Aalto University
School of Business

A TEXT-BASED APPROACH TO INDUSTRY CLASSIFICATION

Word2Vec application in constructing an industry classification to describe the economic relatedness between companies

Master's Thesis
Taeyoung Kee
Aalto University School of Business
Information Service Management
Spring 2018

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of master's thesis

Author Kee Taeyoung	
Title of thesis A text-based approach to industry classification	
Degree Master of Science in Economics and Business Administration	
Degree programme Information Service Management	
Thesis advisor(s) Pekka Malo	
Year of approval 2018 **Number of pages** 54	**Language** English

Abstract

Industry classification schemes are a critical topic in academic research due to their use in combining companies into smaller groups that share similar characteristics. Although many studies in the domains of economics, accounting and finance depend heavily on these schemes, existing ones have significant limitations mainly due to their stagnant nature, which makes the schemes incapable of adapting to constant innovation and technological development.

The objective of this thesis is to propose an automated, text-based industry classification scheme that can reflect constant changes in industry scope. This thesis approaches the research problem by answering two research questions. First, it studies whether it is possible to build an industry classification scheme by using word-embedding vectors extracted from news article. Second, this thesis identifies the benefits of a text-based industry classification scheme in comparison with existing classification schemes. To identify benefits, both qualitative and quantitative assessments are conducted to measure the performance.

In the construction of an industry classification scheme, word-embedding vectors generated from news articles are used. The vectors are built using the Word2Vec algorithm. Word2Vec is a recently developed text-mining tool and is excellent at capturing the relationships between words and expressing them in a quantifiable format.

The key findings of this thesis are twofold. First, it is technically possible to build an automated, text-based industry classification scheme by using word-embedding vectors. Two methods of building the scheme are proposed. Second, the proposed text-based scheme performs well in classifying companies into relevant business categories. In addition, the cluster-based scheme exhibits better performance in grouping companies into financially homogenous groups when parameters are optimized.

The results suggest that a text-based industry classification scheme can serve as an alternative to existing industry classification schemes if parameters are optimized to the purpose of its use. The usefulness of the scheme is expected to increase due to the accelerating speed of innovation and technological development.

Keywords Industry classification, Word2Vec, text mining, cluster analysis

Table of Contents

List of Tables

List of Figures

1 Introduction

In academic research, industry classification schemes play a critical role: on average, 30% of top accounting, economic and finance journals have employed industry classification schemes in their papers (Weiner, 2005). They are important tool for researchers since they often need to combine companies into peer groups, which are expected to share similar characteristics.

Currently, researchers use industry classification schemes developed and maintained by governments or financial institutions, such as the SIC, NAICS, GICS and FF. Despite playing a critical role in research, existing industry classification schemes entail several limitations that may distort the results of studies. One of the limitations comes from the fact that there is no absolute classification scheme that can be considered a general standard. Thus, different organizations have developed schemes with different focuses and purposes. As a result, these schemes do not correspond to each other (Bhojraj, Lee & Oler, 2003), which causes confusion. In addition, existing classification schemes cannot adapt to constant innovation and technological changes due to their stagnant nature; further, they are created and managed by either governments or financial institutions, which need significant resources to revise and update them (Dalziel, 2007). Last, existing industry classification schemes do not sufficiently fulfill their role of describing financial homogeneity between firms within the same industry.

Studies have suggested that the quality of static industry classification schemes is decreasing (Weiner, 2005) due to the rapid changes in modern society. Thus, an automated, easy-to-maintain industry classification scheme is necessary to improve the quality of industry classification schemes in serving their purpose in academic research.

Several studies have proposed new form of industry classification schemes that address the limitations of existing ones. Some of them suggest text-based industry classification schemes due to their advantage in creating up-to-date industry classification by using computer-based automated calculation. This thesis contributes to the existing literature by examining the possibilities of building automated industry classification schemes using text analysis methodology.

1.1 Objectives and Research Questions

The objective of this thesis is to propose an automated industry classification scheme by employing the latest text analysis methodology. Specifically, this thesis utilizes word-embedding vectors extracted from news articles to capture up-to-date information about companies and industries.

The theoretical contribution of this thesis is the application of the text-mining method, word-embedding vectors based on a neural network, to the industry classification scheme, the critical topic in economic research. The empirical aim of this thesis is to propose a classification scheme that performs better than existing industry classification schemes in explaining homogeneity between companies.

Based on these objectives, the research questions are as follows:

1. Can we build an automated industry classification scheme by employing word-embedding vectors derived from news articles?
2. If yes, what are the benefits of the scheme in comparison with existing industry classification schemes?

To identify the benefits of a text-based industry classification scheme, proper assessment must be conducted. In this thesis, the proposed industry classification schemes are evaluated based on qualitative and quantitative criteria in comparison with the GICS classification system. Figure 1 presents an overview of the methodology used.

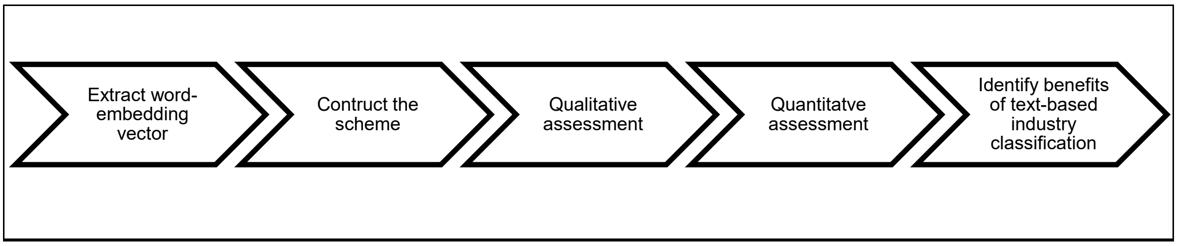

Extract word-embedding vector
Contruct the scheme
Qualitative assessment
Quantitatve assessment
Identify benefits of text-based industry classification

Figure 1. Research methodology overview

This thesis contributes to both theoretical and empirical domains by generating a text-based industry classification scheme and examining the benefits of the proposed scheme using quantitative analysis. The quantified assessment of the proposed scheme can benefit researchers in determining a suitable scheme for their research.

1.2 Main Findings

The main findings of this thesis are twofold. First, it is possible to build an industry classification scheme by employing word-embedding vectors extracted from news articles. By using Word2Vec, firms are represented as vectors, which enables the application of numeric calculation and quantitative analysis. Two different industry classification schemes are proposed using cosine similarity and cluster analysis. Second, the suggested schemes are better than existing schemes in accommodating constant changes. Additionally, the proposed text-based schemes can perform better in explaining economic relatedness than existing classification schemes when the involved parameters are optimized.

1.3 Structure

This thesis is structured in five chapters. Chapter 2, which follows the introductory chapter, consists of a literature review that serves as the foundation of the thesis. The literature review is divided into two parts: the first presents past articles about industry classification schemes and establishes the validity of this thesis; the second introduces the theoretical background of text mining, starting from a broad perspective that narrows to Word2Vec, which serves as grounding for the research methodology selection. In Chapter 3, the research methodology is presented. I discuss the data and how the text-based industry classification scheme is constructed by employing word-embedding vectors. Chapter 4 presents the research results, focusing on the validity of the suggested schemes. Finally, the conclusions, including empirical contributions, are covered in Chapter 5.

2 Literature Review

2.1 Importance of Industry Classification Schemes

Industry classification systems are used widely in different domains, from academic research to more practical applications. Industry classification is needed to combine firms that are homogenous in terms of certain characteristics, which may vary depending on the purpose of the research or its application. Weiner (2005) has examined the usage of industry classifications in accounting, finance and economic journals and discovered that, on average, 30% of top finance and economic journals employ industry classification systems in their papers. For instance, *The Journal of Accounting Research* exhibits an industry classification in more than half of its papers.

Industry classification schemes serve various purposes. Table 1 displays the major purposes of industry classification schemes for papers published in journals between 1995 and 2003. According to Weiner (2005), the first is to select suitable peer groups that are expected to share similar financial characteristics, and the second is to restrict samples based on industry. For instance, many papers exclude banks or utilities due to distinctive financial characteristics that may hinder analysis. The third most important purpose is to develop industry dummies. Average 8% of the papers conducted regressions with industry dummies to identify industry effects. A similar purpose is the coverage of industry effects, which was used by 12 % of the papers, which present descriptive statistics of industry members. Several papers examine diversification of companies whether there is a premium or discount. The international use is not an original purpose, but it is included in this categorization because many papers use Worldscope classification instead of Compustat SIC codes.

Table 1:Purpose of industry classifiction scheme

Purpose	AER	JAE	JAR	JoF	JFE	Average
Comparable selection	34%	25%	16%	23%	42%	27%
Sample restriction	22%	39%	44%	46%	29%	40%
Industry dummies	*11%*	*12%*	*10%*	*2%*	*9%*	*8%*
Industry effects	*15%*	*4%*	*16%*	*15%*	*14%*	*12%*
Industry distribution	*7%*	*6%*	*3%*	*5%*	*17%*	*8%*
Diversification	*2%*	*3%*	*1%*	*8%*	*2%*	*4%*
International use	*8%*	*10%*	*5%*	*5%*	*4%*	*6%*
Other	*9%*	*4%*	*5%*	*3%*	*3%*	*4%*

The results are based on all papers in included journals (1995 – 2003). Since one paper can have one or more purposes, the percentage value does not sum to 100. AER is the American Economic Review*, JAE is the* Journal of Accounting and Economics*, JAR is the* Journal of Accounting Research*, JoF is the* Journal of Finance *and JFE is the* Journal of Financial Economics.

Industry classification schemes are important to analysis; choosing the right classification scheme for a study is critical due to measurement errors or selectivity biases. For example, a diversification discount (typically manifested by a comparison between the value of total sectors of a diversified company and corresponding single-sector companies) is calculated based on the average multiples of firms belonging to same industry group, and the level of discount varies from 18% to 0% depending on the selected classification scheme and the level of industry.

2.2 Current Industry Classification Schemes

Either government or financial institutions have developed the industry classification schemes currently used by financial and accounting research. I first discuss the historical background, structure and methodology of four commonly used classification schemes to help understand further discussion concerning their limitations.

Standard Industrial Classification (SIC)

SIC, the oldest of the four, was developed in the 1930s by the Interdepartmental Committee on Industrial Classification with the aim of suggesting a classification scheme as the standard classification of the federal government (Bhojraj, Lee & Oler, 2003). SIC is widely used by government agencies and financial researchers it has a hierarchical structure represented by four-digit codes that begin with broad industries and narrows to specifics. Each firm is connected to one specific code based on sales in the largest segment.

North American Industry Classification Systems (NAICS)

NAICS was developed in 1999 through the joint effort of government statistical agencies in Canada, Mexico and the United States to reflect rapid changes in world economies. NAICS aimed to improve SIC by employing a production-based framework (SAUNDERS, N. C., 1999) NAICS has the same top-down hierarchical structure as SIC, wherein 1,170 country-specific sub-industries are defined by six-digit codes.

SIC and NAICS were both developed by government agencies targeting the collection of industrial statics. They were "erected on a production-oriented or supply-based conceptual framework in that establishments are grouped into industries according to similarity in the process used to produce goods or services" (OMB, 1998, p.11). Neither was developed in consideration of financial characteristics.

Fama and French (FF)

FF was developed in 1997 by the financial academics Fama and French in their study of the industrial costs of capital. They developed an algorithm that classifies existing SIC groups into 48 industry groups based on shared common risk characteristics. Although their validity has never been tested, several researchers use FF systems researchers (Bhojraj, Lee & Oler, 2003).

Global Industry Classification Standard (GICS)

GICS was developed as a collaboration between the financial service providers Morgan Stanley Capital International (MSCI) and Standard Poor's (S&P). As leading providers of investment decision-making support tools, MSCI and S&P developed an industry classification system based on the needs of finance professionals. According to the GICS guide book (Global Industry Classification Standard [GICS®], 2006), its aim is to support investment research and the asset management process for financial experts.

The GICS system differs from the SIC and NAICS systems in terms of their basis in classifying companies. While SIC and NAICS use a production and supply-based approach, GICS employs information about a firm’s revenues, earnings and market perception from various sources, including annual reports, financial statements and other industry reports.

2.3 Limitations of Current Industry Classification Schemes

Although they are widely used for both academic and practical purposes, current industry classification schemes are limited for the reasons I discuss in this section.

The first limitation derives from the concordance between classification systems. Currently, the SIC system is most commonly used in economic journals, but researchers also use other systems. As discussed above, classification systems were developed by different institutions with different focuses, which has resulted in different structures and methodologies. Bhojraj, Lee and Oler (2003) have examined the similarity between classification systems with the SIC system based on comparison. For each two-digit SIC code, they present the equivalent industry in other systems (NAICS, FF and GICS) based on the number of matching firms. The match is quite poor; for instance, the SIC industry 50 contains 30 firms, of which only five are found in GICS industry 452030. Overall, if one chooses the NAICS classification based on that company’s SIC industry, he or she will be correct 80% of time on average. For FF, 84% will be correct.

The mismatch between classification systems can be manifested not only between different systems but also within systems. Although public institutions have developed the methodology and structure of the classification systems, they do not classify companies. Data vendors and commercial organizations link firms to these systems based on their own interpretation of systems and company information. Weiner (2005) has presented a concordance of 0.46 to 0.79 between a single SIC system based on two different vendors (Worldscope and Compustat) and based on their own concordance measure ranging from 0 to 1.

The second limitation occurs due to consistent innovation and technological change, to which current industry classification systems cannot adapt in a timely manner. Dalziel

(2007) has examined changes that may pose challenges to current classification systems. The primary example is knowledge-based components such as software and microsystems. Although there is a wide range of knowledge-based components with different purposes, all companies that produce software are classified with a same code (5110 – Software publishers), and all firms that produce microsystems are classified as 3344 – Semiconductor and Other Electronic Component Manufacturing. In addition, technological change linked to production process can have a significant impact on industry classification. For example, the invention of electronic computing and semiconductors invalidated the original SIC code that classifies firms that produce computers in SIC Group 35 - Machinery (except electrical), which was appropriate prior to the emergence of electrical computing. Last, blurring the distinction between firms that produce products and those that provide services makes it difficult to embrace supply-based classification systems.

The challenges that static classification systems face in accommodating a changing industry scope are visible in Weiner's research (2005, p 24) on concordance between classification systems between 1990 and 2002; concordance level decreased drastically, while the number of valid industries almost doubled because of the emergence of new firms. Regardless of this visible challenge, existing classification systems cannot keep pace with industry evolution due to the serious cost involved in revisions. According to the U.S census bureau, the SIC system, the most widely used in accounting research, was last updated in 1987. The NAICS system is reviewed every five years and, and the last revised version was published in 2017 (NAICS update process fact sheet, 2017). The Economic Classification Policy Committee (ECPC), which is responsible for the maintenance and review of NAICS, mentions "balancing the costs of change against the potential for more relevant and accurate economic statistics requires significant input from data providers, data producers, and data users" (NAICS update process fact sheet, 2017).

Finally, existing industry classification systems do not adequately address the homogeneity of firms that belong to single industry group, especially in terms of financial characteristics. In capital market research, an industry classification scheme is used to group companies into more homogenous groups, with the expectation that this grouping provides better context for analysis. Another common application concerns quantitative trading, in which stocks are grouped into baskets and firms belonging to same basket are expected to be highly correlated in returns (Kakushadze and Yu, 2016), information indispensable in various modelling and statistical strategies.

One study (Bhojraj, Lee & Oler, 2003) has evaluated industry classification systems focusing on the main applications in financial research. It compares SIC, NAICS, GICS and FF in their capacity to explain homogeneity between firms by measuring similarities in financial metrics such as stock returns, valuation multiples, growth rates, R&D expenditure and other firm-level ratios from financial statements. Based on OLS regression with a firm-level dependent variable and a within-industry average independent variable, the study examined the adjusted R squared for all S&P 1,500 firms in terms of the above metrics. The level of adjusted R squared is low, with a maximum 26.5% for GICS classification system. The result is lower for valuation multiples (the GICS system yields an adjusted R squared of 23.3%, 37.4% and and 15% for price-to-book, enterprise value-to-sales and price-to-earnings, respectively).

Various approaches have been proposed to address the limitations in widely used industry classification systems. Dalziel (2007) has introduced a systems-based approach to identify industry groups. Instead of grouping companies based on their supplies, the author focuses on the demand that companies try to address.

While some have focused on the structural relationships between firms, other researchers have suggested industry classification schemes based on a quantitative approach, building scheme based on economic relatedness between firms based on quantitative financial characteristics. In development, many researchers used cluster analysis, a collection of statistical methods to classify single entities into distinct groups. For instance, Guptar and Huefner (1972) use hierarchical cluster analysis based on key financial ratios in explaining industry characteristics. Jensen (1971) also employs cluster analysis to develop a statistical classification technique useful in performance comparison. Weiner (2005) has developed an independent classification system using cluster analysis, based on the agglomerative hierarchical method and value drivers.

2.4 Theoretical Background of the Text-Based Approach

In this section, I briefly discuss background information about text mining and explain why it is relevant in addressing limitations involved in existing industry classification systems. Additionally, this section explains the relevance of Word2Vec, the critical tool used in developing word vectors, for this research.

2.4.1 Text Mining and Natural Language Processing (NLP)

Due to the development of computer networks, the amount of machine-readable documents has increased tremendously. According to one estimate, 85% of business information resides in text form (Text mining summit conference). However, the significant amount of information contained in unstructured texts cannot be used for normal data mining processes conducted by computer, which recognize text as a simple sequence of strings. Thus, various preprocessing methodologies and algorithms are necessary to extract useful information from text data. The process of extracting knowledge from unstructured text data is generally referred to as text mining.

Text mining was first mentioned by Feldman and Dagan (1995), and it concerns the machine-supported analysis of text. Currently, the field of text mining research addresses problems including text representation, information extraction, classification, clustering or the modeling of underlying patterns (Hotho, Nurnberger & Paaß, 2005). To achieve this, known data mining algorithms must be adapted to text data to select appropriate characteristics and gain domain knowledge.

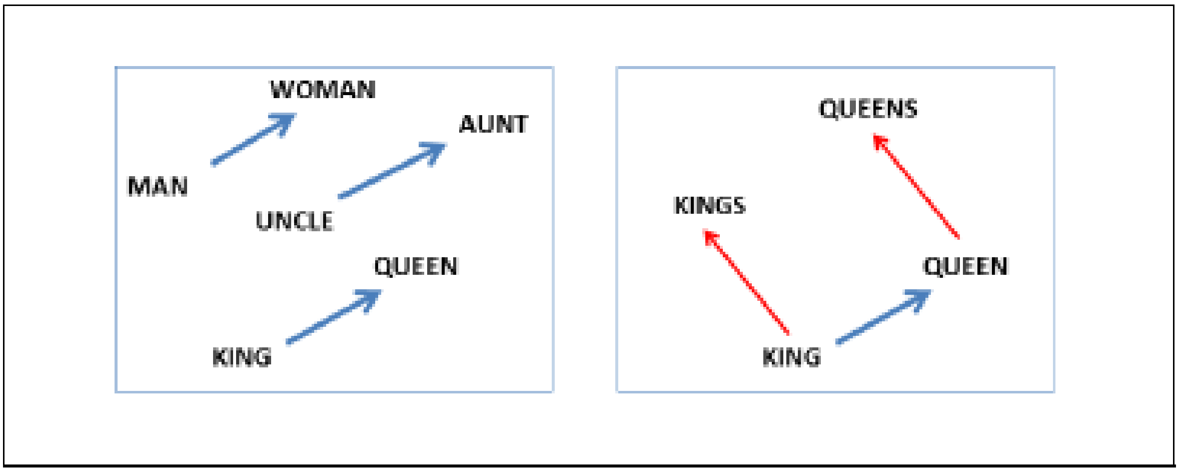
The objective of Natural Language Processing (NLP) is to achieve better understanding of natural language with the use of computers. It aims to summarize and organize natural language to improve comprehensibility and usefulness. More specifically, NLP aims to extract 'simpler representation' that is easier for the computers to manipulate (Collobert et al., 2011). Although a complete understanding is not yet possible, researchers have developed several tasks beneficial to developing application and analysis. These tasks can be syntactic (e.g., part-of-speech tagging, parsing and chunking) and semantic (e.g., semantic-role labeling, word-sense disambiguation and named entity recognition).

2.4.2 Distributed Representations of Words Using Neural Networks

The distributed representation of words in a vector space is one of the areas developed through the application of artificial neural networks to NLP problems. The earliest use of the idea dates to 1986, with a study from Rumelhart, Hinton and Williams. The idea has successfully been introduced to statistical language modeling (Bengio, Ducharme, Vincent & Jauvin, 2003) and further employed in variety of NLP tasks, including speech recognition and machine translation (Mikolov et al, 2013). Researchers have developed algorithms to make such calculation more efficient. Recent achievements by Mikolov et al. (2013) in 'Efficient Estimation of Word Representations in Vector Space' (Mikolov, Chen, Corrado & Dean, 2013) enabled learning distributed vector representations that capture highly accurate syntactic and semantic word relationships, with extreme efficiency.

Word2Vec refers to the specific algorithms used to produce word embeddings that were created by a team of researchers led by Tomas Mikolov of Google. Word2Vec has a relatively simple structure that allows for the efficient computation of high dimensional word vectors from a large corpus of text data. The word vectors are located in vector space, and words that share common features in the corpus are located close to each other. When measured by word similarity task, Word2Vec manifests significantly better performance in both accuracy and computational efficiency than earlier algorithms, such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).

One interesting aspect of word embedding is its strong capability of analogy reasoning, discussed in research by Mikolov et al (2013). Word embedding captures both syntactic and semantic regularities surprisingly well, and the relationships between words are presented by vector-offset. In embedding space, the same type of relationship shares the same offset. Since the space is high-dimensional, a single word can embed multiple relations. By using the embedding, analogy questions can be solved by linear format vector calculation. For instance, to answer the question 'What is King minus Man plus Woman?', the vector calculation vec("King") – vec("Man") + vec ("Woman") can be used. Although it is not possible to find a word located in the exact position, cosine similarity is used to find the word located the closest with the result of vector calculation. In the paper, word embedding demonstrates 40% accuracy in answering analogy questions.



a) Gender relation b) Singular/plural relation

Figure 2. Vector offsets for three word pairs illustrating the relation. (Adapted from Mikolove, Yih & Zweig, 2013)

2.4.3 The Relevance of Word2Vec in Industry Classification Scheme

Below there is a quote from finance news article by Fontana (2018). This quote exemplifies how a company's co-appearance in news articles can capture a variety of relationships such as joint ventures, mergers and acquisitions and various relatedness: economic, product, market and so on. Word2Vec is capable of capturing these relationships from text data and expressing them as multi-dimensional vectors, a quantifiable format making it possible to apply statistical analysis. Since the purpose of this research is to examine the possibilities of constructing an automated, well performing industry classification scheme that captures various relationships of companies by using unstructured data, Word2Vec is considered a suitable tool.

Given the size and strength of tech giants, Facebook Inc. (FB - Get Report), Apple Inc. (AAPL - Get Report), Amazon.com Inc. (AMZN - Get Report), Netflix Inc. (NFLX - Get Report) and Alphabet Inc. (GOOGL - Get Report) (Apple is often added, making the group the FAANG stocks), sometimes a bad week for the group of stocks can lead the market down. Several experts referred to the power of these stocks as "double-edged sword" for the market.

For instance, Facebook led a steep decline in the FAANG stocks in March, after its data misuse scandal came to light on Sunday, March 18. By Friday, March 23, Facebook's stock had dropped 13.5% since the previous Friday, losing more than $70 billion in market cap. Alphabet shares fell 9.5% during that five-day period, while Apple fell 7.3%, Netflix fell 5.5% and Amazon dropped 4.7%. The rest of the market fell with them, with the S&P 500 dropping about 6%, the Dow dropping 5.7% and the Nasdaq falling 6.5%.

Why the FANG Stocks' Dominance May Not Be So Bad for the Market
The tech giants bring added volatility to the market -- but that's not necessarily a bad thing.
Francesca Fontana
May 27, 2018 9:34 AM EDT

("Why the FANG Stocks' Dominance May Not Be So Bad for the Market," 2018)

2.5 Earlier Research in Text-Based Industry Classification

In this section, I cover earlier studies that attempt to utilize text data to discover the relationships between companies.

Shi, Lee & Whinston (2016) aimed to propose a new measure of firms' business proximity by utilizing a text mining technique. They have used unstructured text data collected from Crunchbase, an open, free database of tech companies, people and investors, regarded as the Wikipedia of the venture industry. The database retrieves high-tech-related information automatically from various news sources.

The research used a text mining technique called 'topic modeling' to build their model. This technique is based on the LDA algorithm, and it represents each firm's text description as a probabilistic distribution of a group of underlying topics (interpretable as industries in the case of this research). Next, it calculates business proximity between a pair of firms by calculating cosine similarity of the two corresponding probabilistic topic distributions. To evaluate, it employs Exponential Random Graph Models (ERGMs) as a modeling framework to examine which subset, among all pairs of companies, has a higher chance of engaging in Mergers and Acquisition (M&A) transactions. As a result, they show that the business proximity exhibited by their measure is strongly associated with the likelihood of M&A matching.

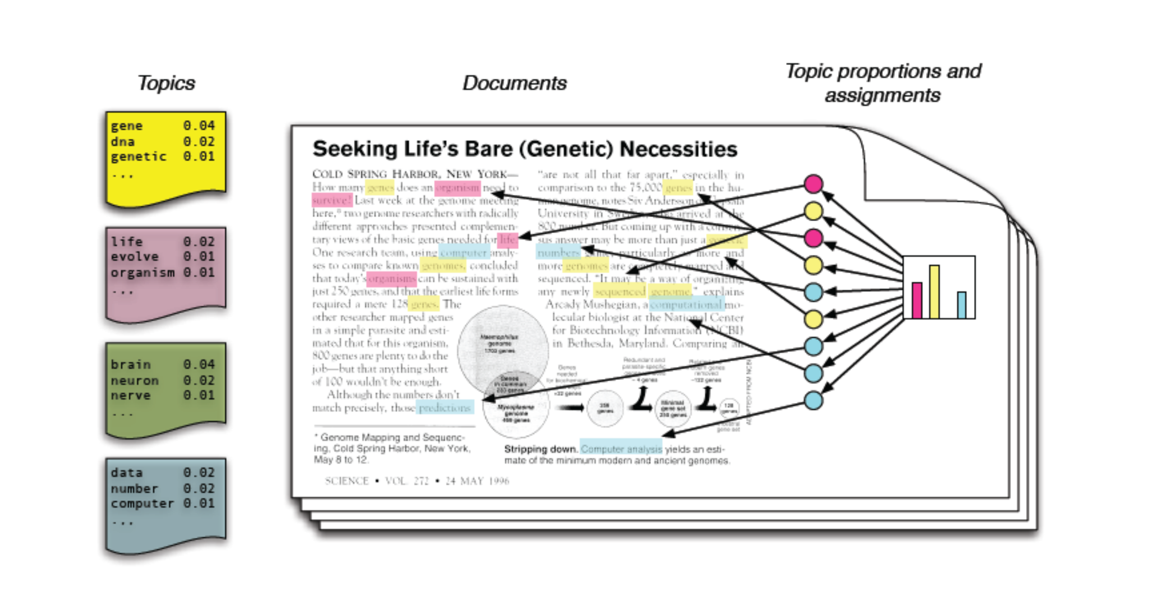


Figure 3. The intuitions behind LDA (Adapted from Blei, 2012)

Another example of earlier research that applies text mining methodology in company grouping comes from Bernstein, Clearwater & Provost, whose paper aims to develop a model that classifies companies by taking knowledge from existing industry classification schemes. Their source of data is a collection of news stories from 1999 to 2002. In building a model, they used a relational vector-space (RVS) model, which abstracts the linked structure by representing companies by vectors of weights. The linkages are based on the co-occurrence of firms in business news articles, and the strength of linkages is measured by the frequency of those co-occurrences. They evaluated the model based on ROC analysis and discovered that suggested classification procedures can be effective, and that classification performance correlates with the relational autocorrelation of the data set.

This thesis shares objectives with above studies at a broad level; all are interested in measuring the proximity between companies by utilizing text data and aim to utilize that information in a business context. To my knowledge, however, there has been no research that aims to identify the proximity between firms based on Word2Vec, the text mining algorithm developed relatively recently. Further, no research has attempted to evaluate the model’s validity in industry classification based on financial characteristics, the most widespread application of industry classification schemes.

3 Methods and Data

This section consists of the research methodology and shows how text-based industry classification schemes are constructed. I introduce two different schemes, one built using cosine similarity and the other built by employing cluster analysis. A brief introduction of mathematical concepts is presented together with the scheme design process.

3.1 Data

Data used in this thesis was collected from three different, publically available sources. The extraction was conducted with an open source tool, such as python library. In this section, I discuss why, how and what was extracted as data for the research.

List of S&P 500 companies with GICS classification

To answer research question, data about companies that will be classified into industry groups is necessary. To build an industry classification scheme, I used companies that belong to the S&P 500 list, which refers to companies that belong to the S&P 500 stock market index. The index is maintained by S&P Down Jones Indices and consists of 505 stocks issued by 500 large-cap business entities. The index is traded on the America stock exchange and covers approximately 80% of the American equity market. Companies that belong to this index are updated regularly based on rules governed by S&P Dow Jones Indices.

The data was extracted on May 7, 2018 from Wikipedia website ("List of S&P 500 companies," 2018). Table 2 was extracted using a python script that scraps table from an html page. The table consists of nine variables, of which four are used in the analysis (ticker symbol, security, GICS sector, GICS sub-industry).

Table 2:Form of S&P 500 company data extracted from wikipedia

Ticker symbol	Security	SEC filings	GICS Sector	GICS Sub-Industry	Address of Headquarters	Date first added[3][4]	CIK
MMM	3M Company	reports	Industrials	Industrial Conglomerates	St. Paul, Minnesota		66740
ABT	Abbott Laboratories	reports	Health Care	Health Care Equipment	North Chicago, Illinois	31.3.1964	1800
ABBV	AbbVie Inc.	reports	Health Care	Pharmaceuticals	North Chicago, Illinois	31.12.2012	1551152
ACN	Accenture plc	reports	Information Technology	IT Consulting & Other Services	Dublin, Ireland	6.7.2011	1467373
ATVI	Activision Blizzard	reports	Information Technology	Home Entertainment Software	Santa Monica, California	31.8.2015	718877
AYI	Acuity Brands Inc	reports	Industrials	Electrical Components & Equipment	Atlanta, Georgia	3.5.2016	1144215
ADBE	Adobe Systems Inc	reports	Information Technology	Application Software	San Jose, California	5.5.1997	796343
AMD	Advanced Micro Devices Inc	reports	Information Technology	Semiconductors	Sunnyvale, California	20.3.2017	2488

Word-embedding vectors

Word-embedding vectors allow us to utilize information about how companies appear in news articles. As mentioned in Chapter 2, I used Word2Vec, a recently developed tool to build word vectors from a text dataset. Fortunately, Word2Vec is part of Google's open source project that is publicly available for research purposes ("Google Code Archive," 2018). Researchers can download the library and build vectors out of the text data of interest. In addition to the tool, pre-trained word vectors are published online for public usage. The word vectors are trained using part of the Google News dataset (about 100 billion words), which contains 300-dimensional vectors for 3 million words and phrases.

In this thesis, I used pre-trained word vectors for the analysis. Using python script, I downloaded the model and extracted word-embedding vectors for each company and industry category. If the company name contained more than one word, I averaged the vectors of each word. Of 500 companies, 480 company word vectors were extracted. Among industry categories, all 11 main categories and 122 GICS sub-categories were identified and extracted.

Financial measures

To assess the industry classification scheme in explaining the economic relatedness of companies belonging to same industry group, data containing the financial measures of S&P 500 companies is required. For each company, this information includes stock return and valuation measures (book value, enterprise value divided by EBITDA, price to earnings, price divided by earning to growth and price to book value) selected as financial measures for the analysis.

Each company's up-to-date financial metrics are publicly available online by service providers such as Yahoo! Finance. To gather the necessary figures for each company, I used an open source python library called Beautiful Soup, which enables pulling data from HTML or XML files. Data was extracted on May 8, 2018. Not all financial measures were available for all 500 companies. For each financial measure (stock return, book value, enterprise to EBITDA, price to earnings, price divided by earning to growth and price to book value), 35, 26, 59, 18, 18 and 45 values were missing respectively.

3.2 Methods for Building Text-Based Industry Classification

This chapter discusses the research methodology used to answer the first research question defined in Chapter 1:

1. Can we build an automated industry classification scheme by employing word-embedding vectors derived from news articles?

An industry classification scheme aims to classify individual companies into a smaller number of groups so that members belonging to the same industry group are similar to each other. In this thesis, we attempt to build a classification scheme based on the assumption that company similarity can be represented as similarity between word-embedding vectors. To reflect the quantitative, measurable nature of word-embedding vectors, two different methodologies were used to build an industry classification scheme.

3.2.1 Scheme Based on Cosine Similarity

The first scheme was built by employing cosine similarity, the standard proximity measure in vector space modeling. In a word-embedding vector model, the word is represented as a vector in high-dimensional space, and the similarity between two words is measured by the similarity between the vectors. In general, the angle between two vectors is utilized as a measure of divergence, and the cosine of the angle is used as the numeric proximity. Cosine similarity is a useful characteristic as a measurement criterion because it is 1.0 for identical vectors and 0.0 for orthogonal vectors (Singhal, 2001).

The cosine similarity is calculated as a normalized dot product of the two vectors, and the normalization is usually Euclidean. For given vectors a and b, the cosine similarity measure between the two is calculated as follows (Sidorov, Gelbukh, Gómez-Adorno & Pinto, 2014):

The dot product is calculated as

$$a \cdot b = \sum_{i=1}^{N} a_i b_i \tag{1}$$

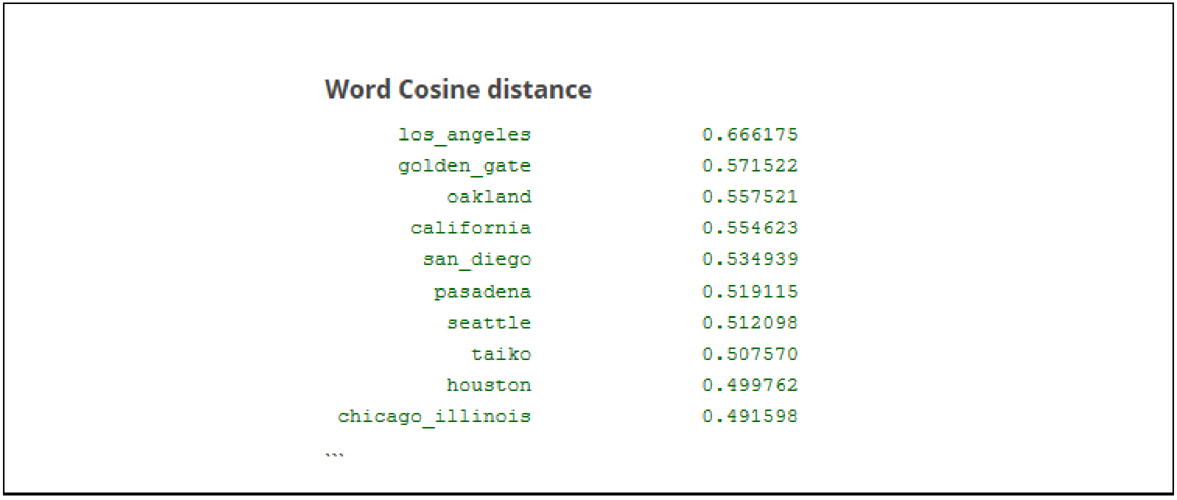
The norm is defined as

$$\|x\| = \sqrt{x \cdot x} \tag{2}$$

And the cosine similarity measure is calculated as

$$cosine(a, b) = \frac{a \cdot b}{\|a\| \llbracket b \rrbracket} \tag{3}$$

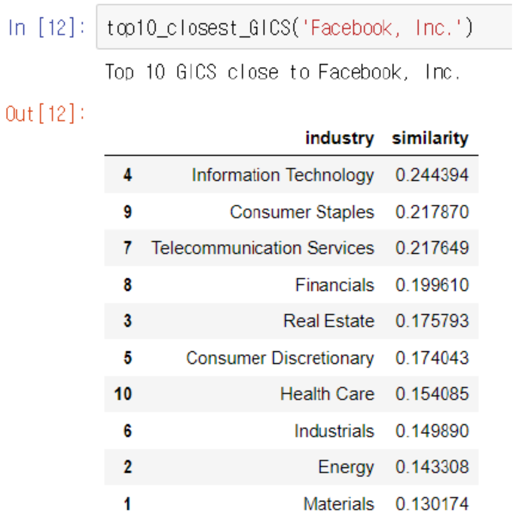
The objective of the thesis is to generate an automated industry classification scheme by using text analytics methodology. Industry classification aims to group companies based on similar products, services or similar behavior in financial markets. Thus, I have developed an industry classification scheme that uses cosine similarity to reflect the similar properties used in a typical industry classification. In Word2Vec applications, a high cosine similarity implies strong semantic similarity between two words, as shown in the figure 4, which is an example script from Google that outputs the words closest to “San Francisco.”



~~~
Word Cosine distance
         los_angeles              0.666175
         golden_gate              0.571522
             oakland              0.557521
          california              0.554623
           san_diego              0.534939
            pasadena              0.519115
             seattle              0.512098
               taiko              0.507570
             houston              0.499762
     chicago_illinois             0.491598
```
~~~

Figure 4. Word2Vec output: list of words close to “San Francisco”

In reflection of the above quality, I calculated cosine similarity between each company and word that describes industry. For each company, I categorized its industry group with the one that manifested the highest cosine similarity between words. For instance, the figure 5 shows the output of the list of industries with the highest cosine similarity with the company Facebook, Inc. In this case, Facebook, Inc. is classified in the “Information Technology” industry group.



```
In [12]: top10_closest_GICS('Facebook, Inc.')

         Top 10 GICS close to Facebook, Inc.
Out[12]:
```

	industry	similarity
4	Information Technology	0.244394
9	Consumer Staples	0.217870
7	Telecommunication Services	0.217649
8	Financials	0.199610
3	Real Estate	0.175793
5	Consumer Discretionary	0.174043
10	Health Care	0.154085
6	Industrials	0.149890
2	Energy	0.143308
1	Materials	0.130174

Figure 5. Word2Vec output: List of industry words close to "Facebook, Inc."

The words that describe industry are taken from the GICS categorization, which was extracted together with the S&P 500 companies' data; there are 11 main categories and 122 sub-categories. I used both to build an industry classification scheme. The industry category words used for this analysis are presented in the appendix.

3.2.2 Scheme Based on Cluster Analysis

The second scheme was constructed by applying cluster analysis, which refers to the task of dividing data into groups that are either meaningful or useful. Dividing objects into meaningful groups or classes that share common properties plays a critical role in analyzing and describing a given situation. Cluster analysis has long served an important role in various fields, including social sciences such as psychology, biology, statistics, information retrieval, pattern recognition, data mining and machine learning.

In cluster analysis, objects in a data set are grouped based only on information included in the data, which describes the objects and the relationships between them. The objective is to build clusters so that objects within the same group are similar to one another and different from the objects in other groups: the greater the homogeneity within a group and the greater the difference between groups, the better the clustering.

Cluster analysis is suitable to build an industry classification scheme due to its ability to sort data objects into meaningful groups. Since our data (word embedding vectors) contains information about each data object, we can use it as a basis for clustering. In addition, by using unsupervised machine learning techniques (the task of finding hidden structure from a given dataset), it is possible to discover hidden patterns within the data.

To build this industry classification scheme, I experimented with three different clustering techniques: K-means, agglomerative hierarchical clustering and DBSCAN. **K-means** is partitional clustering, a division of data objects into non-overlapping clusters, that attempts to find a user-specified number of clusters. **Agglomerative hierarchical clustering** is a technique that generates a set of nested clusters organized as a tree. This technique produces clusters by beginning with each data object as a single cluster and repeatedly merging the two nearest clusters until only a single, large cluster remains. **DBSCAN** is a density-based clustering algorithm in which a cluster is a dense region of objects surrounded by a low-density region. Unlike other algorithms, DBSCAN automatically determines the number of clusters. For K-means and agglomerative hierarchical clustering, different numbers of clusters (from two to 100) are generated and evaluated to determine a suitable number of clusters.

As input data for the clustering algorithm, I used two different formats for each company: vector representation and a cosine similarity matrix. The former contains a list of vectors extracted from the Google News data set, which expresses each company as 300 dimensional vectors. Thus, the data forms a matrix of 500 by 300 dimensions. The second format, the cosine similarity matrix for each company, is constructed by pairwise calculation of the cosine similarity for each individual company. The latter forms a matrix of 500 by 500 dimensions.

3.3 Methods for Assessing Text-Based Industry Classification

Assessing the performance of the scheme is significant to evaluate whether text-based industry classification schemes bear any advantages compared to existing industry classification schemes. This section covers the methodology used in answering the second research question:

2. If yes, what are the benefits of the scheme in comparison with existing industry classification schemes?

The assessment of an industry classification scheme consists of two perspectives: qualitative and quantitative. Qualitative assessment aims to evaluate whether text-based industry classification schemes are capable of grouping companies into qualitatively homogenous industry groups. The second assessment is conducted by applying quantitative analysis to evaluation measures.

3.3.1 Qualitative Assessment

The qualitative assessment of the constructed industry classification scheme has a high level of complexity. First, it is difficult to determine each company's industry group, especially when the basis of judgment is obscure. As discussed in the literature review, the traditional approach to industry classification has limitations. Studies have suggested different approaches to industry classification, but none of the methodologies is accepted as a worldwide standard. Next, examining the validity of industry classification involves the individual investigation of each company. Our dataset contains 500 companies, a challenging amount for individual investigation.

Thus, I took two approaches to evaluating the scheme. I conducted in-depth investigation for a few sample companies. For those samples, I checked the results of the industry classification scheme and compared them with the product and market information of the sample companies. To estimate overall performance, I conducted statistical analysis in comparison with the GICS industry classification. Although this thesis builds on the assumption that existing classification systems entail several limitations, the comparison can provide an idea of how much the proposed scheme deviates from traditional schemes and where that deviation originates.

The scheme based on cluster analysis does not include information about which industry each company belongs to; it only indicates which companies are close to each other and the best way of grouping them. Thus, I described each cluster in terms of industry by applying same methodology used in the first scheme. For each cluster, I calculated the average vector for all the vectors included in the cluster. Next, I calculated the cosine similarity for each average vector with industry word vectors and determined the closest industry word vectors to each cluster. As a result, each cluster describes the industry aspects of companies belonging to the cluster group.

3.3.2 Quantitative Assessment

As mentioned above, an industry classification scheme is expected to describe similarities between companies belonging to the same industry group in terms of financial measures. To measure similarities, I calculated two different evaluation metrics.

1. The median of the standard deviation within each industry group (Weiner, 2005)

Standard deviation (SD, or σ) is used to measure the variation or dispersion of a data set in statistics. For every company belonging to same industry group, the standard deviation of each financial metric is calculated. The standard deviation represents the variation of financial measures for the companies included in same industry group. Next, the median of the standard deviation from every industry is calculated to compare industry classification schemes.

2. The average adjusted R squared for OLS regression (Bhojraj, Lee & Oler, 2003)

The second metric is used to measure a scheme's ability to explain firm-level financial measures with industry-level equivalents. The dependent variable y is the financial metric for firm i. The independent variable, xi, is the average financial metric for all companies in that industry classification.

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{4}$$

R squared refers to the proportion of variance in the dependent variable, which is predictable from the independent variable. The adjusted R squared is a modified version of R squared to remove the effect of the unnecessary increase of R when extra explanatory variables are added. A perfect fit results in a 1, while 0 indicates no explanatory power of the independent variable.

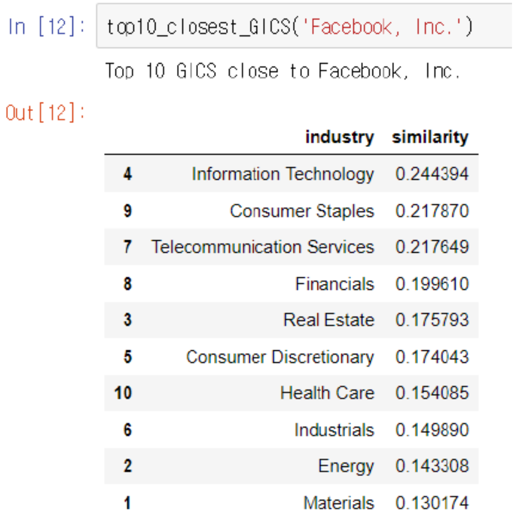
4 Results and Discussion

4.1 Text-Based Industry Classification Scheme

In Chapter 3, two different approaches to building industry classification schemes are presented. Although, technically, the possibility of building an automated, text-based industry classification scheme was confirmed, we cannot answer the first research question without qualitative assessment of the resulting schemes. As discussed in the methodology section, two different approaches were used. This section presents the results of those approaches.

Scheme 1, based on cosine similarity

First, qualitative assessment was conducted for several sample companies. I created the script so that the output exhibits 10 industry categories ordered by cosine similarity for the given input, company names. Figure 6 shows the output of the script for the sample companies.



```
In [12]: top10_closest_GICS('Facebook, Inc.')
         Top 10 GICS close to Facebook, Inc.
Out[12]:
```

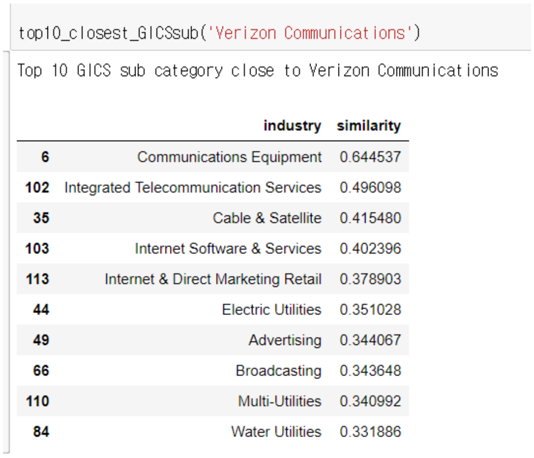
	industry	similarity
4	Information Technology	0.244394
9	Consumer Staples	0.217870
7	Telecommunication Services	0.217649
8	Financials	0.199610
3	Real Estate	0.175793
5	Consumer Discretionary	0.174043
10	Health Care	0.154085
6	Industrials	0.149890
2	Energy	0.143308
1	Materials	0.130174

Figure 6. Word2Vec output: List of industry words close to “Facebook, Inc.”

Facebook is an online social media website that provides a social networking service. Thus, out of 11 GICS categories, information technology is the most suitable industry classification for Facebook. This classification coincides with the GICS classification.

Verizon Communications is a multinational telecommunications conglomerate that has products such as mobile, broadband, digital television and IPTV. Thus, telecommunication services is the most suitable categorization, which is consistent with the GICS classification.

Additional experiments were conducted with GICS sub-categories that contain 122 sectors. Figure 7 is the result for Verizon Communications. The company has the highest cosine similarity with the communications equipment category, which refers to hardware used for telecommunications. In the GICS sub-categorization, it belongs to the category integrated telecommunication services, which is the second-closest industry based on this thesis's definition. Since Verizon now focuses on voice, data and video services and wireless/wireline network solutions, integrated telecommunication services seems to be the more suitable categorization (Verizon annual report, 2017).



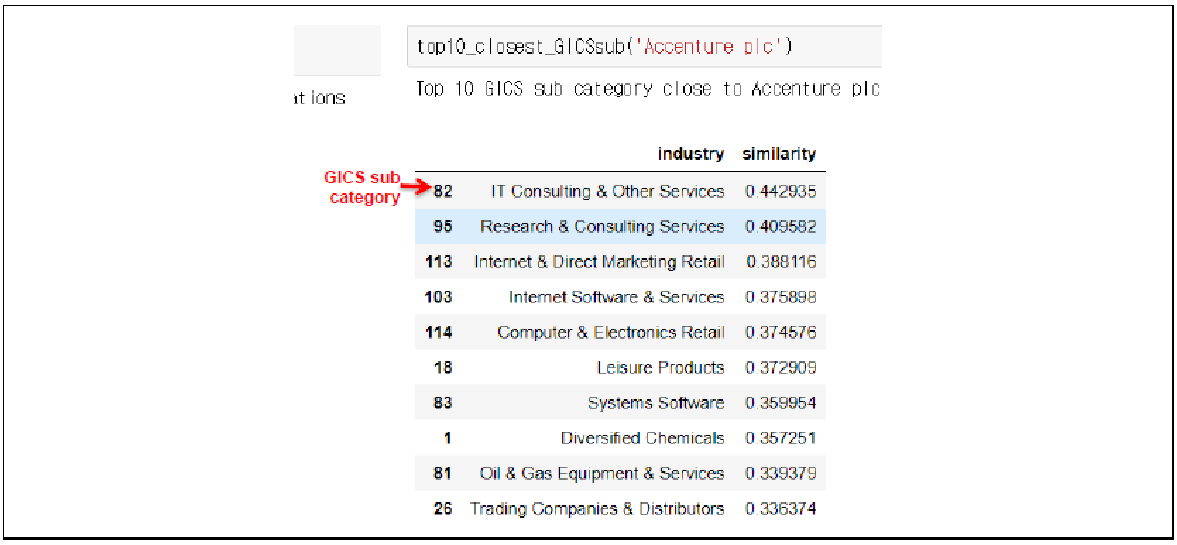
```
top10_closest_GICSsub('Verizon Communications')
```

```
Top 10 GICS sub category close to Verizon Communications
```

	industry	similarity
6	Communications Equipment	0.644537
102	Integrated Telecommunication Services	0.496098
35	Cable & Satellite	0.415480
103	Internet Software & Services	0.402396
113	Internet & Direct Marketing Retail	0.378903
44	Electric Utilities	0.351028
49	Advertising	0.344067
66	Broadcasting	0.343648
110	Multi-Utilities	0.340992
84	Water Utilities	0.331886

Figure 7. Word2Vec output: List of industry words close to "Veirzon Communications"

Figure 8 is the result of Accenture PLC. Accenture is a management consulting and professional service firm that offers strategy, consulting, technology, digital and operations services (Accenture annual report, 2017). Considering that a large part of its business focuses on information technology-related services, the classification result – IT Consulting & Other Services – seem to be suitable. The result accords with the GICS classification.



```
top10_closest_GICSsub('Accenture plc')
```

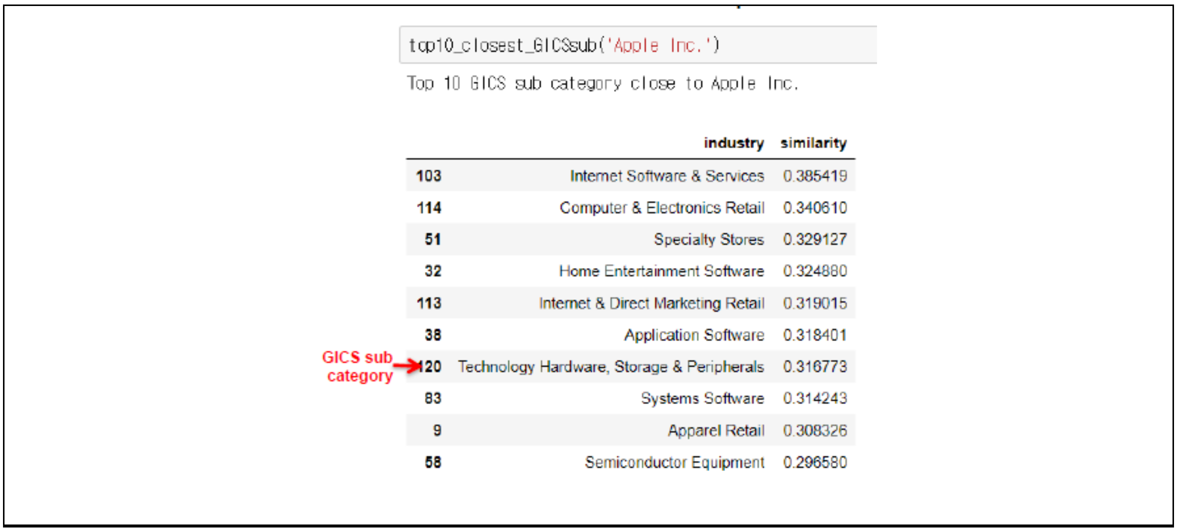
```
Top 10 GICS sub category close to Accenture plc
```

GICS sub category

	industry	similarity
82	IT Consulting & Other Services	0.442935
95	Research & Consulting Services	0.409582
113	Internet & Direct Marketing Retail	0.388116
103	Internet Software & Services	0.375898
114	Computer & Electronics Retail	0.374576
18	Leisure Products	0.372909
83	Systems Software	0.359954
1	Diversified Chemicals	0.357251
81	Oil & Gas Equipment & Services	0.339379
26	Trading Companies & Distributors	0.336374

Figure 8. Word2Vec output: List of industry words close to 'Accenture plc'

Figure 9 shows the results for Apple Inc. Apple is a technology company that develops and sells consumer electronics, computer software and online services. Apple's hardware products include smartphones, laptops, smart watches and tablet computers. It sells software products such as operation systems, media players, web browsers and other professional applications. Thus, the company can be classified in both hardware and software industries if both categories co-exist. The cosine similarity-based classification result shows that the closest industry is Internet software and services, followed by other relevant categories: computer and electronics retail, specialty stores and home entertainment software. The GICS sub-category classifies Apple as technology hardware, storage and peripherals.



```
top10_closest_GICSsub('Apple Inc.')
```

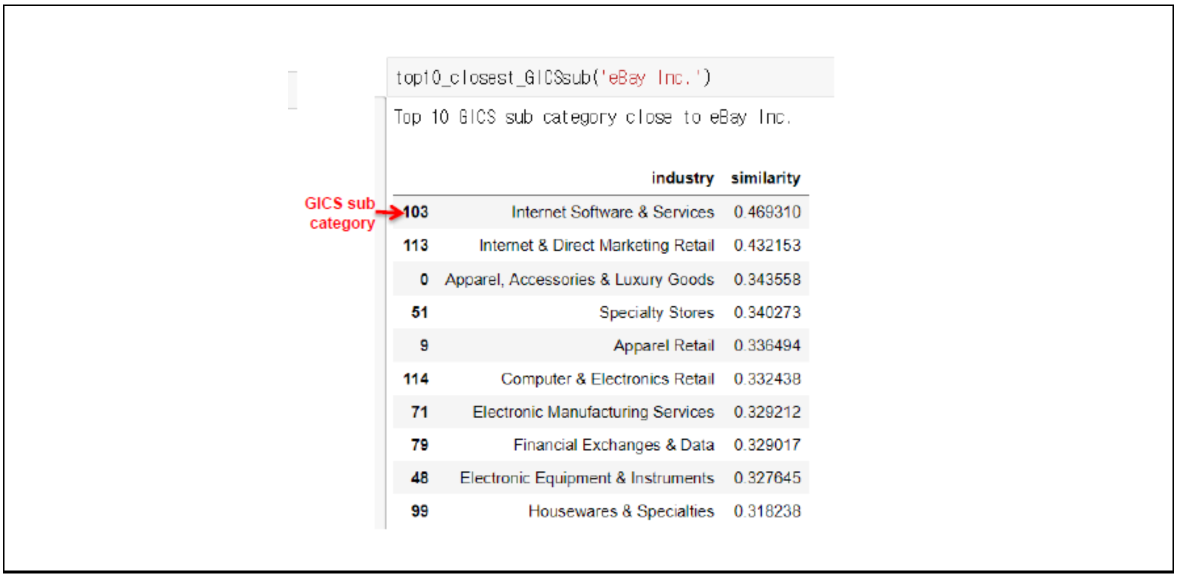
```
Top 10 GICS sub category close to Apple Inc.
```

	industry	similarity
103	Internet Software & Services	0.385419
114	Computer & Electronics Retail	0.340610
51	Specialty Stores	0.329127
32	Home Entertainment Software	0.324880
113	Internet & Direct Marketing Retail	0.319015
38	Application Software	0.318401
120	Technology Hardware, Storage & Peripherals	0.316773
83	Systems Software	0.314243
9	Apparel Retail	0.308326
58	Semiconductor Equipment	0.296580

GICS sub category

Figure 9. Word2Vec output: List of industry words close to "Apple Inc."

The last example is eBay Inc. eBay is an e-commerce company that facilitates business-to-consumer or consumer-to-consumer sales through its website. Thus, the most suitable classification of eBay is Internet software and services, which was selected by both our scheme and the GICS sub-categorization. The next closest industries (Internet and direct marketing retail, apparel, accessories and luxury goods, specialty stores) are interesting since they describe eBay in its market perspective, while the first category, Internet software and services, classifies the company in terms of the products and services it offers.



```
top10_closest_GICSsub('eBay Inc.')
```

```
Top 10 GICS sub category close to eBay Inc.
```

	industry	similarity
103	Internet Software & Services	0.469310
113	Internet & Direct Marketing Retail	0.432153
0	Apparel, Accessories & Luxury Goods	0.343558
51	Specialty Stores	0.340273
9	Apparel Retail	0.336494
114	Computer & Electronics Retail	0.332438
71	Electronic Manufacturing Services	0.329212
79	Financial Exchanges & Data	0.329017
48	Electronic Equipment & Instruments	0.327645
99	Housewares & Specialties	0.318238

GICS sub category

Figure 10. Word2Vec output: List of industry words close to "eBay Inc."

Table 3 shows the concordance between cosine similarity-based industry classification schemes with the GICS classification. In some cases, there was a perfect match (e.g., telecommunication services). While the energy and materials industry has a certain level of matching (72% and 64%), the rest have less than a 50% match. Overall, 43% of companies are classified in the same industry group in both schemes and the level of concordance varies at a high level.

Table 3: Concordance between cosine similarity-based industry classification and GICS classification

Cosine similarity-based scheme

GICS classification

	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	IT	Materials	Real Estate	Telecommunication	Utilities	Total	**Co-occurrence**
Consumer Discretionary	**1**	25	1	5	4	2	8	7	16	9	2	80	1%
Consumer Staples		**20**	2	1	4		1		4		1	33	61%
Energy		2	**21**				1	1		3	1	29	72%
Financials	1	10	7	**22**	1	1	6		1	9	3	61	36%
Health Care		12	1	11	**16**	3	6	7	1	2	1	60	27%
Industrials	1	6	4	2		**3**	5	16	1	20	6	64	5%
IT		15		7			**22**	14	2	10		70	31%
Materials		1	2	2		1		**16**	1	1	1	25	64%
Real Estate	1	1	4	3	2	1	2	1	**11**	6		32	34%
Telecommunication Services										**2**		2	100%
Utilities			14					1			**9**	24	38%
Total	**4**	**92**	**56**	**53**	**27**	**11**	**51**	**63**	**37**	**62**	**24**	**480**	43%

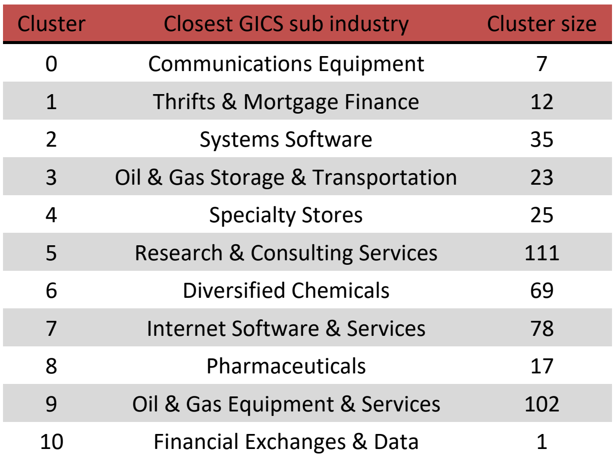
In summary, assigning an industry category based on cosine similarity shows a decent level of accuracy when the result is evaluated based on subjective judgment. If the GICS subcategorization is used, the industry classification can be identified at a more specific level. When compared with its concordance with the GICS classification, scheme 1 has a 43% correspondence with high variance depending on the industry.

Scheme 2, based on cluster analysis

As mentioned above, scheme 2 was built using cluster analysis. Each cluster represents industries and companies that belong to same cluster and are identified as belonging to the same industry group. Unlike scheme 1, the industry scheme based on cluster analysis does not provide information about the business aspects of each company. Thus, the average vector for each cluster is calculated and the closest industry category is identified to determine the industry aspect of each cluster. In industry identification, the GICS subcategory is used due to overlapping industry categorizations in cases where the GICS main categories are used.

Table 4 is the result of K-mean clustering with 11 clusters (the same number as the GICS main categorization). Although I experimented with other algorithms (hierarchical clustering, DBSCAN), I present the results of K-mean clustering this section.

Table 4: Result of K-mean clustering – Cluster distribution



Cluster	Closest GICS sub industry	Cluster size
0	Communications Equipment	7
1	Thrifts & Mortgage Finance	12
2	Systems Software	35
3	Oil & Gas Storage & Transportation	23
4	Specialty Stores	25
5	Research & Consulting Services	111
6	Diversified Chemicals	69
7	Internet Software & Services	78
8	Pharmaceuticals	17
9	Oil & Gas Equipment & Services	102
10	Financial Exchanges & Data	1

Table 5 presents the examples of lists of companies belonging to same industry. The overall accuracy is fine, although not all classifications are plausible.

Table 5:Result of K-mean clustering – Company Classification

Communications Equipment	Thrifts & Mortgage Finance	Systems Software	Specialty Stores
Charter Communications	Ameriprise Financial	Cadence Design Systems	JM Smucker
Dish Network	BlackRock	Cisco Systems	Kraft Heinz Co
F5 Networks	Fifth Third Bancorp	Citrix Systems	Kroger Co.
Juniper Networks	Huntington Bancshares	Cognizant Technology Solutions	Molson Coors Brewing Company
L-3 Communications Holdings	KeyCorp	DXC Technology	Monster Beverage
SBA Communications	Kimco Realty	Electronic Arts	Newell Brands
Verizon Communications	Macerich	FLIR Systems	Nike
	Prologis	General Dynamics	Nordstrom
	SunTrust Banks	Gilead Sciences	Perrigo
	U.S. Bancorp	IDEXX Laboratories	Ross Stores

With the results above, we can now address the first research question:

1. **Can we build an automated industry classification scheme by employing word-embedding vectors derived from news articles?**

Both schemes were built with python code that runs automatically based on input. Additionally, the quality of the resulting industry classification schemes is somewhat plausible based on subjective judgment. Thus, the research result presents a positive answer to the first research question.

4.2 Scheme Assessment

In the previous chapter, the results of the qualitative assessment were presented. In this chapter, I present the results of the quantitative assessment based on two different evaluation metrics discussed in Chapter 3.3.2.

Scheme 1, based on cosine similarity

For scheme 1, the result of the comparison with the GICS classification is presented in table 6. In comparison with the GICS system, scheme 1 has a greater median of standard deviation for five out of six financial measures (stock return, book value, enterprise value divided by EBITDA, price to earnings, price divided by earning to growth). For the adjusted R squared, scheme 1 has a smaller value for five out of six financial measures (stock return, book value, price to earnings, price divided by earning to growth). For both evaluation metrics, the result implies that the GICS system is better than the proposed cosine similarity-based scheme for most measures.

Table 6: The comparison between the GICS system and the cosine similarity-based scheme

Financial metrics	The median of standard deviation within each industry group		Adjusted R squared for OLS regression	
	GICS	Cos-similarity based	GICS	Cos-similarity based
52WeekChange	0.2191	0.2356	0.0963	0.0533
bookValue	18.7907	21.0303	0.1156	0.0373
enterpriseToEbitda	5.4231	9.2278	0.0162	0.0172
forwardPE	4.8932	12.0530	0.1440	0.0336
pegRatio	2.3266	3.1804	0.0694	0.0145
priceToBook	7.3042	5.4141	0.0166	0.0236

Scheme 2, based on cluster analysis

As mentioned in Chapter 3.2.2, various algorithms were tested to build clusters. I present the results of a total of eight different cases (four algorithms times two input data) with a range of clusters numbering from two to 100. The result of the DBSCAN is excluded because it only created unbalanced clusters.

With word vectors (480 by 300)

- K-means
- Agglomerative hierarchical clustering (linkage: average)
- Agglomerative hierarchical clustering (linkage: ward)
- Agglomerative hierarchical clustering (linkage: cosine)

With pairwise cosine similarity (480 by 480)

- K-means with pairwise cosine similarity
- Agglomerative hierarchical clustering (linkage: average)
- Agglomerative hierarchical clustering (linkage: ward)
- Agglomerative hierarchical clustering (linkage: cosine)

Figure 11 shows the results of the experiment in which the financial measure is the 52-week stock return. In the graph, the x-axis presents the number of clusters, and the y-axis presents the calculated evaluation metrics (median of standard deviation, average of adjusted R squared). Each graph shows the result of eight different cases. The black horizontal line refers to the value of evaluation metrics in the case of the GICS classification system. The red vertical line in the left graph represents the value of the median of standard deviation when no group is formed.

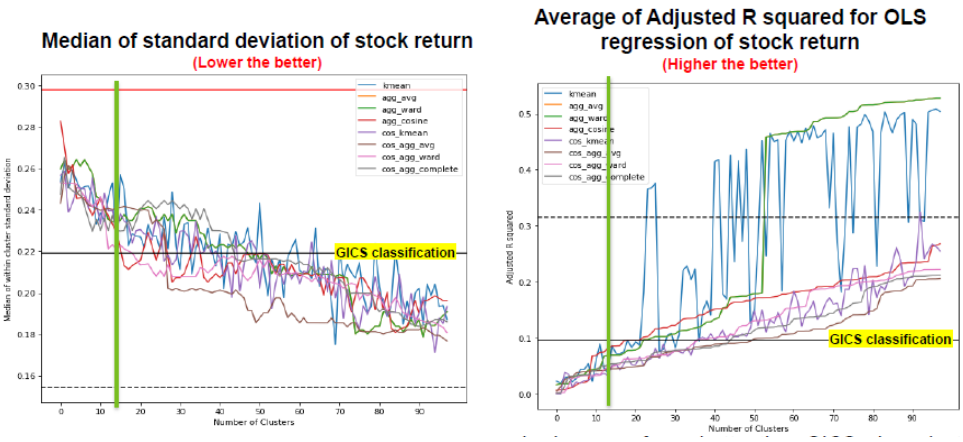


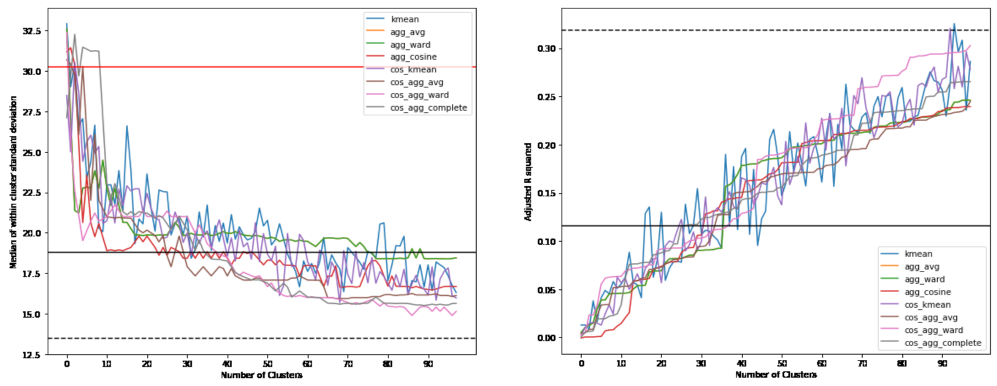
Figure 11. Result of evaluation metrics for cluster-based industry classification – in case of stock return

For both evaluation metrics, the overall performance of the schemes improves as the number of clusters increases. The result is expected since the growth in the number of clusters should be connected to greater homogeneity within the cluster. However, the improvement pattern varies significantly for each case. For instance, in the right figure, the cluster with a k-mean algorithm shows a high level of fluctuation, while other algorithms, such as agglomerative hierarchical clustering, show a steady increase in performance as the number of cluster increases.

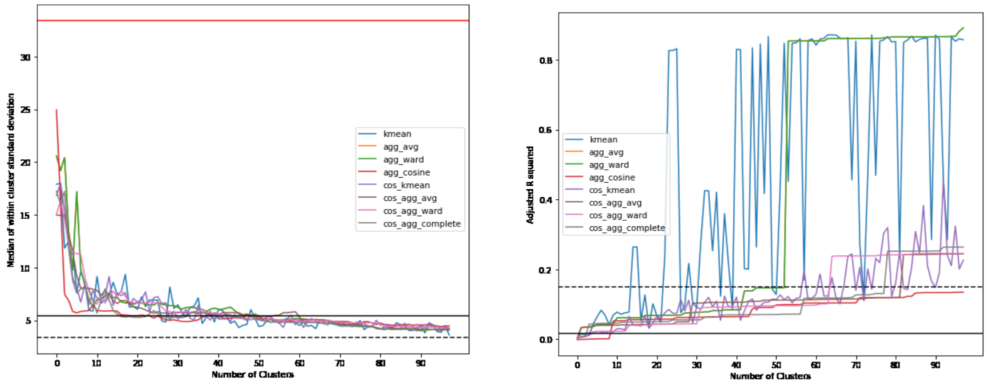
The green vertical line suggests the crossing point between the value of evaluation metrics in the GICS system and the equivalent in the cluster-based classification scheme. For the median of standard deviation, the scheme based on agglomerative hierarchical clustering with cosine linkage exhibits a lower value than the GICS system when the number of clusters is 13. For the average of adjusted R squared, the scheme with the same condition (agglomerative hierarchical clustering with cosine linkage) meets the performance of the GICS classification when the number of clusters is 12. This result implies that a text-based industry classification scheme can meet the performance of existing schemes with a similar number of clusters (the GICS classification system has 11 industries).

Figure 12 presents the results of the remaining five financial metrics. Similar to stock return, the performance of the cluster-based scheme increases as the number of cluster grows. However, the shape of the graph and the point at which the text-based scheme meets the GICS system varies.

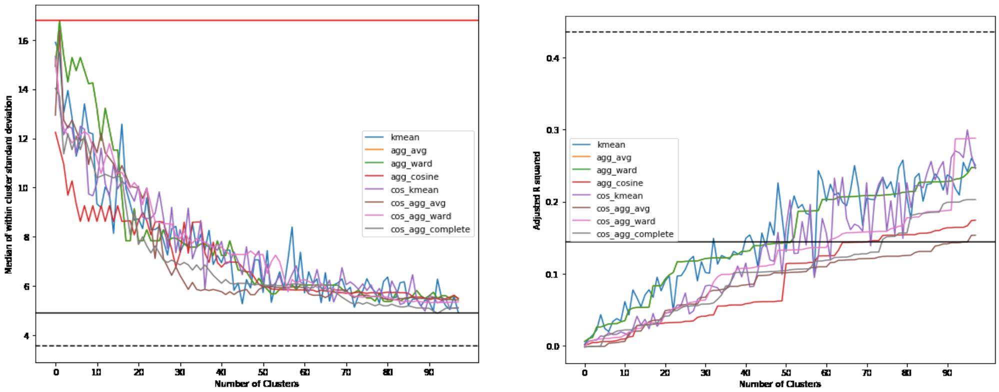
Book Value



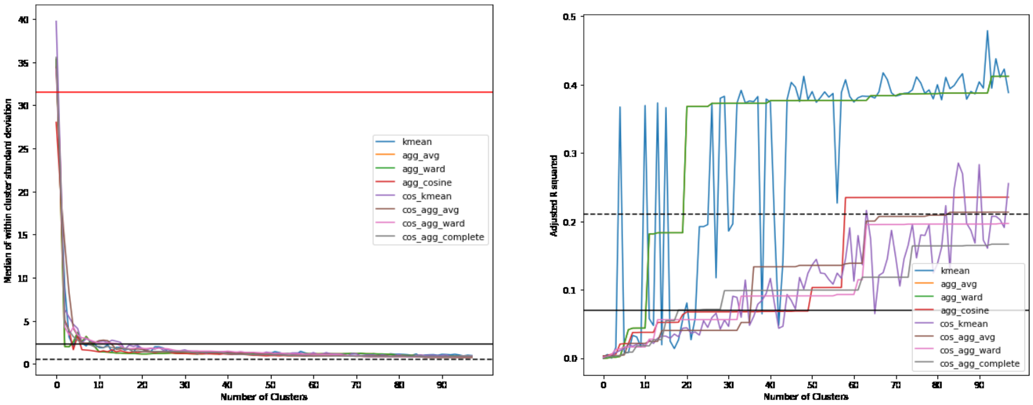
Enterprise Value divided by EBITDA



Price to Earnings



Price divided by Earning to Growth



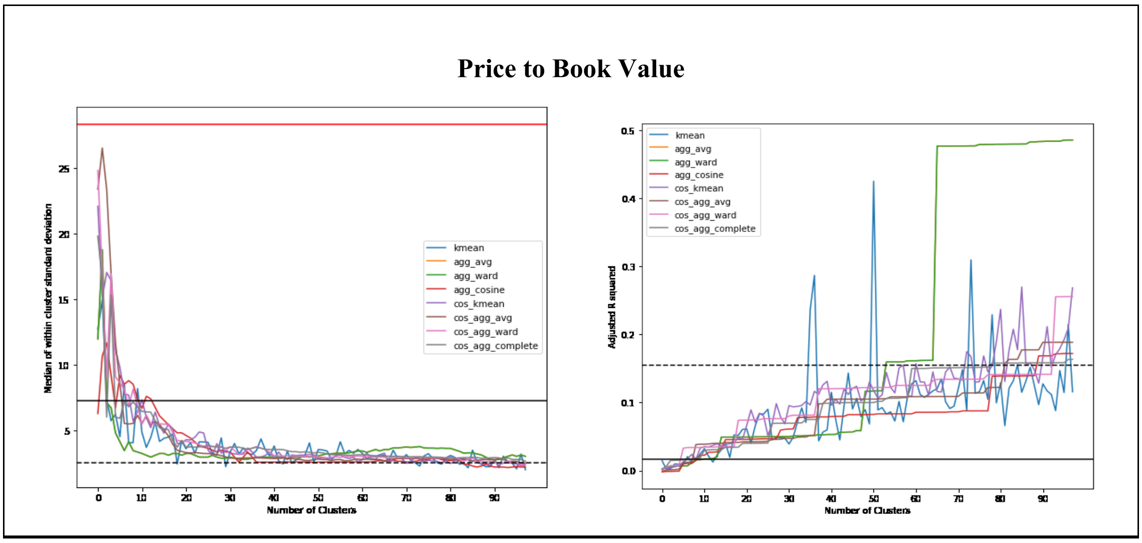


Figure 12. Result of evaluation metrices for cluster-based industry classification – in case of the rest financial metrices. (Left Row: The Median of Standard Devation, Right row: The Average Adjusted R squared)

4.3 Synthesis of Empirical Findings

The empirical findings suggest that it is technically possible to build an automated, text-based industry classification scheme. Two schemes are presented: one based on cosine similarity and the other based on cluster analysis. In addition to the technical possibility, the resulting text-based industry classification schemes were evaluated from both qualitative and quantitative perspectives. When evaluated based on subjective judgment, the text-based classification scheme manifests decent quality in classifying companies into homogenous groups. In quantitative measurement, evaluation metrics were used to estimate how well the scheme explains the economic relatedness between firms that belong to same industry group. The results suggest that the text-based scheme can manifest better performance when parameters are optimized.

Based on the findings, I now answer the second research question:

2. If yes, what are the benefits of the scheme in comparison with existing industry classification schemes?

Before the discussion, it is useful to review the limitations of existing industry classification schemes covered earlier. The limitations of existing schemes include:

1. A mismatch of industry classification of companies between different industry classification systems, which may cause confusion in analysis
2. An inability to adapt to constant technological change and innovation in a timely manner
3. A failure to address homogeneity between firms within the same industry group to a sufficient level, especially in terms of financial characteristics

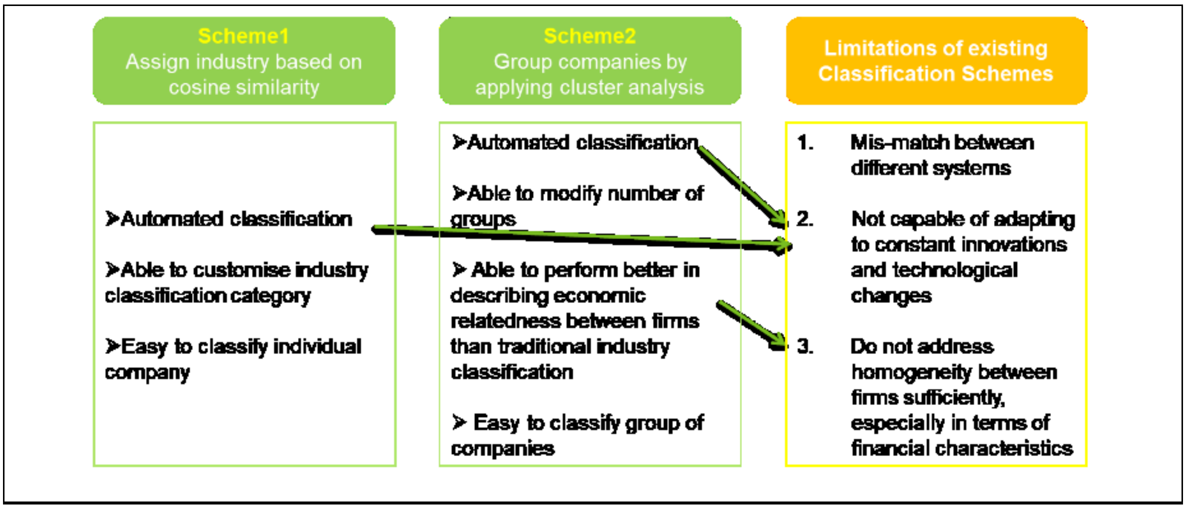
The text-based industry classification schemes can address two of these three points. First, both schemes are computer script-based and easy to maintain and update. Although this thesis used pre-trained vectors from the Google News data set, one can train vectors using Word2Vec. Depending on the purpose of the research, various types (with different topics, publication times and publishers) of text corpuses can be used in training. The ease of maintenance successfully addresses the second point, the current system's inherent difficulty in keeping up-to-date classification.

Next, scheme 2 shows the capabilities of better performance in explaining the economic relatedness of companies belonging to same industry group. As the number of clusters increases, the scheme's performance also increases. The well-performing scheme's attributes (number of clusters, cluster algorithm) vary depending on the financial measures of interest.

Aside from addressing the limitations of existing schemes, scheme 1 shows the ability to explain the business characteristics of the input company. By using simple code built with cosine similarity, the industry group of each company can be presented with a decent level of accuracy. By changing the industry category input, it is possible to create different levels of industry classification.

Scheme 2 provides great flexibility in terms of various attributes included in cluster analysis. By adjusting the cluster number, algorithm and deeper level attributes of cluster analysis, it can generate an industry classification suitable for one's research purposes. If the research has an optimal classification distribution, scheme 2 can be utilized to generate a suitable distribution.

Figure 13 presents a summary of the benefits of each text-based industry classification scheme. The green arrow represents the aspects of each scheme that address the limitations of the existing classification scheme.



Scheme1
Assign industry based on cosine similarity

- ➢Automated classification
- ➢Able to customise industry classification category
- ➢Easy to classify individual company

Scheme2
Group companies by applying cluster analysis

- ➢Automated classification
- ➢Able to modify number of groups
- ➢ Able to perform better in describing economic relatedness between firms than traditional industry classification
- ➢ Easy to classify group of companies

Limitations of existing Classification Schemes

1. Mis-match between different systems
2. Not capable of adapting to constant innovations and technological changes
3. Do not address homogeneity between firms sufficiently, especially in terms of financial characteristics

Figure 13. Benefits of text-based schemes that address the limitations of existing classification

5 Conclusion

5.1 Summary of Key Findings

The objective of this thesis was to propose an automated, text-based industry classification scheme that could address the limitations of existing industry classification schemes. The research problem was addressed by answering two specific research questions. To find answers, the latest text mining tool and suitable quantitative analysis methodologies were applied.

The first research question asks whether it is possible to build an automated, text-based industry classification scheme using word-embedding vectors extracted from news articles. By using quantitative aspects of word-embedding vectors, I built two different classification schemes. The first was built based on cosine similarity by assigning company words to industry categories based on the calculated proximity between the company-word vector and the industry-word vector. The second, meanwhile, was built with cluster analysis by applying a cluster analysis algorithm to company vectors to group the individual companies into a smaller number of groups.

Enabling the construction of the scheme does not create meaningful contribution unless proper assessment is made to measure the effectiveness of the proposed scheme, however. The second question asks whether a text-based scheme is beneficial in comparison to existing schemes. To answer the questions, both qualitative and quantitative assessments were conducted. In qualitative assessment based on subjective judgment, the text-based scheme manifests adequate performance in classifying companies into industry categories that are seemingly correct when judged based on existing information, such as a company's annual report, search results, etc.

In quantitative assessment, two metrics were used to measure the scheme's performance in explaining the economic relatedness of companies within the same industry group. Financial metrics such as stock return and company valuation measures were used in this calculation.

In comparison with the GICS system, the scheme based on cosine similarity manifested poor performance. However, the scheme based on cluster analysis suggests that it may perform better than the GICS system when the involved parameters are optimized.

Thus, the text-based scheme successfully addresses the two limitations identified in existing schemes. First, the script-based, automated scheme is more flexible in accommodating industry changes coming from innovation and technological development unlike static, existing schemes that require adapting many resources. Second, a text-based scheme can have superior performance in explaining the economic relatedness of companies that belong to the same industry group when the involved parameters are optimized.

5.2 Limitations of the Study

Although the study has reached its aims, there are still limitations. In this section, I discuss the inevitable limitations included in this study in terms of data, methodology and validity assessment.

The first limitation involves the scope of the research. In this study, I focused on companies belonging to the S&P 500 list, which consists of 500 large-cap companies traded on American stock exchanges. Although it covers 80% of the American market, the data scope is limited in that it is based on enterprises in only one country. In addition, the pre-trained word vectors used in this research are extracted from the Google News data set, and 300 dimensional vectors were built based on the skip-gram model. However, there are other modifiable parameters in training word vectors: model architecture, the dimension of the vectors, the subsampling rate and the size of training windows. By using a single pre-trained vector, the research is thus limited.

The second limitation occurred while building scheme 2 in the usage of cluster analysis. Although I experimented with different types of clustering algorithms and with cluster numbers, there are other adjustable variables for each cluster algorithm. For instance, a K-means algorithm contains various parameters such as initiation methods, number of initiation points, maximum number of iterations and tolerance. Because I did not experiment with those parameters, the research may have not fully discovered the performance of the algorithms included.

Next, the validity assessment process involves certain limitations. Unlike assessments based on quantitative financial measures, assessing classification schemes in terms of business explanations is highly dependent on subjective judgment. This derives from the complex nature of industry classification schemes, in which a multi-dimensional approach is required. Although a comparison with the GICS classification scheme was presented, the GICS scheme cannot be used as an absolute measure in qualitative assessment due to its inherent limitations, discussed in Chapter 2.

Last limitation comes from the selection of industry words in building the word embedding vector. The industry words, such as IT, real estate and communications, are derived from GICS categorization. Although these words are the official categories that are widely used in academic research, they hardly present in news articles, which I used in this study to construct the industry classification schemes. Thus, the official GICS industry categorization words may not be the most suitable words to capture the relationship between word embedding vectors.

5.3 Suggestions for Future Research

A number of areas would benefit from future research, including the ones I have developed here:

1. Test the scheme with various data settings
   As mentioned, the thesis focuses only on companies that belong to the S&P500 list, with 300-dimensional word embedding vectors trained on the Google News data set. Thus, it would be beneficial to explore the possibilities of text-based industry classification schemes with other types of data: companies outside of the S&P 500 list, word embedding vectors trained on various types of text data not restricted to news articles and vectors trained on algorithms other than a skip-gram model with different parameters.

2. In-depth experiments with parameters in cluster analysis

   Cluster analysis has various parameters depending on its algorithm. Even with the same cluster algorithm, modifying these parameters can result in significantly different clusters, which may impact the result. Thus, it is highly recommended that various possibilities of cluster analysis be explored.

3. Validity assessments from different perspectives

   It was mentioned earlier that qualitative assessment of the scheme depended only on subjective judgment. Further, in assessing the scheme related to a firm's homogeneity in financial characteristics, this thesis focused on few financial measures that were extracted at one single point, which may have led to a lack of generality in the result. Overall, the comparison with existing classification schemes focused only on the GICS industry classification system, but there are many other comparable schemes. Therefore, it would be highly beneficial to examine the scheme's validity in approaches other than the one suggested.

4. Experiment with words that represent industry categories

   As mentioned earlier, industry category words used for this study may not be the most suitable terms since they are rarely used in new articles. To utilize the relationship captured in news articles, it would be better to use terms that can represent industry categories and present frequently in news articles. Although, the selection of terms may involve subjective judgement to decide suitable terms to be used in analysis.

References

Accenture annual report. (2017). [Ebook].

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. Journal Of Machine Learning Research, 3, 1137–1155.

Bernstein, A., Clearwater, S., & Provost, F. (2003). The Relational Vector-space Model and Industry Classification.

Bhojraj, S., Lee, C., & Oler, D. (2003). What's My Line? A Comparison of Industry Classification Schemes for Capital Market Research. Journal Of Accounting Research, 41(5).

Blei, D. (2012). Probabilistic topic models. Communications Of The ACM, 55(4), 77. doi: 10.1145/2133806.2133826

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. Journal Of Machine Learning Research, 12, 2493-2537.

Dalziel, M. (2007). A systems-based approach to industry classification. Research Policy, 36(10), 1559-1574. doi: 10.1016/j.respol.2007.06.008

Feldman, R., & Dagan, I. (1995). Knowledge Discovery in Textual Databases (KDT). In Proc. Of The First Int. Conf. On Knowledge Discovery (KDD), 112-117.

Fontana, F. (2018). Why the FANG Stocks' Dominance May Not Be So Bad for the Market. TheStreet. [online] Available at: https://www.thestreet.com/investing/stocks/why-the-fang-stocks-dominance-may-not-be-bad-for-market-14594253 [Accessed 16 Jul. 2018].

Google Code Archive. (2018). Retrieved from https://code.google.com/archive/p/word2vec/

Gupta, M., & Huefner, R. (1972). A Cluster Analysis Study of Financial Ratios and Industry Characteristics. Journal Of Accounting Research, 10(1), 77. doi: 10.2307/2490219

Hotho, A., Nurnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining.

Jensen, R. (1971). A Cluster Analysis Study of Financial Performance of Selected Business Firms. The Accounting Review, 46(1), 36-56.

Kakushadze, Z., & Yu, W. (2016). Statistical Industry Classification. SSRN Electronic Journal. doi: 10.2139/ssrn.2802753

List of S&P 500 companies. (2018). Retrieved from https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality.

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. Proceedings Of NAACL-HLT 2013, 746–751.

NAICS Update Process Fact Sheet. (2017). [Ebook].

naics@census.gov, S. (2018). Frequently Asked Questions (FAQs) - NAICS - US Census Bureau. Retrieved from https://www.census.gov/eos/www/naics/faqs/faqs.html#q14

OFFICE OF MANAGEMENT AND BUDGET (OMB). (1998). North American Industry Classification System: United States, 1997.. Bernan Press: Lanham, ND.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536. doi: 10.1038/323533a0

SAUNDERS, N. C. (1999). The North American Industry Classification System: Change on the Horizon. Occupational Outlook Quarterly, 34-37.

Shi, Z., Lee, G., & Whinston, A. (2016). Toward a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence. MIS Quarterly, 40(4), 1035-1056. doi: 10.25300/misq/2016/40.4.11

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., & Pinto, D. (2014). Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. Computación Y Sistemas, 18(3). doi: 10.13053/cys-18-3-2043

Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. Bulletin Of The IEEE Computer Society Technical Committee On Data Engineering, 24(4), 35-42.

Standard & Poor’s. (2006). Global Industry Classification Standard (GICS®) [Ebook]. New York.

Verizon Annual Report. (2017). [Ebook].

Weiner, C. (2005). The Impact of Industry Classification Schemes on Financial Research. SSRN Electronic Journal. doi: 10.2139/ssrn.871173

Why the FANG Stocks' Dominance May Not Be So Bad for the Market. (2018). Retrieved from https://www.thestreet.com/investing/stocks/why-the-fang-stocks-dominance-may-not-be-bad-for-market-14594253

Yahoo Finance. (2018). Retrieved from https://finance.yahoo.com/

Appendix A:GICS classification categories

Industry Group	Sub Category
Energy	Oil & Gas Drilling
	Oil & Gas Equipment & Services
	Integrated Oil & Gas
	Oil & Gas Exploration & Production
	Oil & Gas Refining & Marketing
	Oil & Gas Storage & Transportation
	Coal & Consumable Fuels
Materials	Commodity Chemicals
	Diversified Chemicals
	Fertilizers & Agricultural Chemicals
	Industrial Gases
	Specialty Chemicals
	Construction Materials
	Metal & Glass Containers
	Paper Packaging
	Aluminum
	Diversified Metals & Mining
	Copper
	Gold
	Precious Metals & Minerals
	Silver
	Steel
	Forest Products
	Paper Products
Industrials	Aerospace & Defense
	Building Products
	Construction & Engineering
	Electrical Components & Equipment
	Heavy Electrical Equipment
	Industrial Conglomerates
	Construction Machinery & Heavy Trucks
	Agricultural & Farm Machinery
	Industrial Machinery
	Trading Companies & Distributors
	Commercial Printing
	Environmental & Facilities Services
	Office Services & Supplies
	Diversified Support Services
	Security & Alarm Services
	Human Resource & Employment Services

Industry Group	Sub Category
Consumer Staples	Drug Retail
	Food Distributors
	Food Retail
	Hypermarkets & Super Centers
	Brewers
	Distillers & Vintners
	Soft Drinks
	Agricultural Products
	Packaged Foods & Meats
	Tobacco
	Household Products
	Personal Products
Health Care	Health Care Equipment
	Health Care Supplies
	Health Care Distributors
	Health Care Services
	Health Care Facilities
	Managed Health Care
	Health Care Technology
	Biotechnology
	Pharmaceuticals
	Life Sciences Tools & Services
Financials	Diversified Banks
	Regional Banks
	Thrifts & Mortgage Finance
	Other Diversified Financial Services
	Multi-Sector Holdings
	Specialized Finance
	Consumer Finance
	Asset Management & Custody Banks
	Investment Banking & Brokerage
	Diversified Capital Markets
	Financial Exchanges & Data
	Mortgage REITs
	Insurance Brokers
	Life & Health Insurance
	Multi-line Insurance
	Property & Casualty Insurance
	Reinsurance
Information Technology	Internet Software & Services

	Research & Consulting Services
	Air Freight & Logistics
	Airlines
	Marine
	Railroads
	Trucking
	Airport Services
	Highways & Railtracks
	Marine Ports & Services
Consumer Discretionary	Auto Parts & Equipment
	Tires & Rubber
	Automobile Manufacturers
	Motorcycle Manufacturers
	Consumer Electronics
	Home Furnishings
	Homebuilding
	Household Appliances
	Housewares & Specialties
	Leisure Products
	Apparel, Accessories & Luxury Goods
	Footwear
	Textiles
	Casinos & Gaming
	Hotels, Resorts & Cruise Lines
	Leisure Facilities
	Restaurants
	Education Services
	Specialized Consumer Services
	Advertising
	Broadcasting
	Cable & Satellite
	Movies & Entertainment
	Publishing
	Distributors
	Internet & Direct Marketing Retail
	Department Stores
	General Merchandise Stores
	Apparel Retail
	Computer & Electronics Retail
	Home Improvement Retail
	Specialty Stores
	Automotive Retail
	Homefurnishing Retail

	IT Consulting & Other Services
	Data Processing & Outsourced Services
	Application Software
	Systems Software
	Home Entertainment Software
	Communications Equipment
	Technology Hardware, Storage & Peripherals
	Electronic Equipment & Instruments
	Electronic Components
	Electronic Manufacturing Services
	Technology Distributors
	Semiconductor Equipment
	Semiconductors
Telecommunication Services	Alternative Carriers
	Integrated Telecommunication Services
	Wireless Telecommunication Services
Utilities	Electric Utilities
	Gas Utilities
	Multi-Utilities
	Water Utilities
	Independent Power Producers & Energy Traders
	Renewable Electricity
Real Estate	Diversified REITs
	Industrial REITs
	Hotel & Resort REITs
	Office REITs
	Health Care REITs
	Residential REITs
	Retail REITs
	Specialized REITs
	Diversified Real Estate Activities
	Real Estate Operating Companies
	Real Estate Development
	Real Estate Services