnover causes and volatility, the model could capture the investment opportunity for investors in the stock market especially in Options trading.

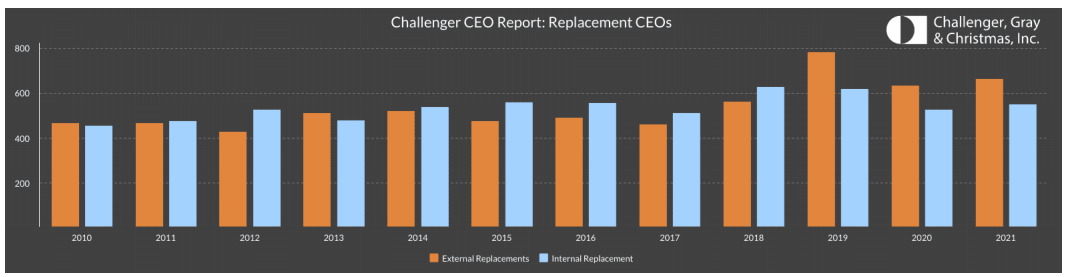


Figure 1.2: The number of CEO replacement cases from 2010 to 2021

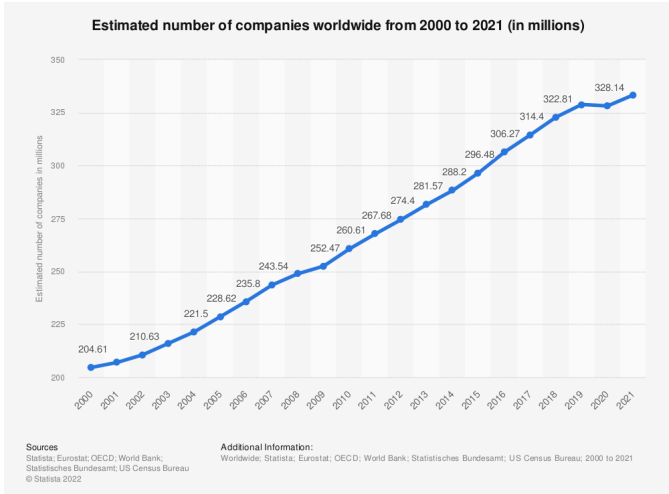


Figure 1.3: The number of companies wordwide

1.4 Organization of the Thesis

The remainder of this paper is structured as follows: Section 2 outlines the related literature. Section 3 introduces the proposed methodology. Section 4 describes the dataset used in this paper and includes the result of the experiments conducted in this paper. Finally, Section 5 includes discussions, limitations and possible future works.

Chapter 2

Literature Review

2.1 CEO Turnover and Volatility

A change in CEO attracts the attention of shareholders and investors as it is a significant event that influences organizational performance and share price ultimately. There is uncertainty in the direction of share price movement caused by CEO turnover. Thus, several studies investigated the impact of CEO turnover on volatility instead of share price [10, 16, 22, 8]. Defond and Park (1999) and Hazarika et al. (2011) showed that the CEO turnovers are likely to take place when prior volatility is high and vice versa. Clayton et al. (2005) and Li et al. (2021) explored the impact of CEO turnover on the post-event stock volatility.

Clayton et al. (2005) established three hypotheses based on two turnover causes. The first factor was whether the turnover was forced or not, and the second factor was whether the successor is from inside or from outside. They tested their hypotheses using a sample of 872 CEO turnover cases and showed that even voluntary turnover followed by successors from inside caused an increase in volatility. As a result, they showed that volatility increased up to 24% when turnover was forced and the successor was from outside the company. Their experiment also added control variables explaining the pre-turnover firm characteristics such as industry median Q

and Net Operating Income in the regression model. Hence, their experiment showed that turnover type explains the changes in volatility even in the presence of those additional causes. Clayton et al. (2005) focused on whether the turnover was forced or not and whether the succession was from inside or outside. In addition to these two types of turnover, Li et al. (2021) estimated the performance of CEO by calculating the industry adjusted CARs in three years prior to the event and used it as the third factor. At the univariate analysis level, the experiment showed that the volatility increases for all turnover types if it was forced. Both studies performed regression analysis on the to analyze volatility changes around turnover events and showed that forced turnovers result in larger volatility increases than voluntary turnovers.

2.2 Machine Learning for Text Classification

Before the emergence of the pre-trained language model, text classification was already an intriguing research area for scholars in various fields. Text classification is a process of automatically assigning categories to a sequence of texts based on its content [17]. A lot of data created and collected in businesses are in unstructured data such as text, and it requires a lot of resources to extract meaningful knowledge from that data. Depending on the context of sample data or categories labeled with data, the text classification has various applications. It could be used in spam filtering [31], opinion mining from online customer reviews and sentiment analysis [20].

In order to perform analysis on textual data, it is necessary to convert textual data into vectors containing numerical values. Prior to the usage of pretrained language models, researchers developed several methods to process textual data. The

traditional approach on handling textual data was based on statistical measurements and called the Statistical Language Model. The main objective of the language model was to estimate the probability distribution of words, phrases or sentences. To estimate the probability distribution of words, the frequency of words in the data corpus is an obvious measurement. Hence, count-based approaches were traditionally utilized in language models through methods such as TF-IDF [1], Bag-of-Words (BoW) [38] or Word2Vec [26] The BoW model simply records the occurrence of words within a document ignoring the order or structure. On the contrary, TF-IDF measures the occurrence of words at document level to put more weights on words which appear in fewer documents compared to words that appear in many documents. For example, if a text corpus contains news articles related to CEO turnover, most articles would contain the word "CEO". However, the word describing the turnover causes such as "Death", "Illness" would appear less. In this case, BoW put more weights on the word "CEO" than "Death" or "Illness" which is not a desired behavior of the model. After textual data successfully transformed into vector representation, various classification models such as K-Nearest-Neighbor algorithms [33], Naive Bayes [37], SVM [18] can be applied to perform text classification.

2.3 Pretrained Language Model for Text Classification

The introduction of transformer architecture utilizing the attention mechanism was a groundbreaking event in the NLP research area [35]. A transformer architecture has solved long-term dependency problems that the traditional seq2seq models have suffered from and significantly shortened the training time of the models. Additionally, transformer architecture enables the model to behave more like a human. With

the development of transformer architecture, language models like ELMo, BERT were introduced. BERT, which stands for Bidirectional Encoder Representations from Transformers, is a language model published in 2018 that is pre-trained on a large corpus and achieved state-of-the-art performance when it was published. BERT was utilized in many NLP downstream tasks successfully. Many researchers have improved BERT and a lot of variants of BERT such as DistilBERT was introduced.

DistilBERT is a distilled version of BERT [8] which has a size of 40% of the size of BERT developed by Hugging Face in 2019 [30]. It retains 97% of BERT performance on the sets of GLUE benchmarks. The distillation process consists of training a model based on a larger model, called the teacher to teach the distilled model, called the student, to produce the behavior of the teacher [30]. BERT is known for the use of the transformer architecture to produce sentence encodings and has been widely used in various NLP related downstream tasks. As a Pre-trained Language Model (PLM), it is capable of learning the contextual word embeddings and converting the textual data into numerical vector representations. The learned vector representation is then used for fine-tuning for desired downstream tasks such as text classification, summarization and generation.

DistilBERT has been widely used in various fields for text classification tasks. It has been used in classifying legal documents [3], sentiment analysis on banking financial news [12] and in distinguishing between fake news and satire [24]. DistilBERT achieved better performance in text classification than the traditional machine learning approach with relatively low resources due to its distilled training procedure. On top of the improvement with respect to performance of the model, DistilBERT along

with other pre-trained language models is capable of extracting contextualized embeddings of the input text that can be utilized for further experiments.

Chapter 3

Proposed Method

3.1 Overall Architecture

The proposed method's architecture is composed of a series of text classification models and a regression model for utilizing the contextual embeddings obtained from the language model used in classification task. The overall framework is illustrated in Figure 3.1. The objective of this paper is to aid shareholders and investors to capture the investment opportunity when CEO turnover occurs. Hence, the expected scenario is as follows.

1. A financial news article is given as an input to the proposed model.
2. The model determines if the financial news article is related to CEO turnover.
3. If the input article is classified as CEO turnover related, the model then identifies the turnover cause from the same input and obtains the contextualized embedding. Otherwise, the model stops processing.
4. The contextualized embedding is utilized as features for predicting stock volatility.

The proposed model is capable of classifying the turnover causes into predefined

categories. There are two sets of categories where the first set has two categories which are voluntary and involuntary and the second set has more fine-grained categories as described in Figure 4.2. A post-turnover volatility change is analyzed and forecasted so that shareholders and investors can adjust their strategic decisions accordingly.

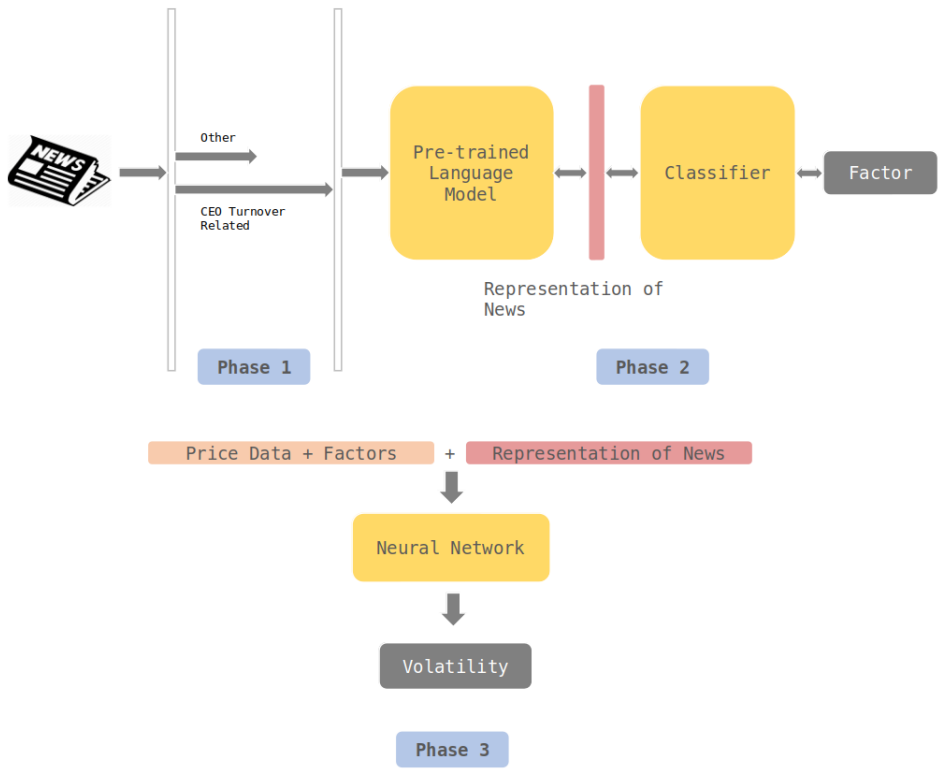


Figure 3.1: The overall framework proposed by the paper

3.2 Machine Learning Text Classification

Although there have been several studies on the relationship between stock volatility and CEO turnovers, most studies were done with relatively small dataset. Also, only

statistical approaches were attempted with turnover events by marking causes manually. In this paper, all available news data related to CEO turnovers were utilized to apply both machine learning and deep learning techniques to fully automate the labelling process. Since the main focus of the proposed method is to identify the turnover causes from news articles, both machine learning and deep learning techniques are used and compared in the accuracy and precision score of the models. In addition, the machine learning techniques used as a benchmark model for comparing the performance of the deep learning methods.

Text classification task is composed of multiple stages as described in Figure 3.2. In general, text data is collected and then preprocessed to ensure text data contains meaningful information and to provide more machine-readable data to the model. After the preprocessing, text data must be transformed into vectors containing numerical values through feature extraction and selection. Finally, the vector representation of text data is then fed into various classifiers to fit the model. Preprocessing text data is required and the most difficult task in the process of natural language processing because there are no specific guidelines or state-of-the-art algorithms for it. Depends on the contents of the text data, different strategies should be considered for the preprocessing step.

Common preprocessing techniques are stemming, lemmatization, tokenization and removal of stopwords [36]. Sentences are composed of multiple words. In order to represent text into vector space, sentences must be split into words through the tokenization process. During the process of splitting the sentence into words, stopwords such as “a”, “the”, “he” are removed since these words do not help in distinguishing two sentences. For a vast amount of text corpus, even after removing

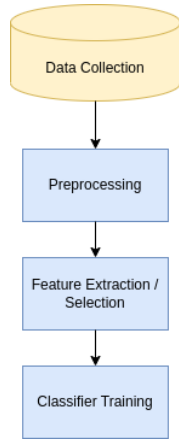


Figure 3.2: General Pipeline for Text Classification Task

stopwords in the sentence, it requires high dimensional vector space to represent all text in the corpus. Stemming is a process of transforming a word to its root form such as transforming the words “changing”, “changed” and “change” into “chang” while Lemmatization transforms the words into a word existing in the language such as transforming the words “changing”, “changed” into “change” without chopping the suffix “e” off as stemming does as illustrated in Figure 3.3.

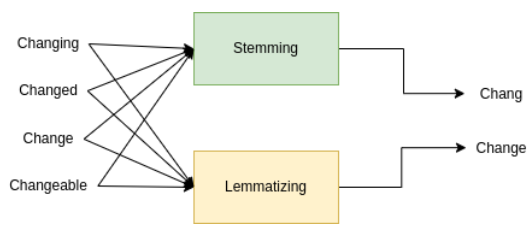


Figure 3.3: Example of Text Preprocessing using Stemming and Lemmatization

After the proper preprocessing, preprocessed text should be converted into numerical vector representation to be used as input for various classifiers. Common text feature extraction methods for machine learning are TF-IDF, BoW or Word2Vec as described in section 2. In this paper, stopwords were removed and PorterStemmer

and WordNetLemmatizer from NLTK library package [5] were used to preprocess the text data and TF-IDF method was utilized to extract features. In addition to the general preprocessing stages, the name of executives and the name of companies have been removed from the news articles. Since the size of the dataset is relatively small, when company names and executive names are included in the corpus, those words tend to be selected as the important feature words. After removal of the names of companies and executives, the TF-IDF method could capture the reasonable features for each class as shown in Table 3.1.

Table 3.1: Top 5 Important Features Extracted from TF-IDF Method

Turnover Cause	Important Features				
Death	died	officer died	died unexpectedly	ceo died	unexpectedly
Illness	leave absence	absence	jbt	santi	shuffle master
Job Performance	loss	stock	poor	changed	yahoo
Legal Violations or Concerns	investigation	allegation	sexual	consensual	false
Retired	resigned	loss	stock	fired	changed
New Opportunity	accept position	accept	riddle	greg	stationer

The last stage of text classification in machine learning is to feed the vector representation of text into a classifier. There exist various classifiers such as Random-Forest classifier, Support Vector Machine (SVM) classifier, Naive Bayes classifiers. In this experiment, two classifiers, SVM and Multinomial Naive Bayes classifiers, are used and compared since SVM and Naive Bayes methods are robust against overfitting problems that can be caused by the smaller size of the dataset and they are more suitable for the sparse data which is what TF-IDF method usually outputs. For each model, the experiment started with the default parameters and performed hyperparameter tuning if the performance can improve.

3.3 Fine-Tuning DistilBERT for Text Classification

Pretrained Language Models have achieved state-of-the-art performance on most downstream NLP tasks. It has been widely accepted that the increase in number of parameters in the model led to the increase in the performance. The base model of BERT is known to have approximately 110 million parameters with 12 transformer encoder blocks, 12 attention heads and 768 hidden dimension size while the large model of BERT has 340 million parameters with 24 transformer layers, 16 attention heads and 1024 hidden dimension size. Researchers have attempted to modify the architecture of BERT and applied various techniques to improve the performance of BERT and came up with variants of BERT such as ALBERT [21], SBERT [29] and DistilBERT [30].

DistilBERT, a distilled version of BERT, originated from BERT which shares the exact same architecture with BERT with fewer transformer encoder blocks. Unlike BERT pre-trained on Masked Language Model (MLM) and Next Sentence Prediction

(NSP), DistilBERT is only pre-trained using MLM but using three loss functions. From the combination of three loss functions, the training process of DistilBERT mimics a student-teacher learning with BERT as a teacher model. As a result, the DistilBERT model retains 97% of BERT functionality with 40% smaller size and 60% faster training speed. In this paper, the base architecture of DistilBERT is utilized for the classification task due to its efficiency with respect to hardware resources.

In the process of converting the sequence of embeddings into one sentence embedding, called pooling, it is important to capture the meaning of the entire text. Most common method is to use a special token, called CLS token, inserted in front of the tokenized input. Since BERT is trained on Next Sentence Prediction (NSP) tasks, the CLS token embedding is fine-tuned on sentence-level tasks and thus could be used for downstream tasks such as classification. As features extracted from TF-IDF described, each news article for turnover causes contains a distinct set of words. Hence, the experiment utilized three different strategies for pooling methods as follows. The first is identical to the pooling method of BERT, which was using the output of the CLS token. The second strategy was to compute the man of all output vectors, called MEAN strategy and the last one was to compute a max-over-time of the output vectors called MAX strategy [29].

After the pooling method to extract the vector representation of news data of 768 dimensions, it is fed into the fully connected dense layer for fine-tuning process. The experiment conducted hyperparameter tuning on the size of the dense layer dimension, dropout ratio, training batch size, and pooling method. In this paper, one fully connected dense layer with dimensions of 32 followed by the last classifier layer and softmax layer for classification purposes are used. The dense layer

except the last classifier layer and softmax layer is followed by activation layer and dropout layer to prevent overfitting. The experiment significantly reduced the size of the dense layer because with larger size of dimension the model tends to overfit. Since the size of the labeled dataset is relatively small, the proposed method also decided to freeze weights in a pre-trained model to prevent overfitting. For the learning rate, the experiment utilized schedulers provided by the PyTorch python library [27]. The experiment attempted two different schedulers, which was LambdaLR and CosineAnnealingWarmRestarts [23]. The model is trained after the datset is partitioned into three dataset: Training, Validation and Test with ratio at 6:2:2. During the training process, the experiment implemented an early-stopping method to prevent overfitting and to speed up the training time. The early stopping method prevents the model from continuing training if the validation loss does not decrease for more than 3 times in a row.

3.4 Regression Model for Stock Volatility Prediction

Based on the previous studies, it is clear that involuntary turnover led to the increase in volatility. Hence, shareholders and investors could adjust their decision based on the output of the classification model. In this paper, a further experiment was conducted on the embeddings obtained from the language models to predict the value of expected volatility after a turnover event. In order to use the embedding vectors of news articles as input to regression models, dimension reduction was necessary because the size of the dataset is small in the experiment. Hence, Principal Component Analysis and Scree plot are used to determine the number of the reduced dimension size. After the classification model is trained and tested, news articles

related to CEO turnover are passed into the model and the embeddings of the articles are obtained. Then, the experiment performed PCA to check the number of principal components that explains at least 80% of the variances in the data as illustrated in Figure 3.4.

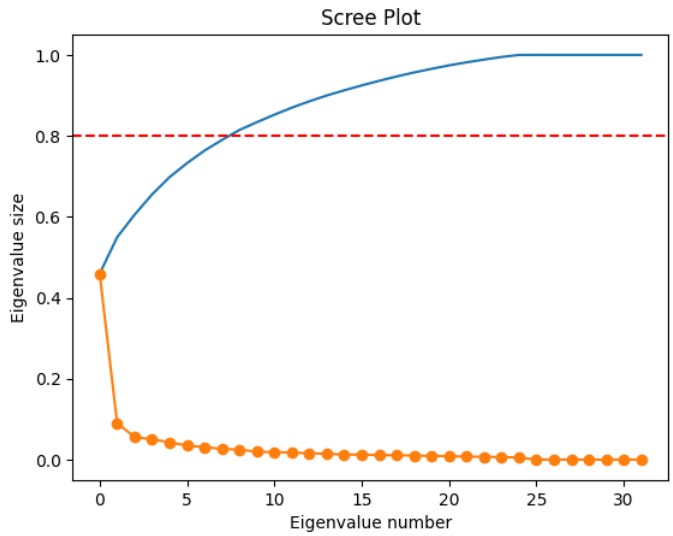


Figure 3.4: Scree Plot of PCA

For each turnover case, data contains the CEO turnover cause, date of CEO left his or her office, GVKEY of the firm. With the given information, the historical closing price of the firm before the turnover events could be obtained. The objective of this model is to verify the necessity of the features extracted from the embedding vectors of the news article related to the turnover event. Thus, the regression models are designed with and without the features extracted from the embedding and checked if the accuracy of the model with the embedding feature is higher than the model without it. For the regressor model, the ridge regression model and regression Artificial Neural Network are used.

For both models, the adjusted close price of 90 days prior to the turnover event and the features from the embedding vectors are used as input. In order to check the necessity of the embedding vector, other variables such as the firm's size or CRAs are not used in the models. For the regression models, the dataset was partitioned at a ratio of 7:3 into train and test dataset. Then, utilized scikit-learn package's cross validation ridge regression model to find the best alpha value. For the artificial neural network model, the experiment designed a simple fully connected neural network with one hidden dense linear layer of size 16 and a ReLU activation layer to introduce non-linearity in the model.

Chapter 4

Experiments and Results

4.1 Data

In this paper, three types of dataset are used. The first dataset is CEO turnover dataset. The dataset is obtained from an open-source database documenting the causes for CEO departure in S&P 1500 firms from 2000 through 2018 [11]. There are 8,194 samples of CEO turnover and each sample contains information including the Company Name, GVKEY, Departure Code and the date of departure from the office of CEO. The sample of data can be found in Figure 4.1 and the description of each data is explained in Figure 4.2 & 4.3.

	dismissal_dataset_id	coname	gvkey	fyear	co_per_rol	exec_fullname	departure_code	ceo_dismissal	interim_coceo
1054	1184	DUN & BRADSTREET CORP	4094	2009	23924	Steven W. Alesio	5.0	0.0	NaN
4841	5489	KLX INC	22343	2017	49900	Amin J. Khoury	7.0	0.0	NaN
1440	1618	GREAT LAKES CHEMICAL CORP	5306	2003	16586	Mark P. Bulriss	3.0	1.0	NaN
436	487	BLOUNT INTL INC	2271	2009	13414	James S. Osterman	5.0	0.0	NaN
3484	3902	WESTWOOD ONE INC -OLD	11450	2010	38055	Roderick M. Sherwood III	7.0	0.0	NaN

	tenure_no_ceodb	leftofc	notes	sources	eight_ks	cik
1054	1	2010-01-01 00:00:00	Retired and then starting working as a senior ...	https://www.businesswire.com/news/home/2011012...	NaN	1799208.0
4841	1	2018-09-10 00:00:00	On October 9, 2018, the acquisition of KLX Inc...	https://www.streetinsider.com/SEC+Filings/Form...	https://www.sec.gov/Archives/edgar/data/161789...	1617898.0
1440	1	2004-11-05 00:00:00	CEO Resignation - On November 5, 2004, Mark Bu...	https://www.sec.gov/Archives/edgar/data/43362/...	https://www.sec.gov/Archives/edgar/data/43362/...	43362.0
436	1	2009-12-18 00:00:00	Retired after 50 years of service amd would st...	https://www.marketwatch.com/story/blount-names...	https://www.sec.gov/Archives/edgar/data/100160...	1001606.0
3484	1	2011-10-21 00:00:00	merged Westwood One Network Radio business wit...	https://www.bloomberg.com/profile/person/40784...	NaN	NaN

Figure 4.1: The sample of CEO dismissal database of firms in S&P 1500

The dataset also contains the source link for each sample redirecting to the relevant sources from the Internet. However, most of the source links for turnover events in the past are deleted. Thus, this paper also utilized news data in S&P 1500 during the same time period. The news dataset contains headline, content and topic for each sample as shown in Figure 4. To fine-tune the language model for classification tasks, each sample from the dismissal dataset has to be mapped with appropriate news data instead of the source data provided in the dataset. In order to map the news data with dismissal data, this paper checked if the content of news data includes the name of the incumbent CEO and checked if the difference between the announced date of the news and the date of CEO left the office is no larger than a month. Finally, the dismissal dataset mapped with news data has 7,295 samples and unnecessary columns for fine-tuning the DistilBERT are dropped and the result is shown in Figure 4.4.

CEO Departure Reasons and Definitions

Code	Title	Brief Description
1	Involuntary - CEO death	The CEO died while in office and did not have an opportunity to resign before health failed.
2	Involuntary - CEO illness	Required announcement that the CEO was leaving for health concerns rather than removed during a health crisis.
3	Involuntary – CEO dismissed for job performance	The CEO stepped down for reasons related to job performance. This included situations where the CEO was immediately terminated as well as when the CEO was given some transition period, but the media coverage was negative. Often the media cited financial performance or some other failing of CEO job performance (e.g., leadership deficiencies, innovation weaknesses, etc.).
4	Involuntary - CEO dismissed for legal violations or concerns	The CEO was terminated for behavioral or policy-related problems. The CEO's departure was almost always immediate, and the announcement cited an instance where the CEO violated company HR policy, expense account cheating, etc.
5	Voluntary - CEO retired	Voluntary retirement based on how the turnover was reported in the media. Here the departure did not sound forced, and the CEO often had a voice or comment in the succession announcement. Media coverage of voluntary turnover was more valedictory than critical. Firms use different mandatory retirement ages, so we could not use 65 or older and facing mandatory retirement as a cut off. We examined coverage around the event and subsequent coverage of the CEO's career when it sounded unclear.
6	Voluntary - new opportunity (new career driven succession)	The CEO left to pursue a new venture or to work at another company. This frequently occurred in startup firms and for founders.
7	Other	Interim CEOs, CEO departure following a merger or acquisition, company ceased to exist, company changed key identifiers so it is not an actual turnover, and CEO may or may not have taken over the new company.
8	Missing	Despite attempts to collect information, there was not sufficient data to assign a code to the turnover event. These will remain the subject of further investigation and expansion.
9	Execucomp error	If a researcher were to create a dataset of all potential turnovers using execucomp (co_per_rol != l.co_per_rol), several instances will appear of what looks like a turnover when there was no actual event. This code

Figure 4.2: The description of each cause labeled in the database from [14]

Data Dictionary

Variable name	Type	Brief description
dismissal_dataset_id	int	The primary key. This will change from one version to the next. gvkey-year is also a unique identifier.
coname	str30	The Compustat Company Name.
gvkey	numerical long	The Compustat Company identifier.
cik	%10.0f	The company's Central Index Key
fyear	numerical long	The fiscal year in which the event occured.
co_per_rol	numerical long	The executive/company identifier from Execucomp.
exec_fullname	str50	The executive full name as listed in Execucomp.
departure_code	byte	The departure reason coded from criteria above.
ceo_dismissal	byte	A dummy code for involuntary, non-health related turnover (Codes 3 & 4).
interim_coceo	str7	A descriptor of whether the CEO was listed as co-CEO or as an interim CEO (sometimes interim positions last a couple years).
tenure_no_ceodb	byte	For CEOs who return, this value should capture whether this is the first or second time in office.
leftofc	Int (formatted %td)	Left office of CEO, modified occasionally from execucomp but same interpretation. The date of effective departure from the office of CEO. For companies that were acquired and thus dropped from Execucomp, *leftofc* is an attempt to capture the merger's effective date.
notes	strL	Long-form description and justification for the coding scheme assignment.
sources	strL	URL(s) of relevant sources from internet or library sources.
eight_ks	strL	URL(s) of 8k filing from the Securities and Exchange Commission from 270 days before through 270 days after the CEO's leftofc date which might relate to the turnover. Included here are any 8k filing 5.02 (departure of directors or principal executives) or simply item 5 if it is an older filing. These were collected without examining their content.

Figure 4.3: The description of each columns in database from [14]

	date	headline	content	Company Name	GVKEY
256267	2004-08-02 00:00:00	Dell Inc. Promotes Kevin Rollins	Dell Computer founder Michael Dell has passed ...	DELL INC	014489
145991	2011-06-01 12:00:00	Cedar Shopping Centers Inc. Announces Executiv...	Cedar Shopping Centers Inc. announced that Leo...	CEDAR INCOME FUND 1 LTD	013189
52127	2014-05-05 12:00:00	Aegion Corporation Announces Management Changes	Aegion Corporation announced that Charles R. "...	AEGION CORP	005978
38080	2019-07-01 20:45:00	Highwoods Properties, Inc. Announces Executive...	Highwoods Properties, Inc. announced Ed Fritsc...	HIGHWOODS PROPERTIES INC	030298
247915	2009-08-17 11:30:00	Corrections Corp. of America Announces Executi...	Corrections Corp. of America announced that Jo...	CCA PRISON REALTY TRUST	065084

Figure 4.4: The sample of news dataset related to CEO turnover

	text	label
3178	Veeco Instruments Inc. Announces Changes to Ex...	2
3972	Noven Pharmaceuticals Inc. Announces Board and...	6
5328	DSP Group Inc. Announces Management Changes. D...	2
2446	Phillips-Van Heusen Corp. Announces Executive ...	4
658	Coca-Cola Company Nominates Muhtar Kent as Dir...	2

Figure 4.5: The sample of CEO dismissal dataset mapped with news data

4.1.1 Label Engineering & Imbalance Dataset

In the CEO turnover dataset mapped with news data, the reasons for turnovers are categorized into 9 classes as described in Figure 4.2. However, since this study aims to automate the process of classifying the turnover causes based on the results of previous studies, irrelevant labels, i.e., Other, Error and Missing, were ignored. After the removal of irrelevant labels, the dataset now contains 7,163 samples. Unfortunately, as shown in Figure 4.6, the number of samples for each turnover factor is severely imbalanced.

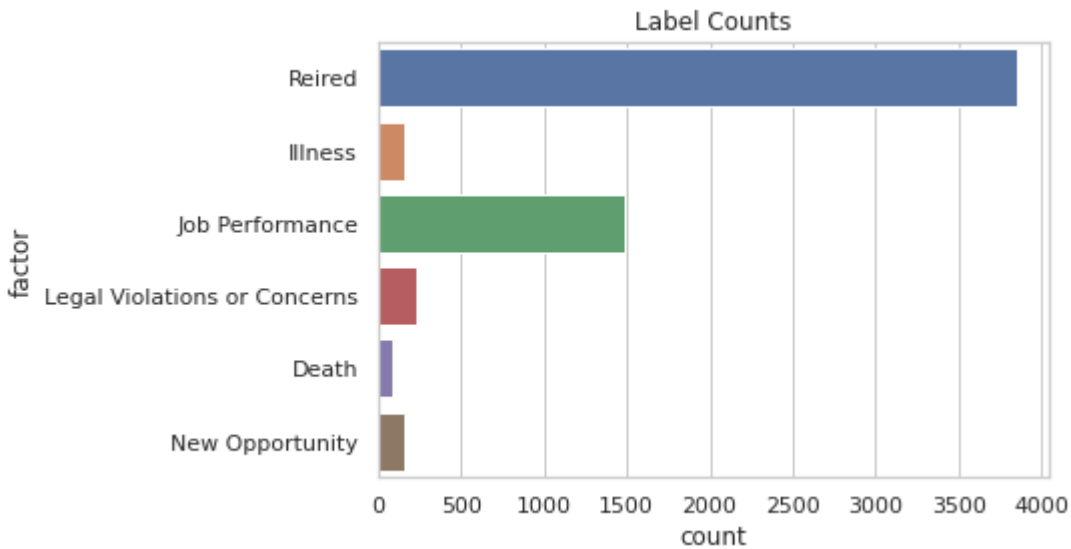


Figure 4.6: The distribution of classes in the database before label engineering

Prior to balancing the dataset, this study grouped the 6 labels into 2 based on the results of previous studies. In order to validate this decision, a post-event stock realized volatility was calculated for each sample. To calculate the annualized volatility, this study prepared closing prices of firms and index level data in S&P

1500 from 1996 to 2022. The index level data were used to neutralize the impact of the market on each firm's volatility. For all 6 classes, the distributions of post-event volatility per class is illustrated in Figure 4.7. Clearly, volatility after involuntary turnover events was higher than that of voluntary turnover events as expected. Thus, 6 classes of turnover causes could be grouped into 2 groups, i.e., voluntary and involuntary as shown in Figure 4.9 and the distribution of post-event volatility grouped by these two causes are illustrated in Figure 4.8.

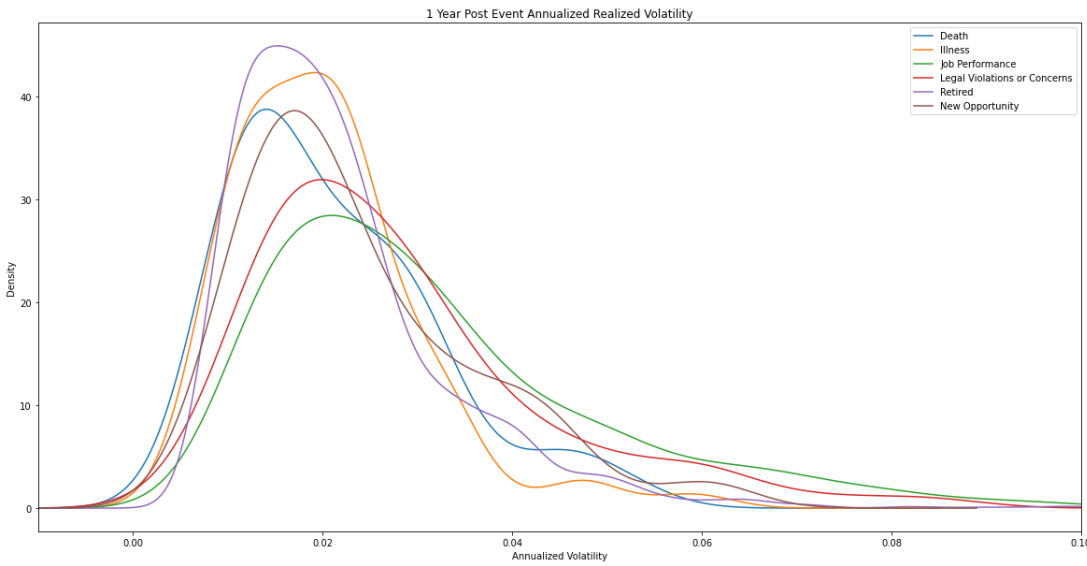


Figure 4.7: The distribution of post-event volatility by 6 turnover causes

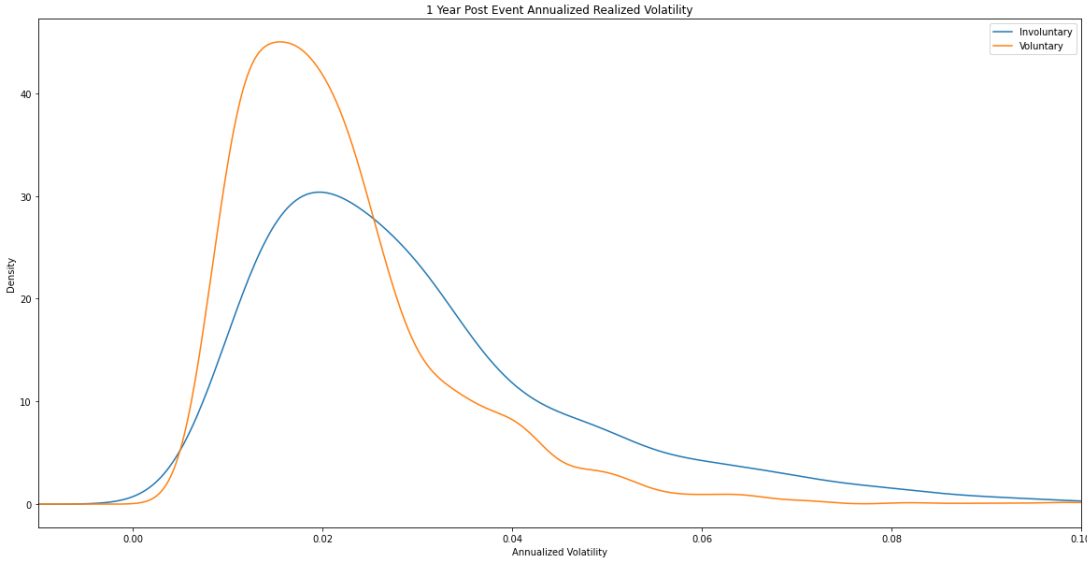


Figure 4.8: The distribution of post-event volatility by 2 grouped turnover causes

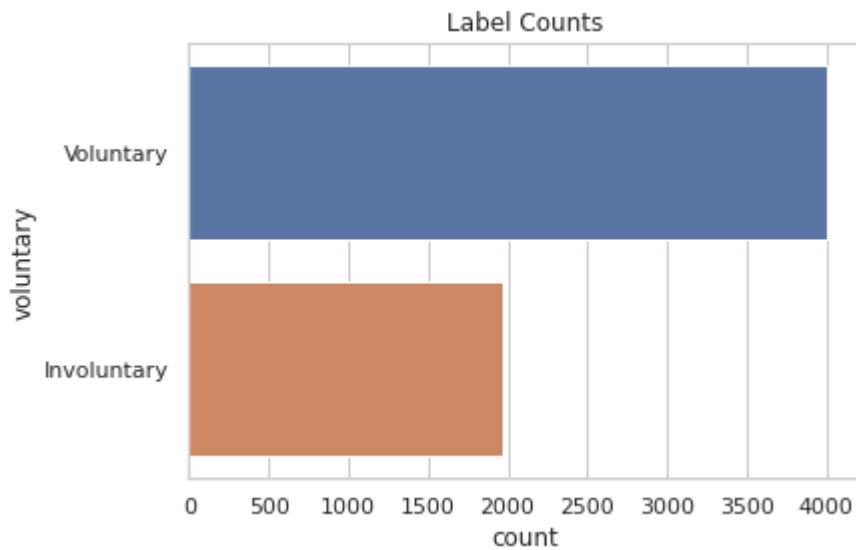


Figure 4.9: The histogram of grouped label count

Without balancing the departure code label, the model tends to classify every news data into the majority class. In order to balance the dataset, this paper

utilized a random oversampling method and text augmentation method. Random Oversampling method duplicates samples from minority classes. Among several NLP augmentation techniques such as random deletion, random swap as described in [13], this paper utilized a back translation method and synonym replacement method to balance the dataset. The final augmented dataset has 28,381 datasets with labels distributed as shown in Figure 4.10. However, the NLP augmentation decreased the performance of machine learning method. Since the size of minority classes is seriously smaller than that of majority classes, the random replacements of words tend to generate a duplicated sample which is not ideal to the frequency-based machine learning techniques. On the other hand, the NLP augmentation techniques improved the performance of deep learning method. The replacements and back translation provide more data to the machine to learn context of sentences without the frequency problem occurred in machine learning technique. Therefore, the datasets used in TF-IDF method both 2-classes and 6-classes news mapped turnover events with random oversampling method applied on the training dataset while the DistilBERT model utilized both random oversampling and NLP augmentation to balance the training dataset.

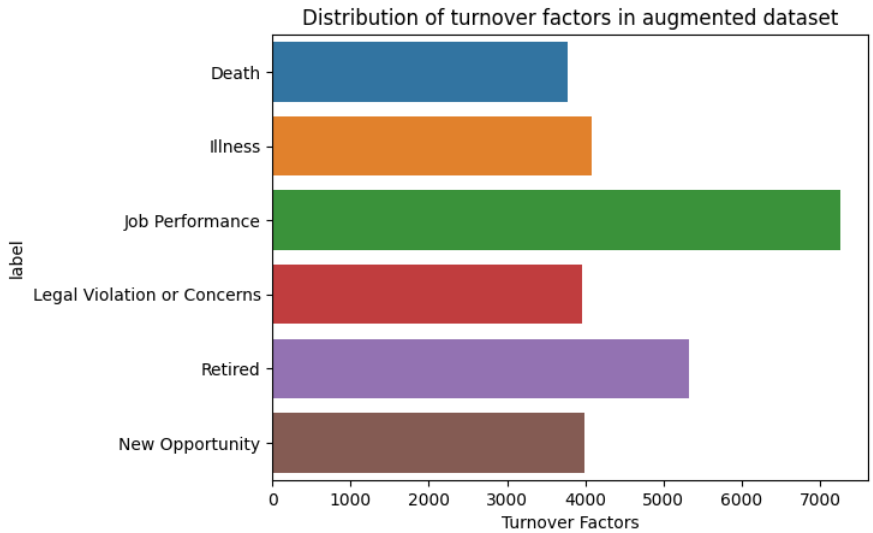


Figure 4.10: The histogram of dataset by turnover factor after NLP augmentation

4.2 Evaluation

The proposed framework contains two sequence classification tasks and one regression task.. The first classification task is to filter out CEO turnover related news from a vast amount of news on various topics. The second classification task is to classify whether the turnover was voluntary or involuntary from news data. Classification tasks are usually evaluated using a confusion matrix. Confusion matrix for binary classification has four components, which are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) as illustrated in Figure 4.11. There are four major metrics using these four values called Accuracy, Precision, Recall and F1 Score. In this experiment, if a turnover event is classified as involuntary, shareholders and investors would anticipate the volatility would increase. Hence, it is important to minimize the type I error, which is equivalent to getting a higher precision score. Moreover, since the dataset is originally imbalanced, it is important to get a higher F1 score.

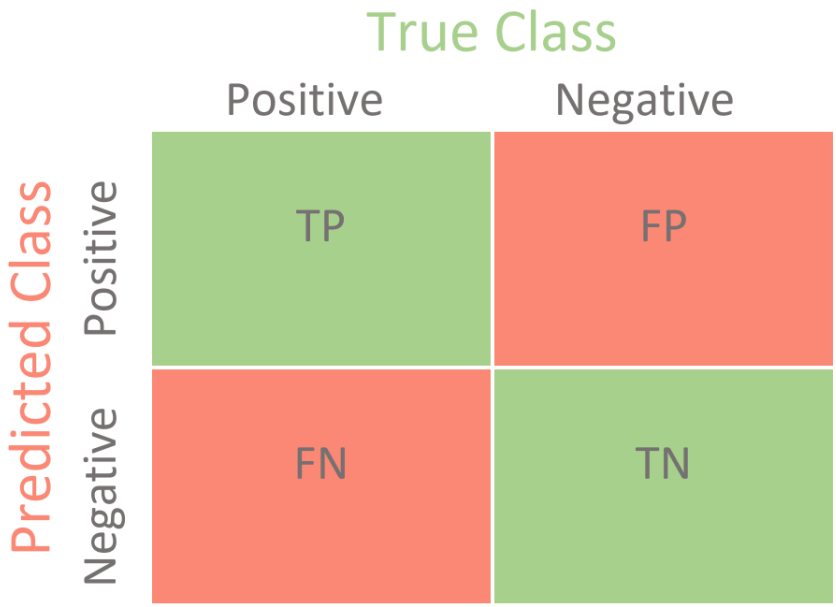
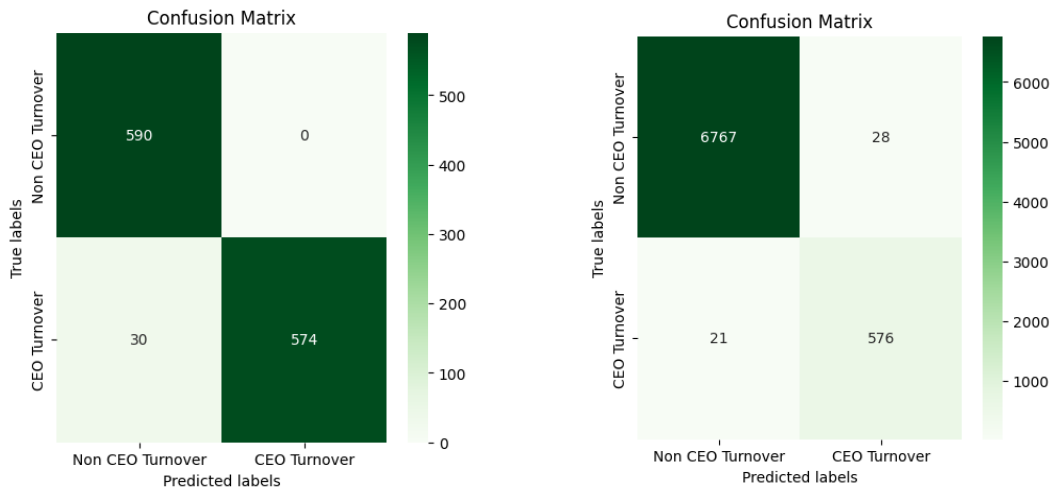


Figure 4.11: The components of a Confusion Matrix

The regression model is evaluated with R-Squared, Mean of Absolute value of Error(MAE) and Root Mean Squared Error (RMSE) metrics. R-squared is a measure that represents the proportion of the variance for a dependent variable, which is volatility, that's explained by independent variables which are close prices and news article embeddings. MAE and RMSE both measure the difference between true and predicted values. To test the performance of the regression models, the data from a different universe of securities are utilized. Although the classification model could not be tested on the different universe of security due to the absence of required labels, the regression models could be tested by estimating the post turnover volatility. For the evaluation part only, news articles and adjusted close price data are collected for the firms that are not in the S&P 1500 Index but in the Russell 3000 Index.

4.3 Results

For shareholders and investors to quickly capture the investment opportunity with CEO turnover events announced through news data, the first step is to filter out news related to CEO turnover from news published everyday with various topics. The result of the first classification model is described with the table and figure below. The performance of this model is remarkable in both balanced (4.12a) and imbalanced dataset (4.12b). The loss history of the model and the classification report are described in the Figure 4.13 & Table 4.1.



(a) Balanced

(b) Imbalanced

Figure 4.12: The confusion matrix of classification task for news related to turnover

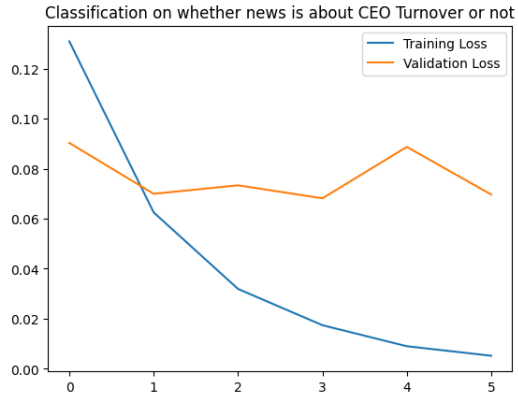
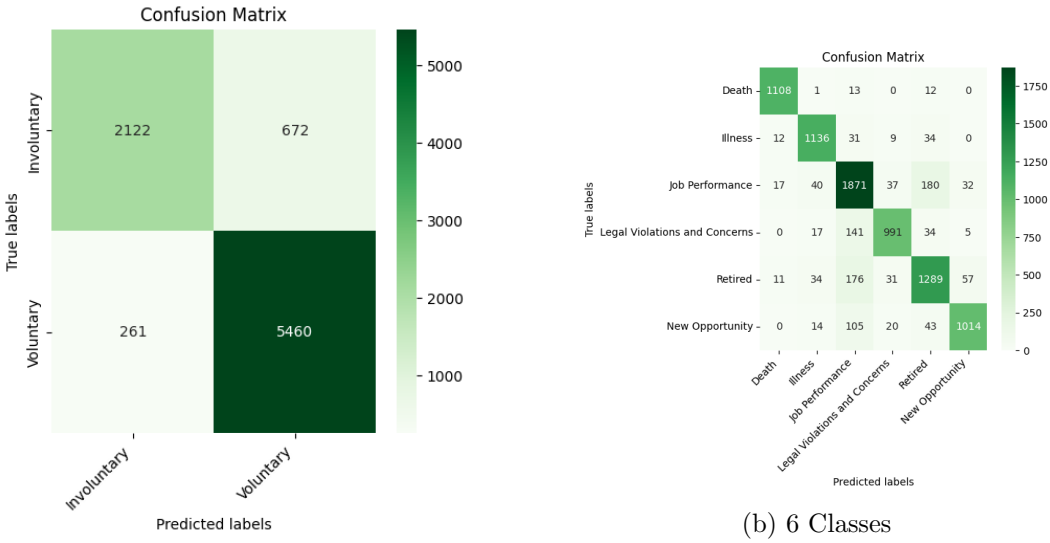


Figure 4.13: The training and validation loss of the first classification task.

Table 4.1: The classification report of the first classification task

	Precision	Recall	F1 Score	Support
Non CEO Turnover Related	0.951613	1.000000	0.975207	590.000000
CEO Turnover Related	1.000000	0.950331	0.974533	604.000000
Accuracy	0.974874	0.974874	0.974874	0.974874
Macro Avg	0.975806	0.975166	0.974870	1194.000000
Weighted Avg	0.976090	0.974874	0.974866	1194.000000

The second classification task is to identify turnover causes from news articles. The first attempt was to classify news data into 6 different labels. Then, the performance is compared with a model with 2 labels as described in the Data section. The experiment was conducted with TF-IDF and Naive Bayes Classifier and Linear SVC at first. The classification performance of TF-IDF on 6 labels and 2 labels are both remarkable. The accuracy scores are 87% for 6 labels and 88% for 2 labels, and the other score metrics are shown in Figure 4.14 and Table 4.2.



(a) 2 Classes

(b) 6 Classes

Figure 4.14: The confusion matrix of classification task for turnover causes

Table 4.2: The classification report for 6 labels classification using TF-IDF Naive Bayes Classifier

	Precision	Recall	F1 Score	Support
Death	0.965157	0.977072	0.971078	1134.000000
Illness	0.914654	0.929624	0.922078	1222.000000
Job Performance	0.800599	0.859440	0.828977	2177.000000
Legal Violations or Concerns	0.910846	0.834175	0.870826	1188.000000
Retired	0.809673	0.806633	0.808150	1598.000000
New Opportunity	0.915162	0.847826	0.880208	1196.000000
Accuracy	0.870112	0.870112	0.870112	0.870112
Macro Avg	0.886015	0.875795	0.880220	8515.000000
Weighted Avg	0.872058	0.870112	0.870389	8515.000000

From the results of studies by Clayton et al. (2005) and Li et al. (2021), if turnover events can be classified into either voluntary or involuntary correctly, it can still aid shareholders and investors to capture the investment opportunity and decide their strategic decisions. However, the TF-IDF embedding merely contains

the information of frequency of each term in the document and contextual meaning is ignored. Thus, in addition to the TF-IDF method, the experiment also conducted Deep Learning method by fine-tuning DistilBERT language model for sequence classification task in order to obtain sequence embeddings for each news article that contain contextual meaning so that it can be used as additional features in predicting stock volatility with a simple Neural Network and Ridge regression model.

The performance of DistilBERT with respect to accuracy, f1 score and precision is significantly lower than that of TF-IDF method especially for 6 label classification. However, for the 2 label classification task, the fine-tuned DistilBERT model could classify the turnover events into voluntary and involuntary with accuracy of 60.8% and precision score of 0.84 as can be inferred from Figure 4.15. Although the accuracy is not comparable to that of TF-IDF model, the precision score is in a considerable range and the contextual embeddings of news articles can be utilized for further research. Moreover, the vector representation obtained from DistilBERT compared to the one from the TF-IDF method has learned contextual meaning as Figure 4.16 describes. When the high dimensional vector representations of news articles from both methods are plotted on the 2 dimensional space using t-SNE [34], the embeddings from the DistilBERT classification model could cluster the articles by its label while the embeddings from the TF-IDF could not.

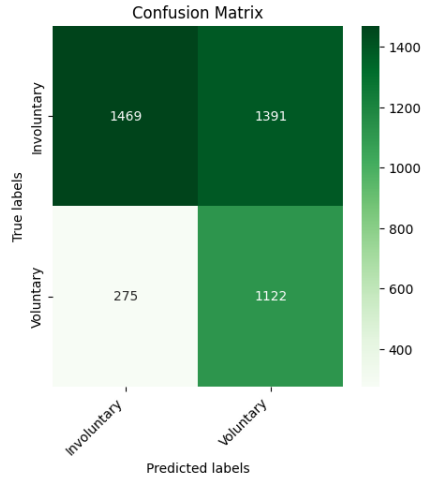
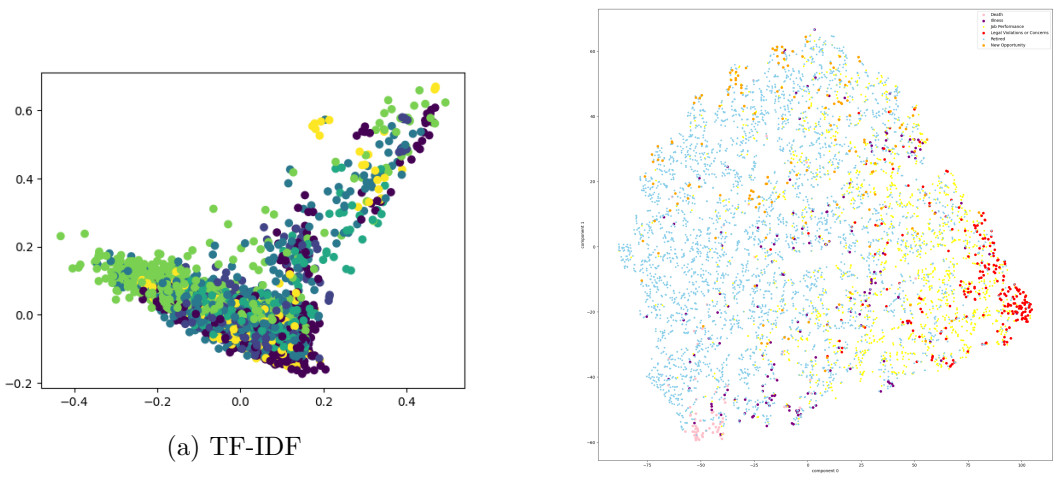


Figure 4.15: The confusion matrix for fine-tuned DistilBERT classification model.



(a) TF-IDF

(b) DistilBERT

Figure 4.16: News Articles Embedding Visualization using tSNE

For the final component of the proposed framework, a simple Neural Network with dense linear layer with activation function and the ridge regression model have been utilized to predict the stock volatility after the turnover events using historical closing price. As described in the Section 3, this paper conducted an experiment

on predicting stock volatility with and without news article embedding. There are 5,974 turnover events that are either voluntary or involuntary as explained above. For each turnover case, the experiment estimated the post turnover volatility for the target value. In the previous studies, volatilities before and after the turnover events are compared to investigate the impact of turnover on the volatility. However, in both studies by Clayton et al. (2005) and Li et al. (2021), post turnover volatility is calculated over a year or two. In this experiment, volatility after the post event for various time periods are compared and found out that volatility reacts quickly to the turnover events. Thus, the target variables of the regression model is used with the annualized volatility with rolling window of size 63 (3 months) after the turnover events. The regression model's result is described in Table 4.3.

Table 4.3: Regression Model Result with Embedding / without Embedding

	period	R^2	MAE	RMSE
Ridge	1 Month	0.022119/0.002738	0.016915/0.017065	0.023219/0.023448
	3 Month	0.014361/0.004389	0.010861/0.010915	0.014067/0.014138
	6 Month	0.013266/0.003352	0.007467/0.007533	0.009429/0.009477
Lasso	1 Month	0.020166/-0.000142	0.016927/0.017095	0.023242/0.023481
	3 Month	0.010853/-0.000479	0.010882/0.010929	0.014092/ 0.014172
	6 Month	0.010841/-7.48999	0.007476/0.007545	0.009441/0.009493
ANN	1 Month	-0.002869/-0.262589	0.017207/0.015375	0.024208/0.016999
	3 Month	-0.073778/-0.006411	0.011657/0.011335	0.015691/0.015797
	6 Month	-0.140540/0.013476	0.007269/0.007100	0.009708/0.009141

The influence of concatenation of embedding vectors into the input data was not effective as expected. However, the experiment trained the model with 10 days of log returns of adjusted close price for eacth turnover case with and without the embedding vectors reduced to dimension of size 2 with tSNE method. Thus, if additional features such as market variables would improve the performance of forecasting the

volatility. From this experiment, the model trained with embedding vectors had better R-squared score and slightly low RMSE values. Also, the experiment found out that if model tries to forecast volatility for longer period in the future, the performance increased. It could mean that the volatility increases after turnver events for a short period, but becomes stable and predictable as consumers and the company get used to the new CEO.

Chapter 5

Conclusion

In a fluctuating or volatile market, investing is always accompanied by risks. Stock volatility is a measure of variation in the stock price in a given period. Thus, higher volatility means higher risk. There are a plethora of unpredictable reasons for stock prices to change. However, there are also various trading strategies that aim to achieve profitable returns in this volatile market, such as the straddle strategy. Several studies have shown that CEO turnover impacts firm performance and hence its share price changes. Also, it has been thoroughly examined that CEO turnover causes affect differently on post turnover equity volatility. However, the problem this paper attempted to solve was the manual labeling process. By applying NLP to financial news, it was possible to capture turnover event causes from the news articles. With TF-IDF and Naive Bayes classifiers, 6 different causes were successfully classified. With the machine learning method for text classification, the experiment achieved 87% or higher accuracy, precision and F1-score. This machine learning model can signal to investors and shareholders that the stock volatility would go higher promptly.

The limitation of the proposed model is that it can only label the CEO turnover factor into one factor. However, CEO turnovers can have multiple causes or other

factors that are not described in the dataset. For example, the recent pandemic situation and gender equality and women's empowerment movement also caused CEO turnover. Another limitation is the size of the dataset. The dataset used in this experiment only contains turnover events that occurred in S&P 1500 from 2000 to 2018. Hence, this paper could extend the experiment by preparing a larger dataset or improve current labels to reflect changes in society. Moreover, the experiment found out that the magnitude of increase in volatility differs by the time period after the turnover events. Companies and investors get to know the new CEO after the turnover event and the share price becomes stable. However, in the Option trading strategies, the expiration time is also important. Thus, the experiment could further investigate on how long does the impact of turnover on volatility remains to add more value to the proposed method. On top of that, the result of the comparison between regression models with and without the embeddings of news articles related to the turnover event showed that the embeddings did not significantly improve the accuracy of forecasting. There are other language models besides DistilBERT such as Sentence-BERT that are more suitable for generating embeddings of sentences.

The main objective of this paper is to automate the process of examining the cause of turnover events. Previous studies analyzed the turnover event with respect to the firm performance, industry, market and news to identify what had caused the turnover events. Thus, the size of dataset was limited to less than 1,000 for big firms. However, since there are more than 300 million of companies worldwide and the turnover events are continuously occurring, it is valuable to automate the labelling process of turnover events so that shareholders and investors become agile. Although the performance of regression model is not attractive, the performance

of classification models are remarkable. Therefore, combing the results from previous studies, the proposed model is capable of signaling the increase in volatility to shareholders and investors to start with new strategies.

Bibliography

[1] A. Aizawa, *An information-theoretic perspective of tf–idf measures*, Information Processing & Management, 39 (2003), pp. 45–65.

[2] P. C. Andreou, C. Louca, and A. P. Petrou, *Ceo age and stock price crash risk*, Review of Finance, 21 (2017), pp. 1287–1325.

[3] P. Bambroo and A. Awasthi, *Legaldb: Long distilbert for legal document classification*, in 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), IEEE, 2021, pp. 1–4.

[4] O. Bandiera, A. Prat, S. Hansen, and R. Sadun, *Ceo behavior and firm performance*, Journal of Political Economy, 128 (2020), pp. 1325 – 1369.

[5] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*, " O'Reilly Media, Inc.", 2009.

[6] I. Challenger, Gray & Christmas, *December 2021 ceo final report; 1,337 ceo changes in 2021, up 22021.* `https://www.challengergray.com/blog/december-21-ceo-report-1337-ceo-changes-in-2021-up-2-over-2020-26-of-new-ceos` 2022.

[7] H. Chang, J. Chen, W. M. Liao, and B. K. Mishra, *Ceos'/cfos' swearing by the numbers: Does it impact share price of the firm?*, The Accounting Review, 81 (2006), pp. 1–27.

[8] M. C. Clayton, J. C. Hartzell, and J. Rosenberg, *The impact of ceo turnover on equity volatility*, The Journal of Business, 78 (2005), pp. 1779–1808.

[9] E. Dedman and S. W.-J. Lin, *Shareholder wealth effects of ceo departures: Evidence from the uk*, Journal of Corporate Finance, 8 (2002), pp. 81–104.

[10] M. L. DeFond and C. W. Park, *The effect of competition on ceo turnover*, Journal of Accounting and Economics, 27 (1999), pp. 35–56.

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

[12] V. Dogra, A. Singh, S. Verma, N. Jhanjhi, M. Talib, et al., *Analyzing distilbert for sentiment classification of banking financial news*, in Intelligent Computing and Innovation on Data Science, Springer, 2021, pp. 501–510.

[13] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, *A survey of data augmentation approaches for nlp*, arXiv preprint arXiv:2105.03075, (2021).

[14] R. Gentry, J. Harrison, T. Quigley, and S. Boivie, *Open sourced database for ceo dismissal 1992-2018*, Strategic Management Journal, (2021).

[15] C. R. Harvey and R. E. Whaley, *Market volatility prediction and the efficiency of the s & p 100 index option market*, Journal of Financial Economics, 31 (1992), pp. 43–73.

[16] S. Hazarika, J. M. Karpoff, and R. Nahata, *Internal corporate governance, ceo turnover, and earnings management*, Journal of Financial Economics, 104 (2012), pp. 44–69.

[17] R. Jindal and S. Taneja, *A lexical approach for text categorization of medical documents*, Procedia Computer Science, 46 (2015), pp. 314–320. Proceedings of the International Conference on Information and Communication Technologies, ICICT 2014, 3-5 December 2014 at Bolgatty Palace & Island Resort, Kochi, India.

[18] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*, in European conference on machine learning, Springer, 1998, pp. 137–142.

[19] M. Joydwip, *Confusion matrix for your multi-class machine learning model*. `https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826`, 2020.

[20] K. S. Kumar, J. Desai, and J. Majumdar, *Opinion mining and sentiment analysis on online customer review*, in 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), IEEE, 2016, pp. 1–4.

[21] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, *Albert: A lite bert for self-supervised learning of language representations*, arXiv preprint arXiv:1909.11942, (2019).

[22] H. Li and P. Farah, *Ceo turnover and equity volatility*, (2021).

[23] I. Loshchilov and F. Hutter, *Sgdr: Stochastic gradient descent with warm restarts*, arXiv preprint arXiv:1608.03983, (2016).

[24] J. F. Low, B. C. Fung, F. Iqbal, and S.-C. Huang, *Distinguishing between fake news and satire with transformers*, Expert Systems with Applications, 187 (2022), p. 115824.

[25] A. Mackey, *The effect of ceos on firm performance*, Strategic management journal, 29 (2008), pp. 1357–1367.

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

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *Pytorch: An imperative style, high-performance deep learning library*, in Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.

[28] S.-H. Poon and C. W. Granger, *Forecasting volatility in financial markets: A review*, Journal of economic literature, 41 (2003), pp. 478–539.

[29] N. Reimers and I. Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks*, arXiv preprint arXiv:1908.10084, (2019).

[30] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter*, arXiv preprint arXiv:1910.01108, (2019).

[31] D. Sculley and G. M. Wachman, *Relaxed online svms for spam filtering*, in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 415–422.

[32] Statista, *Estimated number of companies worldwide from 2000 to 2021 (in millions) [graph].* `https://www.statista.com/statistics/1260686/global-companies/`, 2022.

[33] B. Trstenjak, S. Mikac, and D. Donko, *Knn with tf-idf based framework for text categorization*, Procedia Engineering, 69 (2014), pp. 1356–1364.

[34] L. Van der Maaten and G. Hinton, *Visualizing data using t-sne.*, Journal of machine learning research, 9 (2008).

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

[36] S. Vijayarani, M. J. Ilamathi, M. Nithya, et al., *Preprocessing techniques for text mining-an overview*, International Journal of Computer Science & Communication Networks, 5 (2015), pp. 7–16.

[37] S. Xu, *Bayesian naïve bayes classifiers to text classification*, Journal of Information Science, 44 (2018), pp. 48–59.

[38] Y. Zhang, R. Jin, and Z.-H. Zhou, *Understanding bag-of-words model: a statistical framework*, International journal of machine learning and cybernetics, 1 (2010), pp. 43–52.

국문초록

대표이사 변경은 기업에서 발생하는 이벤트 중의 하나이며 해당 기업에 큰 영향을 준다. 대표이사의 역할은 기업의 전반적인 경영 전략 등을 담당하며, 때문에 대표이사의 변경은 기업의 경영 전략뿐만 아니라 소비자 인식, 투자 전략 등에 영향을 주며 이는 해당 기업의 주가에도 반영된다. 그렇기 때문에 투자자들에게도 대표이사 변경은 눈여겨볼 이벤트이며, 특히 변경 사유는 투자자들이 주의하는 부분이다. 대표이사 변경 사유는 이벤트 발생 이전의 주가의 변동, 기업 실적 등을 통해서도 대략적으로 유추할 수 있다. 하지만 대표이사 변경에 관련된 뉴스에는 보다 직접적으로 사유에 대해서 찾아볼 수 있다. 특별한 이유없이 나이가 들거나 그로인해 생긴 질병으로 인해 본인의 의지로 대표이사직에서 물러나거나 특별한 이유로 인해 강제적으로 물러나는 등의 사유가 있을 수도 있고, 또 다른 경우에는 다음 후임자에 대한 정보도 파악할수 있다. 본 논문에서는 뉴스로 부터 자연어처리를 통하여 대표이사 변경의 사유를 분류하는 모델을 제안한다. 기존의 수기로 레이블링 하는 방식을 자동화하는 것에 의의를 둔다. 단어의 빈도와 역 문서 빈도를 활용한 TF-IDF 모델을 변경 사유 분류 모델의 벤치마크 모델로 활용하고, 트랜스포머 구조의 사전학습된 언어모델을 사용하여 대표이사 변경 사유를 분류하는 테스크를 통하여 파인튜닝하는 과정에서 뉴스의 임베딩을 추출한다. 대표이사 변경 사유 분류를 통하여 사유에 따라 변경 이후 주가 변동성이 증가할 것이란 신호를 투자자들에게 제공함으로써 빠르게 투자 전략을 조정할 수 있도록 기여한다. 또한, 언어모델에서 얻은 맥락을 포함한 벡터 임베딩을 활용하여 이벤트 발생 이후 해당 기업의 주가 변동성을 예측하는 모델을 구축하여 사유 분류 모델의 활용도를 실험하였다.

주요어: 텍스트 분류, 대표이사 변경, 대표이사 변경 사유 분류 자동화, 주가 변동성 예측

학번: 2020-23018

감사의 글

서울대학교 산업공학과의 모든 식구들께 감사드립니다.