information

Article

Comparative Analysis of NLP-Based Models for Company Classification

Maryan Rizinski [1,2,*], Andrej Jankov [2], Vignesh Sankaradas [1], Eugene Pinsky [1], Igor Mishkovski [2] and Dimitar Trajanov [1,2,*]

[1] Department of Computer Science, Metropolitan College, Boston University, Boston, MA 02215, USA
[2] Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, 1000 Skopje, North Macedonia
* Correspondence: rizinski@bu.edu (M.R.); dtrajano@bu.edu (D.T.)

Abstract: The task of company classification is traditionally performed using established standards, such as the Global Industry Classification Standard (GICS). However, these approaches heavily rely on laborious manual efforts by domain experts, resulting in slow, costly, and vendor-specific assignments. Therefore, we investigate recent natural language processing (NLP) advancements to automate the company classification process. In particular, we employ and evaluate various NLP-based models, including zero-shot learning, One-vs-Rest classification, multi-class classifiers, and ChatGPT-aided classification. We conduct a comprehensive comparison among these models to assess their effectiveness in the company classification task. The evaluation uses the Wharton Research Data Services (WRDS) dataset, consisting of textual descriptions of publicly traded companies. Our findings reveal that the RoBERTa and One-vs-Rest classifiers surpass the other methods, achieving F1 scores of 0.81 and 0.80 on the WRDS dataset, respectively. These results demonstrate that deep learning algorithms offer the potential to automate, standardize, and continuously update classification systems in an efficient and cost-effective way. In addition, we introduce several improvements to the multi-class classification techniques: (1) in the zero-shot methodology, we TF-IDF to enhance sector representation, yielding improved accuracy in comparison to standard zero-shot classifiers; (2) next, we use ChatGPT for dataset generation, revealing potential in scenarios where datasets of company descriptions are lacking; and (3) we also employ K-Fold to reduce noise in the WRDS dataset, followed by conducting experiments to assess the impact of noise reduction on the company classification results.

Keywords: company classification; industry classification; natural language processing; machine learning; deep learning; finance; fintech

Citation: Rizinski, M.; Jankov, A.; Sankaradas, V.; Pinsky, E.; Mishkovski, I.; Trajanov, D. Comparative Analysis of NLP-Based Models for Company Classification. *Information* **2024**, *15*, 77. https://doi.org/10.3390/info15020077

Academic Editors: Agnes Vathy-Fogarassy and János Abonyi

Received: 9 January 2024
Revised: 20 January 2024
Accepted: 25 January 2024
Published: 31 January 2024

1. Introduction

Over the past few years, machine learning (ML) and natural language processing (NLP) have become increasingly prominent in various business domains, revolutionizing the way organizations operate. The surge in the availability of enormous textual data, coupled with the increasing complexity of financial markets, has necessitated the application of advanced NLP models to extract meaningful insights. Many finance-related areas benefit from these technologies, including financial research [1–3], business analytics [4–7], risk assessment [8–10], stock market prediction [11–13], and financial sentiment analysis [14–17].

NLP-based models offer a distinct advantage by enabling the automated extraction, comprehension, and analysis of textual information from diverse sources, including news articles, financial reports, social media, and more. The integration of these models not only enhances the efficiency of data processing but also contributes to substantial cost reduction through the automation of previously manual tasks. In the context of business applications, this translates into improved risk assessment, more comprehensive market analysis, and more informed decision-making processes.

In financial research, a company classification is a popular approach that involves grouping similar companies into categories or clusters [18,19]. The process of classifying companies into discrete categories has numerous practical applications for financial researchers, analysts, decision-makers, and investors. For example, it can help manage portfolio risk, facilitate relative valuation, and enable peer-group comparisons [20]. Additionally, it can aid in analyzing the effects of corporate reorganizations, changes in financial and investment policies, and the evaluation of the performance of a specific company against a set of similar companies. Beyond the financial sector, company classification can generate prospective leads for sales and marketing teams, identify new clients for insurance companies, and pinpoint competitors for corporations. Investment banks and venture capital firms can also benefit from company classification by understanding the distribution of companies among different industries [21].

The task of company classification has traditionally relied on established standards, such as the Standard Industrial Classification (SIC), the North American Industry Classification System (NAICS), the Fama French (FF) model, and the Global Industry Classification Standard (GICS), among others. Nonetheless, these methods have several limitations that hinder their effectiveness. These standards require time-consuming and effort-intensive manual analysis and data processing by domain experts. The human-based results can be subjective and prone to inaccuracies, as the existing classification schemes are often constructed and maintained by domain experts. Another significant challenge is the lack of interoperability between different schemes, which results in classification inconsistencies due to vendor-specific assignments. The lack of unified standardization further emphasizes the limitations of the existing approaches. Additionally, the classification process across different data vendors is prone to inconsistencies, leading to issues with accuracy and homogeneity. An effective classification scheme should ensure a high degree of homogeneity within each company cluster, which may not always be streamlined due to the existence of various data vendors.

As products and services become increasingly complex, updating classification schemes becomes a challenging task. The dynamic market environment in which companies operate causes frequent changes in their business, affecting their industry affiliation. The existing classification standards are static and cannot keep up with the fast-changing environments. The current schemes heavily rely on self-reporting and manual entry, resulting in slow, costly, and ineffective updates when adapting to the changed business landscape. In the absence of capabilities for real-time updates, these existing standards may not be the optimal choice in various application settings, thereby emphasizing the need to explore automation techniques. Ultimately, selecting an appropriate classification scheme becomes a non-trivial task due to the multitude of available standards. A notable limitation in this context is the potential for discrepancies among different data vendors employed for classification, even when adhering to the same classification standard. The discrepancies may result in the classification of the same company into different clusters, even within the same classification standard.

The recent advancements in ML and NLP can be explored to address the limitations of the traditional standards for company classification and hold promises for reducing costs, complexity, and manual labor. In particular, text classification using NLP methods has significantly progressed over the past decade. Large-scale pretrained transformer models have revolutionized the field of text classification, enabling successful implementations across various application domains, such as machine translation, text summarization, and sentiment analysis. These models witnessed successful deployments in systems that need scalability and real-time analysis, making them also valuable in addressing the problem of company classification.

In this paper, we explore the use of various deep learning techniques, including zero-shot learning, One-vs-Rest classification, multi-class classifiers, and ChatGPT-aided classification, on the Wharton Research Data Services (WRDS) dataset. The WRDS dataset contains names and textual descriptions of 34,338 companies classified per the GICS index. In the classification experiment, we calculate standard metrics, such as precision, recall, F1

score, and support, for each classification category, as well as for the overall model. We aim to evaluate the potential of the studied techniques for company classification.

We introduce several improvements to the standard multi-class classifiers. First, we enhance the zero-shot methodology by leveraging term frequency-inverse document frequency (TF-IDF) to extract common words associated with each company sector. Our findings reveal that this approach helps obtain a more precise representation of the sector names, resulting in improved accuracy compared to a conventional zero-shot classifier without TF-IDF. Secondly, we utilize ChatGPT to create a dataset of company descriptions to evaluate its impact on classification performance. Due to ChatGPT's ability to provide more detailed company descriptions, we conduct a series of experiments, including zero-shot, multi-class, and One-vs-Rest classifiers. While this method results in overall inferior results, it showcases ChatGPT's potential for dataset generation, particularly in scenarios where such datasets comprising company descriptions are not readily available. Finally, we employ K-Fold to mitigate the considerable amount of noise in the WRDS dataset. We perform experiments involving both cleaned and uncleaned validation datasets, which help to provide insights into the impact of noise reduction on the company classification results.

The paper is structured as follows: Section 2 focuses on background information and preliminaries, introducing the definition of company classification and providing an overview of mainstream standards. In Section 3, we present a review of related work in the literature. In Section 4, we describe several NLP-based models for the purpose of company classification and introduce the datasets used to train and evaluate the models. Section 5 presents a thorough examination of the models employed in this study, accompanied by a comparative analysis among the models and an in-depth discussion of the results obtained from the experiments. This section aims to offer a comprehensive understanding of the utilized models, emphasizing their strengths and weaknesses in order to contribute valuable insights to the company classification context. Finally, Section 6 offers a concise summary and the conclusions of the paper.

2. Standards for Company Classification

2.1. Definition and Benefits of Company Classification

Company classification, also known as industry classification, involves categorizing companies based on their business activities, industry, and other relevant factors. This process aims to group similar companies together and distinguish them from others based on several comparison parameters [22]. Company classification results in the formation of distinct groups of companies. Each group consists of companies that share similar business types, ensuring a coherent categorization. Simultaneously, these groups should exhibit differences from one another, emphasizing their unique characteristics and diverse nature.

In the context of company classification, homogeneity refers to the degree of similarity or uniformity among the companies within a particular category or group. When classifying companies, analysts and researchers often categorize them into groups based on specific criteria, such as industry, size, business model, or financial performance. Homogeneity implies that the companies within a specific classification share similar characteristics or attributes. Homogeneity is a crucial criterion for selecting an industry classification standard from the available options, and it is typically evaluated using various approaches. For instance, ref. [20] suggests using stock return co-movement, while [18] recommends utilizing 12 fundamental variables. By segmenting the market into partitions with distinct business and financial characteristics [23,24], these classifications provide a framework for understanding the similarities and differences among companies. The goal is to identify groups of businesses that engage in similar market activities and have comparable market conditions.

Company classification serves multiple purposes and provides numerous benefits. It has been demonstrated that companies within the same group tend to experience concurrent movements in their stock returns while exhibiting weaker returns correlation compared to companies in other groups [20]. Performing adequate classification may also facilitate cluster-based research, such as industry analysis and strategy development [25]. Further-

more, it facilitates the identification of peers and competitors, benchmarking of company activities and performance, measurement of economic indicators, quantification of market share, and the construction of exchange-traded funds (ETF) products [23,24]. Ultimately, company classification plays a vital role in various sectors, such as government, private sector, academia, and even the broader public, serving as a fundamental component of business and economic information [23].

2.2. Mainstream Standards

Industry classification standards are an essential tool for economic analysis, financial research, and policy-making. Among the most prominent industry classification systems are the Standard Industrial Classification (SIC), the North American Industry Classification System (NAICS), Fama French (FF), and the Global Industry Classification Standard (GICS). The SIC system, established in the 1930s by the Interdepartmental Committee on Industrial Classification, holds the distinction of being the oldest among the four classification systems. It was developed under the umbrella of the Central Statistical Board in the United States. Over time, the SIC system experienced periodic revisions to adapt to changes in the economy, with the most recent revision being performed in 1987 [18]. However, these efforts were insufficient, prompting the governmental statistical agencies of the United States, Canada, and Mexico to collaborate on a joint initiative to improve the SIC system and create a more comprehensive and unified classification scheme across North America. The result was the creation of the NAICS system in 1999. The 2017 edition of the NAICS taxonomy partitions the North American economy into 1057 industries, each assigned a unique six-digit code. NAICS employs a hierarchical classification scheme, categorizing each industry into distinct levels: industry groups, subsectors, and sectors. Initially, each industry is classified into an industry group, which then belongs to a specific subsector. Each subsector, in turn, is a defined segment within a broader sector. The industry group, subsector, and sector are represented by the first four, first three, and first two digits of the NAICS code, respectively. The NAICS system encompasses a total of 20 sectors, 99 subsectors, and 311 industry groups [21].

The FF system was initially conceptualized by academic researchers in finance as a means to investigate the industrial cost of capital [26]. FF achieves its purpose by reclassifying the existing SIC codes and grouping companies into 48 distinct industry sectors. Despite its prevalence in academic research concerning asset pricing, corporate finance, accounting, and economics, the FF system has not gained much popularity within the financial industry. In contrast, the Global Industry Classification Standard (GICS) was specifically designed by Standard & Poor's (S&P) and Morgan Stanley Capital International (MSCI) to meet the needs of financial professionals, such as investment managers and financial analysts. As shown in Table 1, GICS employs an eight-digit code to classify companies, and its structure is hierarchical, encompassing ten sectors subdivided into 24 industry groups, 64 industries, and 139 subindustries. The leading two digits, four digits, six digits, and the full eight-digit code of GICS are used to represent sectors, industry groups, industries, and subindustries, respectively (a more detailed information about the GICS index is available at https://www.msci.com/our-solutions/indexes/gics, accessed on 15 January 2024). The GICS scheme classifies companies based on their principal business activity, sources of revenue and earnings, as well as market perception concerning their primary line of business [20]. To achieve company assignments into different sectors, S&P and MSCI have leveraged information from annual reports and financial statements, including investment research reports and other information relevant to the financial industry [18].

GICS has been shown to outperform other popular industry classification systems, such as SIC, NAICS, and FF, in various comparison experiments [18,20,27–29]. This is due to its superior ability to capture industry homogeneity, leading to more accurate industry classifications. Additionally, the GICS index has been found to exhibit robust classification performance not only in settings with large and well-known companies (e.g., S&P companies) but also when applied to smaller and less-followed companies [30]. This

strong performance across a wide range of companies and industries makes the GICS index an ideal candidate for deep learning contexts, as utilized in this paper.

Table 1. GICS taxonomy illustrating the names of the four classification levels, from broadest to narrowest, the number of categories for each classification level, and the number of digits used to represent each level.

GICS Taxonomy			
Level	**Title**	**Number of Categories**	**Digits**
Level 1 (broadest)	Sector	11	first 2 digits
Level 2	Industry Group	24	first 4 digits
Level 3	Industry	64	first 6 digits
Level 4 (narrowest)	Sub-industry	139	all 8 digits

Apart from the widely used mainstream classification schemes, there are several accessible alternatives that may not be as popular among institutional practitioners. As highlighted in [19], these schemes share several common features. Firstly, the criteria used to categorize companies into groups are not publicly known. Secondly, the data vendors typically assume the role of assigners in these schemes. Lastly, the primary objective of the schemes is commercial in nature. The interested reader may refer to Bloomberg, Capital IQ (available on https://finance.yahoo.com, accessed on 15 January 2024), Hoovers & First Research, Market Guide, MarketLine, Morningstar, and Thomson Reuters (available on https://www.msn.com/en-us/money, accessed on 15 January 2024) [19]. Other schemes include the Thomson Reuters Business Classification (TRBC), the Industry Classification Benchmark (ICB), and the International Standard Industrial Classification of All Economic Activities (ISIC) [31].

2.3. Issues with the Current Standards

Despite the presence of diverse standards available for company classification and their wide use, the existing classification schemes suffer from several important limitations that deserve attention. Assigning companies to industries is currently performed manually and is vendor-specific [19]. The process is time-consuming, effort-intensive, subjective, and prone to inaccuracies, as existing classification schemes are often constructed and maintained by domain experts. These schemes can also become quickly outdated due to market developments and changes in products, technology, and business patterns, making them inadequate to properly reflect the fast-changing market dynamics [24]. Therefore, relying solely on human-aided classification is not optimal. In fact, even with domain expertise and sufficient data, deciding which companies belong to an industry is not straightforward. The presence of diverse classification schemes and the absence of unified standardization further exacerbate these limitations.

Inconsistencies in the classification process across different data vendors can pose issues in terms of accuracy and homogeneity, as highlighted by [18]. This is exemplified by the findings in [32], which analyzed manually assigned SIC codes for companies and revealed a significant discrepancy between two major data providers (Compustat and SRSP). Specifically, at the two-digit level, approximately 36% of SIC classifications differed, while at the four-digit level, almost 80% differed, as noted in [25]. In this context, the quality of the assignments into groups is essential to ensure cluster quality. An effective classification scheme should ensure the partitioning of companies into clusters that exhibit a high degree of homogeneity within each cluster (referred to as within-class homogeneity) while also ensuring that different clusters are distinctly different from one another (known as between-class heterogeneity) [24].

The timely update of classification schemes to reflect changing business and industry environments is another concern. For example, in [32], it was observed that the SIC codes are not permanently assigned and change over time due to the changes in the business of the companies, leading to changes in their industry affiliation. Thus, regardless of the underlying standard used, the classification scheme should incorporate current and

frequently updated data. However, this poses a major challenge as adjusting and updating classification schemes to reflect changes in company structure and operations requires extensive human efforts, which can be both time-consuming and expensive [24].

With a multitude of standards available, it is not trivial to select an appropriate industry classification scheme. The study [33] highlights several issues in this regard. A major concern is that the different data sources used for classification exhibit mismatch even when applied within the same classification system. As a result, the same company can be classified in different partitions even under the same classification system if different data sources are used for the classification. Moreover, the data are typically extracted from static data sources. This provides a suboptimal strategy for assigning companies into industries, emphasizing the need for historically correct and dynamic data for classification purposes [33].

3. Related Literature

Although some studies have been conducted in recent years, the literature on the application of NLP methods for industry classification remains limited overall.

In [34], the authors investigate the effectiveness of deep learning models on encyclopedic data from the English DBpedia Knowledge Base (https://www.dbpedia.org, accessed on 15 January 2024). Specifically, the study evaluates the performance of two popular models, Glove and ULMfit, against two baseline models (one-hot unigram and one-hot bigram). The dataset used for the experiments includes 300,000 textual descriptions of companies from DBpedia. While the company descriptions are uniform in length and style, the dataset contains a significant variation in industry representation. The dataset comprises 32 industries, showcasing a significant variation in the number of companies, spanning from the largest industry, with 76,000 companies, to the smallest industry, with approximately 300 companies. The findings of the study reveal that the tested models exhibit similar performance, and none of them can be designated as superior. The tested models perform well on the larger classes but exhibit a decline in performance when dealing with the smaller classes. Furthermore, different models show substantial variation in behavior in the smaller classes. Importantly, the study does not use a dataset that is considered a "gold standard", thereby attributing the inferior algorithm performance to errors in the underlying data rather than to the algorithm's inner workings. Another limitation is the absence of an established industry taxonomy in DBpedia.

A relevant study, [31], uses the same dataset and experimental setup as [34], but introduces BERT and XLNet models in addition to Glove and ULMfit. This study compares the four models with the same baseline (one-hot unigram and one-hot bigram) and finds that all the algorithms perform acceptably in well-represented classes, but experience decreased overall performance in less-represented classes. Although no algorithm stands out as the best for small classes, XLNet and BERT demonstrate more stable performance overall, thanks to their superior F1 scores. As noted in [31], there are currently no benchmark datasets for industry classification. Previous studies rely on "Industry Sector" data comprising 6000 company descriptions collected from the web and classified into 70 industry sectors, but they utilize algorithms that are considered outdated in modern big data applications; these algorithms include Naive Bayes (NB), multinomial NB, Maximum Entropy classifier, Support Vector Machine, and k-Nearest Neighbors.

The authors of [35] investigate the usefulness of text-based industry classification using various word and document embedding techniques in conjunction with different clustering algorithms. Their approach is applied to publicly traded companies in both the US and Chinese markets, and the results are compared against the GICS index as the standard is available in both markets. For Chinese companies, the study relies on company descriptions from the China Securities Regulatory Commission (CSRC), while for US companies, it uses data from Yahoo. The study utilizes advanced embedding techniques, such as BERT, but surprisingly, the results show that a simpler technique, latent semantic indexing (LSI), combined with k-means clustering, outperforms BERT on two measures. This finding is remarkable because LSI, an extension of conventional techniques,

such as bag of words (BoW) and TF-IDF, is not commonly used in state-of-the-art (SOTA) text classification applications. As a result, this study sheds new light on the potential usefulness of LSI for text-based industry classification.

The study in [36] employs graph neural networks (GNNs) to facilitate the classification of Chinese firms based on supply chain network information. It harnesses the advantages of GNNs and aims to categorize companies according to the CSRC classification scheme. While this method aligns with the conventional approach adopted by Chinese scholars for analyzing the Chinese economy, it is important to note that the CSRC classification is country-specific. The paper, however, has not explored the application of GNNs in the context of the GICS standard.

A method for fine-tuning a pretrained BERT model is proposed in [37], which is then evaluated on two datasets consisting of US and Japanese company data. The US dataset includes 2462 annual reports from 2019 of companies listed on the US stock market (Form 10-K documents), while the Japanese dataset contains 3016 annual reports from 2018 of companies listed on the Tokyo Stock Exchange. The paper's objective is to explore the extent to which companies with similar vector representations operate in comparable industries, as well as to evaluate how effectively companies can be classified within a given industry based solely on the industry name. The study compares BERT to two baseline models, namely, BoW representation and skip-gram Word2Vec embedding, and demonstrates BERT's superior performance. The findings confirm the effectiveness of the proposed approach and suggest the integration of additional sources of business data, such as the price earnings ratio (PER) and the price book-value ratio (PBR), to augment annual reports for the purpose of industry classification.

In the study [24], a novel classification scheme called business text industry classification (BTIC) is introduced. BTIC is developed on a dataset comprising Form 10-K documents of S&P500 companies. To categorize companies into distinct clusters, the authors employ Doc2Vec for document embedding and Ward's hierarchical clustering method. The study examines different factors to assess the homogeneity of each industry cluster. The findings indicate that BTIC performs comparably to established classification schemes like GICS and SIC in terms of grouping companies into homogeneous clusters. Furthermore, the paper showcases the potential of BTIC to surpass existing classification schemes in additional areas, such as process automation, objectivity, flexibility, and result interpretability.

A novel knowledge graph enriched BERT (KGEB) model, which is capable of loading any pretrained BERT model and fine-tuning it for classification, is presented in [38]. KGEB enhances word representations by incorporating additional knowledge through learning the graph structure of the underlying dataset. The model is tested on a dataset of publicly listed companies on the Chinese National Equities Exchange and Quotations (NEEQ). The dataset consists of 17,604 annual business reports and their corresponding industry labels. KGEB is shown to outperform five models that are selected as baselines, such as the graph convolutional network (GCN), logistic regression, TextCNN, BERT, and K-BERT. The findings highlight that enriching word representations with knowledge graphs is beneficial as it takes into account the structure of the extracted graph, thereby improving the classification of domain-specific texts.

The authors of [21] employed a deep neural network based on a multilayer perceptron architecture with four fully connected layers to predict the industries of novel companies. The model is trained on a dataset sourced from the proprietary EverString database (EverString was acquired by ZoomInfo in 2020). Due to the considerable size of the dataset, each company is represented by a sparse feature vector to facilitate the model training. This representation consists of a weighted combination of the most relevant keywords present in the company's description, resulting in a sparse vector that assigns weights exclusively to keywords within the description. The study shows that this approach achieves higher precision and outperforms premier databases in classifying companies into six-digit NAICS codes, although it comes at the expense of sacrificing recall. The authors also evaluated the model's performance using LinkedIn industry codes with satisfactory results, indicating

the adaptability of this neural network approach to other classification schemes. However, while company codes on LinkedIn pages are largely self-reported and, thus, tend to be highly reliable, the authors acknowledge that LinkedIn industry classification is a far simpler task than performing six-digit NAICS classification. The study would have benefited from analyzing GICS as it is more popular than NAICS for industry classification among financial analysts and investment managers. The authors highlight the presence of highly noisy labels in their training dataset, resulting in suboptimal classification. The reliability of the training data is affected by NAICS taxonomy ambiguity, human error, and the use of naive algorithms by traditional data vendors for automatic industry classification. Due to these issues, the authors call for further efforts to create cleaner datasets.

In their study, the authors of [25] conducted an evaluation of 28 classifiers based on four underlying Word2Vec models with varying window sizes, different SVM kernels, and logistic regression solvers. The Word2Vec models were trained on a corpus of articles from the Guardian newspaper, consisting of 600 million words. The company-industry mappings are extracted from DBpedia for companies that occur in both the news dataset and DBpedia, allowing the development of an industry classification model that works on unseen companies. The analysis is promising in identifying company–industry mappings in news texts but has certain limitations. Firstly, as the authors acknowledge, this approach is not entirely robust for automatic company classification, and further investigation is required to verify the quality of training labels obtained through DBpedia. Secondly, while the dataset in the study is comprehensive, the study lacks a comparison with benchmark schemes like GICS. Lastly, it is worth noting that the use of Word2Vec, while effective, may be somewhat outdated in comparison to SOTA deep learning models that demonstrate superior performance in text classification tasks.

Additional relevant references include [39–44], among others. The study [39] introduces a multimodal neural model aimed at training company embeddings. This approach leverages the similarities found in both historical prices and financial news, enabling the model to capture nuanced relationships that exist between companies, thereby facilitating the identification of related companies. Ref. [42] reports the extraction of distinctive features from business descriptions in financial reports and the application of dimensionality reduction techniques to assess company homogeneity.

A method for company representation, called Company2Vec, based on unstructured textual and visual data from German company webpages, is presented in [45]. Company2Vec relies on Word2Vec and dimensionality reduction, demonstrating its ability to reflect companies' business activities based on the NACE codes. NACE codes, or "Nomenclature of Economic Activities", constitute a standard system commonly used in the European Union for classifying economic activities. The study [46] utilizes large language models to generate company embeddings by analyzing raw business descriptions extracted from Securities and Exchange Commission (SEC) 10-K filings. It assesses the capability of these embeddings to replicate GICS sector/industry classifications when employed as features. The research states a noteworthy limitation: the reduced interpretability of company embeddings generated by language models compared to traditional classification approaches.

The authors of [40] propose a deep learning method that leverages multiple sources of knowledge for company classification. Their model incorporates not only assignment-based knowledge (prior assignments performed by domain experts) but also definition-based knowledge (expert definition of each industry) as well as structure-based knowledge (relationships among industries as defined in a specific classification scheme). The latter two sources are often overlooked in existing methods. Although [41] does not utilize deep learning techniques, it uses the latest advancements in unsupervised machine learning through the integration of t-distributed stochastic neighbor embedding (t-SNE) and spectral clustering. This approach reduces the dimensionality of large datasets and generates visualizations that assist domain experts in making informed decisions regarding company

classification. By harnessing the power of t-SNE and spectral clustering, this methodology offers tools for data exploration and decision support.

The classification of companies using unstructured business news has been explored in [43,44]. In their work, the authors of [43] proposed a relational-vector space model that builds on existing classifications by considering the frequency of co-occurrences of companies within the same news article. Ref. [44], on the other hand, presents a corpus-based method for identifying groups of companies called "collective entities". This approach utilizes linguistic patterns to recognize collective entity names, their members, and the natural relationships among different collective entities. A list of selected papers and the respective models used in them is given in Table 2.

Table 2. A list of selected previous studies on company classification using NLP-based methods, ordered chronologically.

Referenced Paper	Year of Publication	Description
[24]	2016	Introduces a model called Business Text Industry Classification (BTIC), developed on a dataset comprising Form 10-K documents of S&P500 companies.
[21]	2017	Employs a deep neural network based on a multilayer perceptron architecture and trained on a dataset from the proprietary EverString database with the goal of classifying companies into six-digit NAICS codes.
[25]	2018	Uses Word2Vec models with varying window sizes, different SVM kernels, and logistic regression solvers. The models are trained on a corpus of Guardian articles, which consists of 600 million words. Evaluation is performed on company–industry mappings extracted from DBpedia.
[34]	2019	Compares Glove and ULMfit with two baseline models (one-hot unigram and one-hot bigram) using a dataset extracted from the English DBpedia. The dataset comprises 300,000 uniform-length textual descriptions of companies from 32 industries in DBpedia.
[31]	2019	Uses the same experimental setup as in [34] to assess BERT and XLNet, in addition to Glove and ULMfit.
[35]	2020	Investigates various word and document embedding techniques combined with clustering algorithms on datasets comprising publicly traded companies in the US and China. Compares the obtained results with GICS.
[37]	2020	Proposes a method for fine-tuning a pretrained BERT model, which is evaluated on datasets consisting of US and Japanese company data from Form 10-K documents and data from the Tokyo Stock Exchange, respectively.
[38]	2021	Introduces a model called knowledge graph enriched BERT (KGEB), tested on publicly listed Chinese companies. KGEB enhances word representations by learning the graph structure of the underlying dataset, and is capable of loading pretrained BERT and fine-tuning it for company classification.
[39]	2022	Employs a multimodal neural model that facilitates the identification of related companies by leveraging similarities found in historical prices and financial news.
[40]	2022	Proposes a deep learning method that leverages various sources of knowledge for company classification, such as assignment-based, definition-based, and structure-based knowledge.
[41]	2022	Uses unsupervised learning, employing t-SNE and spectral clustering, to reduce the dimensionality of large datasets and generate visualizations that assist domain experts in making informed decisions about company classification.
[45]	2023	Presents a model called Company2Vec based on Word2Vec and dimensionality reduction as well as on unstructured textual and visual data from German company webpages, aiming to predict companies' business activities based on NACE codes.
[46]	2023	Utilizes large language models to generate company embeddings by analyzing raw business descriptions extracted SEC 10-K filings. Assesses the ability of the embeddings to replicate GICS sector/industry classifications.
[36]	2023	Employs graph neural networks (GNNs) for classification of Chinese companies based on the China Securities Regulatory Commission (CSRC) classification scheme.

4. Materials and Methods

4.1. Dataset

We use the Wharton Research Data Services (WRDS) to create the dataset for this research. WRDS is a web-based data management system that provides researchers with access to a vast array of financial, economic, and marketing data from different sources, including Compustat, the Center for Research in Security Prices (CRSP), the Institutional Brokers' Estimate System (IBES), and others. WRDS is a research platform that has been developed and maintained by the Wharton School at the University of Pennsylvania to support researchers in their data-driven research activities. The platform is used by academic researchers, corporate professionals, and financial analysts to retrieve, manage, and analyze large sets of data for their research projects. WRDS also provides a suite of tools for data cleaning, analysis, and visualization to help users get the most out of the data available on the platform. Using WRDS, we extract the Compustat dataset, which contains financial and market data on publicly traded companies across the United States. The original dataset contains data for 44,033 companies, including their names, descriptions, and classification into sectors and industry groups as per the GICS taxonomy. We filter the dataset due to the absence of GICS sector assignment for some entries. After the filtering, the dataset consolidates a total of 34,338 entries (i.e., companies). The distribution of the number of companies across the GICS sectors is shown in Table 3. We use this dataset to perform diverse classification experiments employing various NLP approaches with deep learning, which will be explained in the subsequent parts of the paper (the code for all the experiments is available on GitHub at: https://github.com/nubs4dayz/company-classification-research, accessed on 15 January 2024).

Table 3. Distribution of companies in the WRDS dataset across various GICS sectors.

WRDS Dataset	
GICS Sector	**Number of Companies**
Energy	2822
Materials	3833
Industrials	3934
Consumer Discretionary	4662
Consumer Staples	1433
Health Care	4565
Financials	5363
Information Technology	5192
Communication Services	1285
Utilities	740
Real Estate	509

As presented in Table 4, the WRDS dataset, containing 34,338 entries, is denoted as W-Full to indicate that the full dataset size (after filtering) is employed in a specific experiment. In the analysis to follow, the dataset is partitioned into training and testing sets through an 80–20 split, yielding the W-Train and W-Test datasets. Additionally, we refined the WRDS dataset by removing the company names from the company descriptions to help mitigate bias in various experiments. The refined dataset is denoted as W-Full-R. Partitioning this dataset into training and testing sets using 80–20 split results in the W-Train-R and W-Test-R datasets, respectively. To mitigate the substantial noise present in the WRDS dataset, we applied the K-Fold technique for denoising. Specifically, denoising was performed on the W-Train dataset, resulting in the creation of its denoised counterpart, namely, the W-Train-C dataset. For our experiments, we created a dataset of company descriptions using ChatGPT (the dataset is available on GitHub: https://github.com/nubs4dayz/company-classification-research/blob/main/Datasets/gpt_generated.csv, accessed on 15 January 2024). To construct the dataset, we employed ChatGPT to generate 20 company descriptions for each of the 11 GICS sectors, excluding the company names (the ChatGPT prompts used

to generate the dataset can be found at the following link: https://chat.openai.com/share/0efde6df-6655-4965-8372-03548ebf5365, accessed on 15 January 2024). The resulting dataset, denoted as CG-Full in Table 4, encompasses 220 company descriptions. We then partitioned the ChatGPT-generated dataset into training and testing sets through an 80–20 split, resulting in the CG-Train and CG-Test datasets, respectively. As we will explain later on, we compress the W-Test dataset to 100 dimensions using an autoencoder to obtain the W-AE-Full dataset. We employed an 80–20 split on W-AE-Full to derive the datasets labeled as W-AE-Train and W-AE-Test. Subsequent experiments will illustrate that when evaluating One-vs-Rest classification on the W-Test dataset, the model incorrectly classifies a total of 1445 instances; these misclassified instances form a distinct dataset denoted as W-MC in Table 4. Furthermore, for additional evaluation, we utilize a dataset sourced from Kaggle (the Kaggle dataset with GICS-related data is available at the following link: https://www.kaggle.com/datasets/merlos/gics-global-industry-classification-standard, accessed on 15 January 2024), which comprises company descriptions along with their corresponding classifications based on the GICS taxonomy.

Table 4. Datasets used for the training and evaluating the models.

Dataset	Description	Purpose	Size
W-Full	Full size of the WRDS dataset	Test	34,338
W-Train	80% of the WRDS dataset	Train	27,470
W-Train-C	W-Train cleaned with K-Fold	Train	21,716
W-Test	20% of the WRDS dataset	Test	6868
W-Full-R	WRDS dataset with removed company names	Test	34,338
W-Train-R	80% of the W-Full-R dataset	Train	27,470
W-Test-R	20% of the W-Full-R dataset	Test	6868
CG-Full	Full size of the ChatGPT-generated dataset	Test	220
CG-Train	80% of the ChatGPT-generated dataset	Train	176
CG-Test	20% of the ChatGPT-generated dataset	Test	44
Kaggle	Kaggle dataset	Test	158
W-AE-Full	W-Test compressed to 100 dim. using autoencoder	Test	6868
W-AE-Train	80% of the W-AE-Full dataset	Train	5494
W-AE-Test	20% of the W-AE-Full dataset	Test	1374
W-MC	Dataset of misclassified company descriptions	Test	1445

The WRDS dataset is visualized in Figure 1 using t-SNE to achieve dimensionality reduction. The visualization employs 11 clusters to align with the number of sectors in the GICS taxonomy. Figure 1 illustrates that companies within the healthcare and finance sectors demonstrate notably homogeneous clustering attributed to distinct textual descriptors reflecting their industry-specific terminology. These linguistic nuances effectively differentiate entities within these two sectors from those in other sectors under analysis. Similarly, real estate companies also exhibit well-defined grouping, showing some overlap with the finance sector; this correlation is expected given the interconnected nature of these fields. Meanwhile, companies categorized under information technology display a more dispersed distribution across various clusters, reflecting the pervasive nature of this industry, which finds applications across diverse fields. Additionally, the visual representation of Figure 1 highlights an overlap between the clusters associated with information technology and communication services, emphasizing the close relationship between these two domains.

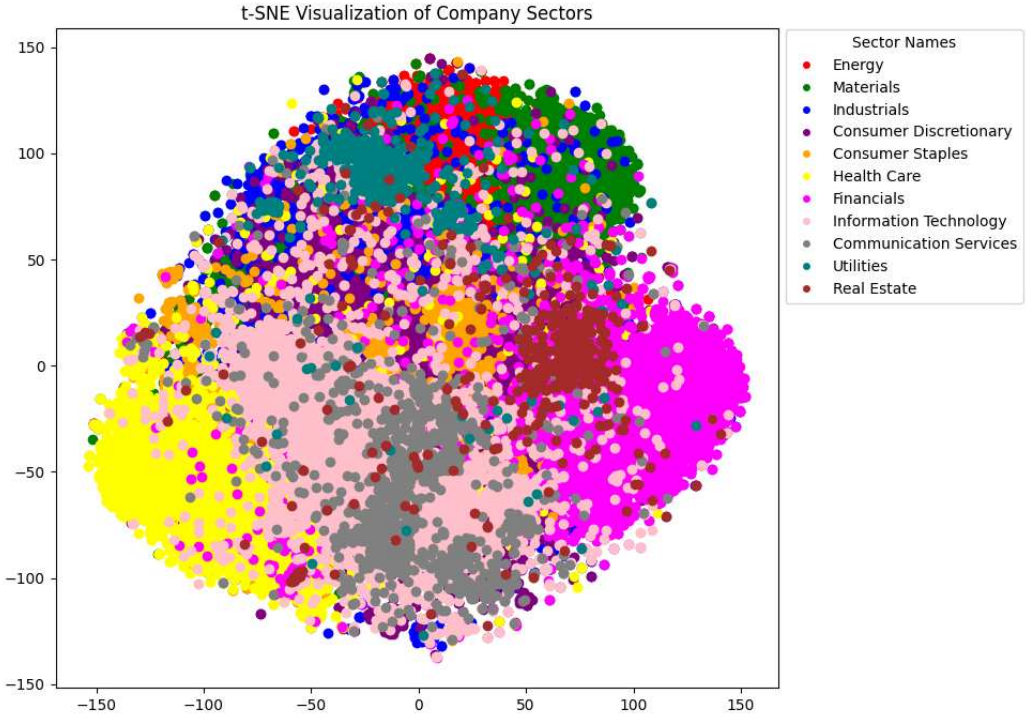


Figure 1. Visualization of the WRDS dataset performed using t-SNE to achieve dimensionality reduction. The clustering employs 11 clusters to align with the number of sectors in the GICS taxonomy.

4.2. Methodology

We utilize several methods for company classification. Firstly, we leverage the power of pretrained transformer models to perform zero-shot classification on the WRDS dataset. Zero-shot classification enables us to categorize data without the need for training data or the traditional division of datasets into train and test sets. Pretrained transformer models are particularly well-suited for this task, as they have been trained on extensive textual datasets, enabling them to classify text into classes even without prior knowledge of examples belonging to those classes. Our classification process involves analyzing both the original GICS sector names and the descriptions obtained through ChatGPT. The descriptions generated by ChatGPT provide more detailed information, and we aim to assess whether they can enhance the model's understanding of sector classification. Additionally, we apply zero-shot classification in conjunction with the third GICS level to categorize companies into industries.

Secondly, we employ an approach with a multi-class classifier, using a transformer trained on a large corpus of textual data in English. Our experiment involves labeling the sector indices, followed by dividing the WRDS dataset into train and test sets. After training the model on the training set, we then observe the model's effectiveness in classifying the GICS sectors on the test set.

We have enhanced the methodological foundation by incorporating additional references that support the comparative research design, mainly focusing on the usage of multi-class classifiers. To strengthen the underlying methodology, the model comparison builds upon prior references that explore multi-class classifiers as viable algorithms for text classification [47–51]. These references contribute to a more comprehensive understanding of the chosen experimental setup. By building upon prior results reported in the literature, our empirical study aims to provide novel insights into the relevance of text classification in areas such as company classification.

The task of company classification inherently belongs to the domain of multi-class classification. Therefore, we also adopt the One-vs-Rest classification approach, which leverages binary classification algorithms to address multi-class classification problems.

Our initial step involves employing the One-vs-Rest classifier with the support vector classifier as the estimator, utilizing the default radial basis function (RBF) kernel and the default number of iterations. Subsequently, we perform dimensionality reduction to assess its impact on the model's performance. For this purpose, we employ two distinct methods: principal component analysis (PCA) and autoencoder architecture.

We conducted an experiment using One-vs-Rest classification on a dataset comprising company descriptions generated by ChatGPT. Our primary objective was to assess the impact of ChatGPT-generated content on classification performance. To create the dataset, we employ ChatGPT to generate 20 company descriptions for each of the 11 GICS sectors, excluding the company names. Consequently, the dataset comprises a total of 220 company descriptions (20 descriptions per sector, across all 11 sectors).

To address the considerable amount of noise in the WRDS dataset, we employed K-Fold to denoise it. We trained a One-vs-Rest classifier using 80% of the cleaned WRDS dataset and evaluated the model's performance on the remaining 20% of the dataset, which was left uncleaned. Additionally, to further evaluate the model, we tested it on a separate dataset obtained from Kaggle, which includes company descriptions and their corresponding classifications based on the GICS taxonomy. Lastly, we compared the predictions made by our model with those made by ChatGPT on the 20% uncleaned dataset.

Furthermore, we investigated the use of a contextual sentence transformer, which presents a novel method for generating text embeddings. This method enables the embedding of textual input alongside instructions that explain the specific use case. To evaluate its efficacy, we conducted tests incorporating different contexts, leveraging the model embeddings in combination with a One-vs-Rest classifier.

These deep-learning approaches for company classification are explained in detail in the subsequent section. All results from the experiments are consolidated in Table 5.

Table 5. Company classification performance of various NLP-based models applied on the WRDS dataset. The performance is evaluated in terms of classification metrics, such as precision, recall, and F1 score. The highest numerical value in each column representing these metrics is highlighted in bold and underlined. All entries with One-vs-Rest are based on the all-mpnet-base-v2 model except for the case with the contextual sentence transformer, which uses the hkunlp/instructor-large model.

	Datasets		**Macro Average**			**Weighted Average**				
	Train	**Test**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Support**	**Final F1 Score**
Zero-shot classification with valhalla/distilbart-mnli-12-3										
Using original GICS sector names	—	W-Full	0.49	0.54	0.48	0.57	0.56	0.55	34338	0.56
Using GICS sector names enhanced with TF-IDF	—	W-Full	0.60	0.64	0.58	0.67	0.64	0.64	34338	0.64
Zero-shot classification on industries (part 1)	—	W-Full	0.51	0.10	0.12	0.72	0.77	0.69	34338	0.77
Zero-shot classification on industries (part 2)	—	W-Full	0.44	0.09	0.12	0.68	0.71	0.61	34338	0.71
Zero-shot classification on industries (part 3)	—	W-Full	0.61	0.13	0.17	0.65	0.60	0.48	34338	0.60
Multi-class classifier based on RoBERTa-base										
Using GICS sectors	W-Train	W-Test	0.77	**0.76**	**0.77**	**0.80**	**0.80**	**0.80**	6868	**0.80**
Using GICS industrial groups	W-Train	W-Test	0.72	0.70	0.71	0.75	0.75	0.75	6868	0.75
One-vs-Rest (OvR) classification with all-mpnet-base-v2										
OvR classifier using SVC estimator with RBF kernel	W-Train	W-Test	**0.78**	0.74	0.75	0.79	**0.80**	0.79	6868	**0.80**
OvR classifier using SVC estimator with cosine similarity kernel	W-Train	W-Test	0.76	0.72	0.73	0.77	0.78	0.77	6868	0.78
Using dimensionality reduction with PCA to 100 dim.	W-Train	W-Test	0.27	0.29	0.27	0.37	0.42	0.39	6868	0.42
Using dimensionality reduction with autoenc. arch. to 100 dim.	W-AE-Train	W-AE-Test	0.74	0.70	0.72	0.77	0.78	0.77	1374	0.78
Trained on W-Train-C and tested on W-Test	W-Train-C	W-Test	**0.78**	0.70	0.72	0.78	0.78	0.77	6868	0.78
Evaluated on WRDS with removed company names	W-Train-R	W-Test-R	0.77	0.73	0.75	0.77	0.78	0.77	6868	0.78
Using contextual sentence transformer (hkunlp/instructor-large)	W-Train	W-Test	**0.78**	0.72	0.73	0.79	0.79	0.78	6868	0.79
ChatGPT-based classification										
Using zero-shot classification with label descriptions	—	W-Full	0.60	0.67	0.58	0.69	0.61	0.61	34338	0.61
OvR with ChatGPT descriptions	CG-Train	CG-Test	0.52	0.55	0.50	0.59	0.52	0.51	44	0.52
OvR with ChatGPT descriptions	CG-Train	W-Full	0.56	0.52	0.40	0.66	0.46	0.39	34338	0.46
ChatGPT predicting on W-Test	—	W-Test	0.71	0.66	0.67	0.82	0.71	0.75	3434	0.71
ChatGPT predicting the W-MC	—	W-MC	0.24	0.21	0.21	0.28	0.22	0.23	1445	0.22

5. Results and Discussion

5.1. Zero-Shot Classification

In this section, we start by assessing the performance of zero-shot classification in two scenarios: one employing the original GICS sector names and the other utilizing GICS sector names enhanced by TF-IDF. We employ a zero-shot classification pipeline using the valhalla/distilbart-mnli-12-3 model. We chose the valhalla/distilbart-mnli-12-3 model for its popularity on Hugging Face, where it ranks as one of the top three models in terms of the number of downloads. To assess its effectiveness, we conducted a comparative analysis with two other models, namely facebook/bart-large-mnli and joeddav/xlm-roberta-large-xnli. Our evaluation showed that the valhalla/distilbart-mnli-12-3 model performed slightly better than the other two models.

The valhalla/distilbart-mnli-12-3 model belongs to the class of transformer models, which are a powerful type of neural network architecture that has been widely adopted in NLP [52]. Transformer models are capable of modeling long-range textual dependencies, thereby effectively capturing relationships between distant words in a sentence [53]. The transformer architecture includes an attention mechanism that allows the model to selectively focus on relevant parts of the input sequence. This enables the model to extract important relationships between words and better capture the meaning of the input text [54].

Zero-shot classification refers to the ability of a model to classify inputs into multiple classes without requiring any training data [55]. Pretrained transformer models have shown potential in zero-shot classification tasks as they have been trained on massive amounts of textual data, enabling them to classify inputs into classes even if they have never seen examples of those classes before. When needed, pretrained models can be fine-tuned on specific zero-shot classification tasks with only a small amount of training data, allowing the models to adapt to the underlying task and improve the overall performance (e.g., accuracy) on that task.

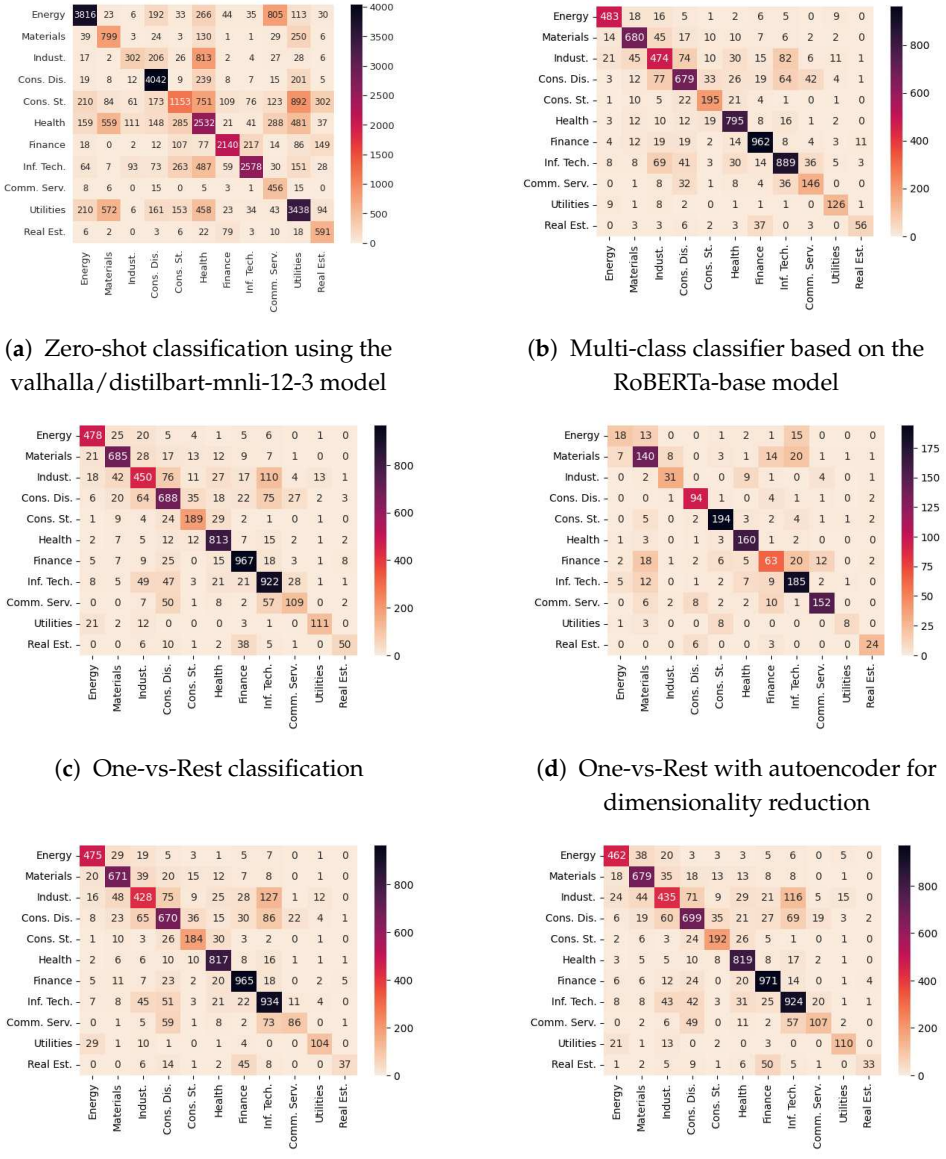
In our experiment, we utilized the valhalla/distilbart-mnli-12-3 model and adopted the zero-shot classification technique [56]. Specifically, we fed the model with the company descriptions available in the WRDS dataset (W-Full) without performing any fine-tuning of the model. As we employed the zero-shot learning approach, we did not need to divide the dataset into train and test sets.

We obtained an F1 score of 0.56 using the original category names from the GICS taxonomy. All F1 scores reported in this paper are rounded to the second decimal place. We aimed to boost the F1 score by modifying the sector names with alternative labels that do not impede the model's classification performance. Specifically, we utilized TF-IDF vectorization to extract the top 30 most common words for each sector to obtain a more precise representation of the sector names and improve the accuracy of the zero-shot classification model. This process involved preprocessing the company descriptions using the NLTK (https://www.nltk.org, accessed on 15 January 2024) library to identify all verbs in the dataset and exclude them as stop words (since verbs are the most frequently occurring words). In addition to default stop words, we also excluded country names and certain abbreviations (e.g., Ltd., LLC) that occur frequently but are not relevant to the classification task. The original and modified sector names are shown in Table 6. This technique increases the F1 score to 0.64. However, the aforementioned change failed to result in any significant improvement in the sectors that had the lowest F1 scores, namely, Real Estate, Consumer Staples, Consumer Discretionary, and Industrials.

The confusion matrix and classification report are presented in Figure 2a and in Table B2, respectively. As can be seen, the weighted F1 score obtained across the dataset is 0.64. The highest individual F1 scores are obtained for Health Care and Oil & Natural Gas with 0.84 and 0.81 F1 scores, followed by Banking & Lending and Raw Minerals & Mining with 0.77 and 0.75 F1 scores, respectively. The lowest F1 scores are observed for Food, Beverages and Household Products with 0.30 F1 score, Real Estate with 0.39 F1 score, and Industrials and Transportation with 0.39 F1 score.

Table 6. Modified sector names of the GICS taxonomy using TF-IDF preprocessing and removal of stop words with the purpose of increasing the F1 score in zero-shot classification.

Sector Names	
Original GICS Names	**Names after TF-IDF**
Energy	Oil, Natural Gas, Consumable Fuels and Petroleum
Materials	Raw Materials, Mining, Minerals and Metals (Gold, Silver and Copper)
Industrials	Industrials and Transportation
Consumer Discretionary	Non-Essential Goods, Retail and E-Commerce
Consumer Staples	Food, Beverages and Household Products
Health Care	Health Care
Financials	Banking and Lending
Information Technology	Software, Technology and Systems
Communication Services	Communications, Telecommunications, Networking, Media and Entertainment
Utilities	Utilities, Energy Distribution and Renewable Energy
Real Estate	Real Estate Properties



(**a**) Zero-shot classification using the valhalla/distilbart-mnli-12-3 model

(**b**) Multi-class classifier based on the RoBERTa-base model

(**c**) One-vs-Rest classification

(**d**) One-vs-Rest with autoencoder for dimensionality reduction

(**e**) One-vs-Rest with K-Fold applied on 80% of the WRDS dataset

(**f**) One-vs-Rest with the instructor-large model

Figure 2. Confusion matrix obtained on the WRDS dataset using various approaches.

We conducted an additional experiment to address the issue of unsatisfactory F1 scores. In this experiment, we employed a zero-shot classification pipeline utilizing the same valhalla/distilbart-mnli-12-3 model. Instead of splitting the dataset into train and test sets, we employed all descriptions from the dataset as input. We specifically focused on sectors that exhibited low F1 scores in our previous experiment. From the dataset, we extracted only the descriptions corresponding to these sectors. The purpose of this approach was to assess whether the model was still noisy when classifying only the selected sectors. Unfortunately, our findings indicate that these sectors remained problematic, resulting in persistently low F1 scores despite isolating their descriptions.

After assessing the GICS sectors, we also evaluated the zero-shot classification of GICS industry groups. In this approach, we employed the third GICS level to categorize companies into industries. As GICS comprises 69 levels (industries) in total, which can be challenging for multi-class classification, we divided them randomly into three separate lists, each containing 23 levels. The partitioning of industries into three lists is given in Table A1 in Appendix A. A possible next step in the experimental setup would be to also divide the dataset according to the respective industries as per their belonging to one of the three created lists. However, this approach may introduce bias and may prevent us from testing the entire dataset. Thus, to avoid these issues, the entire dataset was used as input. To ensure that the dataset is utilized in its entirety, we introduced a helper class called "No Class". This class helps in classifying companies that do not fall into the 23 industries specified in a given list. Without this class, a company belonging to any of the remaining 46 industries would inevitably and erroneously be forced into one of the 23 industries, resulting in ambiguity. Next, we established a cut-off threshold probability of 0.80, which is defined as follows: If a company's probability of belonging to a specific industry is below this threshold, it is assigned to the "No Class" category. Conversely, if the probability exceeds 0.80, the company is classified within the corresponding industry. Our analysis demonstrated that 0.80 is an optimal threshold across all three lists. Finally, the F1 scores obtained for the three lists were 0.77, 0.71, and 0.61, respectively.

5.2. Multi-Class Classifier Based on Roberta-Base

We continue by evaluating the performance of company classification using a multi-class classifier. To train the model for this experiment, we used the pretrained RoBERTa-base transformer from Hugging Face (the model was sourced from Hugging Face's AutoModelForSequenceClassification). RoBERTa, which stands for "A Robustly Optimized BERT Approach", is a large-scale, pretrained language model introduced by Facebook in 2019 [57]. It is based on the bidirectional encoder representations from the transformers (BERT) model architecture but incorporates several modifications and enhancements, resulting in improved performance on various natural language processing (NLP) tasks.

The RoBERTa-base model is one of the variants of the RoBERTa model (additional information about RoBERTa-base can be found on the Hugging Face website at the following link: https://huggingface.co/roberta-base, accessed on 15 January 2024). The base variant refers to a medium-sized version of the model, which is smaller and computationally less expensive compared to the larger variants like RoBERTa-large or RoBERTa-xlarge. With 125 million parameters, the RoBERTa-base model is still of significant size and is capable of achieving strong performance on a wide range of NLP tasks.

RoBERTa-base is trained using a large unlabeled corpus of publicly available text in English obtained from various sources, such as Wikipedia, books, and websites. The model learns to represent words and sentences in a self-supervised fashion by predicting masked tokens in a given input sentence and performing the next sentence prediction. This pretraining process enables the model to capture the contextual understanding of words and sentences, allowing it to encode rich representations of text.

To set up the model for our experiment, we obtained the sector indices and applied a label encoder for labeling them. Next, we divided the dataset into training and testing sets,

employing the widely adopted practice of an 80–20 split. The split results in the W-Train and W-Test datasets.

The model underwent training for two epochs. During testing, it demonstrated an F1 score of 0.80 for classifying the GICS sectors and a score of 0.75 for classifying the industry groups. These results highlight the model's effectiveness in accurately predicting sector classifications. The confusion matrix and classification report are presented in Figure 2b and Table B3. Additionally, we conducted an experiment comparing the performance of the BERT-base-uncased model with that of the RoBERTa-base. The results revealed that both models achieved an F1 score of 0.80, indicating comparable performance between the two.

5.3. One-vs-Rest Classification

One-vs-Rest (OvR) classification is a valuable technique that utilizes binary classification algorithms to address multi-class classification problems. In a multi-class classification scenario where there are more than two classes, the objective is to predict the appropriate class label for each instance accurately. The problem of company classification inherently falls under the domain of multi-class classification, making the One-vs-Rest approach an appropriate strategy for tackling this task. This approach involves transforming the multi-class problem into multiple binary classification subproblems, where each class is distinguished from all the other classes combined using a separate binary classifier. This approach assumes that each class is independent of the others and the decision boundaries between classes are mutually exclusive.

5.3.1. One-vs-Rest Classifier Using SVC Estimator with RBF and Cosine Similarity Kernel

We assessed the One-vs-Rest classification using the support vector classifier (SVC) as an estimator in two scenarios: one employed the default radial basis function (RBF) kernel, and the other utilized the cosine similarity function as the kernel. We initially labeled the indices for various sectors and subsequently divided them into training and testing sets using an 80–20 split, resulting in the W-Train and W-Test datasets. Afterward, we performed preprocessing on the company descriptions to eliminate irrelevant information. The preprocessing involves converting the text into lowercase letters, followed by removing text in square brackets, links, punctuation marks, and words containing numbers. The data in the training and testing sets were then embedded using the all-mpnet-base-v2 model (the all-mpnet-base-v2 model was selected as it is listed as the best-performing model in an extensive evaluation of various models, as found at the following link: https://www.sbert.net/docs/pretrained_models.html, accessed on 15 January 2024). These embeddings were then utilized as input for training the OneVsRestClassifier from the scikit-learn library. The OneVsRestClassifier model employed the SVC estimator, utilizing the default RBF kernel and default number of iterations. The model demonstrated its effectiveness in the classification task by achieving an F1 score of 0.80. The confusion matrix and classification report are presented in Figure 2c and Table B4, respectively. We also tried the same approach with one slight difference. Instead of the RBF kernel, we used the SVC estimator with a cosine similarity function as its kernel and the default number of iterations. The F1 score obtained for this model was 0.78, slightly smaller than the F1 score observed in the previous experiment.

5.3.2. One-vs-Rest Classifier Using Dimensionality Reduction with Principal Component Analysis (PCA) and Autoencoder Architecture

We proceed by evaluating the One-vs-Rest classification in combination with dimensionality reduction using principal component analysis (PCA) and autoencoder architecture. To prepare the data for this experiment, we followed a slightly different approach. We first divided the WRDS dataset into training and testing sets, employing an 80–20 split. The training and testing sets, W-Train and W-Test, were encoded using the all-mpnet-base-v2 transformer. This transformer belongs to the class of sentence transformers and is capable of mapping sentences and paragraphs into a high-dimensional dense vector space of 768 dimen-

sions (more details about the all-mpnet-base-v2 transformer can be found on Hugging Face at the following link: https://huggingface.co/sentence-transformers/all-mpnet-base-v2, accessed on 15 January 2024).

Since the embeddings are vectors in a high-dimensional dense vector space, we wanted to study the effect of dimensionality reduction on the classification performance. We applied PCA to reduce the dimensionality of the embeddings from 768 to 100. However, the resulting model yielded a significantly lower F1 score of 0.42. These findings suggest that the reduction in dimensionality with PCA had an adverse effect on the model's performance.

To decrease the dimensionality of the vectors, we utilized an additional method. We used an autoencoder architecture comprising five stacked layers with the following neuron configurations: 768, 256, 100, 256, and 768. Both the input and output layers consist of 768 neurons, which aligns with the dimensionality of the vector obtained from the company description after encoding with the all-mpnet-base-v2 transformer.

Our objective was to employ the W-Train dataset to train the autoencoder and subsequently utilize the trained autoencoder to reduce the dimensionality of the W-Test dataset to 100 dimensions through its middle (compressed) layer. This procedure effectively eliminates the unnecessary dimensions, leading to dimensionality reduction. After several iterations, we found that the autoencoder achieved optimal results with ten epochs and a batch size of 32. After applying the trained autoencoder to the W-Test dataset, we reduced its dimensions and created a new compressed version of the set called W-Test-Compressed.

After training the autoencoder, the next step was to train the One-vs-Rest model and evaluate its performance. To achieve that, we utilized the newly created W-Test-Compressed dataset along with the original, unused W-Test dataset, combining them together. On this combined dataset, denoted as W-AE-Full in Table 4, we performed a new 80–20 train and test split, resulting in the W-AE-Train and W-AE-Test datasets. It is important to note that the W-Test set was not encoded using the sentence transformer technique.

Finally, we employed the OneVsRestClassifier, utilizing the support vector classifier (SVC) model with default RBF kernel and default number of iterations. The model was trained on the W-AE-Train and evaluated on the W-AE-Test. Through the approach, we achieved an F1 score of 0.78, indicating the model's effectiveness in predicting the classification of company descriptions. The confusion matrix and classification report are presented in Figure 2d and Table B5.

5.3.3. Using K-Fold to Clean the WRDS Dataset from the Incorrectly Predicted Descriptions

Our analysis revealed that the WRDS dataset contains a substantial amount of noise. To address this issue, we utilized the K-Fold approach, which involves splitting the WRDS dataset into two parts using an 80–20 split. The smaller portion, representing 20% of the dataset, was set aside as a test set for future models. The test set is denoted as W-Test in Table 4.

Next, we employed the all-mpnet-base-v2 model to embed the remaining 80% of the dataset, referred to as W-Train. To clean the data and enhance data quality, we applied the K-Fold method by dividing the embedded dataset into five equal parts. Four parts were designated as training sets, while the remaining fifth part served as a testing set. This process was repeated five times to ensure that all possible combinations of training and testing sets were evaluated.

For classification, we utilized the One-vs-Rest approach with the SVC estimator, employing the default RBF kernel and the number of iterations, as in the previous experiments. In each iteration, we identified descriptions that were misclassified by the model and removed them from W-Train. This resulted in a new dataset, which was saved and used as a training set for future models; this dataset is named W-Train-C to indicate that it underwent the K-Fold cleaning process. By following this approach, we successfully reduced the noise present in the WRDS dataset.

5.3.4. One-vs-Rest Evaluated on the WRDS Dataset Cleaned Using K-Fold

In the experiment, we first used 80% of the WRDS dataset (cleaned using K-Fold), designating it as a train set (W-Train-C). The remaining 20% of the dataset left uncleaned and immediately saved served as the test set (W-Test). This test set was also subjected to the same sentence transformer, namely the all-mpnet-base-v2 model, for embedding its content. Subsequently, the One-vs-Rest model was then tested on this test set, resulting in a high F1 score of 0.78. The confusion matrix and classification report are presented in Figure 2e and Table B6. To further assess the performance of our model, we tested the model on a Kaggle dataset comprising company descriptions and their corresponding classifications as per the GICS taxonomy. Remarkably, the model attained an F1 score of 0.85 on this test set, indicating its efficacy in accurately predicting the sectors of companies.

5.3.5. One-vs-Rest Evaluated on the WRDS Dataset with Omitted Company Names

We performed an additional experiment using a refined WRDS dataset, denoted as W-Full-R, from which the company names have been omitted. In particular, we made modifications to the WRDS dataset by excluding the company names from the company descriptions. The objective was to mitigate bias in the decision-making process when utilizing various classification models. This approach extends to ChatGPT-based classification, as ChatGPT inherently incorporates information about different companies and their names. Rather than basing its categorization on the provided description, ChatGPT tends to extrapolate its decision from the name itself, influenced by prior knowledge of the company. This can lead to suboptimal classifications rather than accurately attributing it to its relevant GICS sector. The remaining aspects of the setup mirrored the previous experiments, including labeling indices for different sectors, partitioning the dataset into an 80–20 split of train and test sets (represented by the W-Train-R and W-Test-R datasets, respectively), and preprocessing the company descriptions to remove irrelevant information. Using the One-vs-Rest classifier, we achieved an F1 score of 0.78.

5.3.6. One-vs-Rest Using Contextual Sentence Transformer

Finally, we explore the use of advanced natural language processing techniques, experimenting with a new state-of-the-art contextual sentence transformer. Specifically, we employed the hkunlp/instructor-large sentence transformer to generate embeddings for each company description, taking into account the relevant context based on the specific problem at hand. The instructor model represents a novel approach for computing text embeddings in which each textual input is embedded together with instructions explaining the use case, such as task and domain descriptions [58]. The embedded descriptions derived from this model were fed into the One-vs-Rest classifier employing an SVC estimator with the default RBF kernel and the default number of iterations. We examined various contexts to identify the most effective strategy. Yet, even with the use of this sophisticated model, the highest F1 score that we achieved was 0.79. The confusion matrix and classification report are presented in Figure 2f and Table B7. Interestingly, we observed that using different contexts yielded only minimal differences in the F1 score, indicating that this method exhibits modest improvements.

5.4. ChatGPT-Based Classification

In this approach, we employ a dataset comprising company descriptions generated using ChatGPT, denoted as CG-Full in Table 4. Our primary aim is to assess the impact of ChatGPT-generated content on classification performance because ChatGPT has the ability to provide more detailed company descriptions. To achieve this, we conducted a series of experiments utilizing previously employed techniques, such as zero-shot, multi-class, and One-vs-Rest classifiers. The obtained results are shown in Table 5.

We start by evaluating zero-shot classification on sector names obtained using ChatGPT. Using the zero-shot classification pipeline based on the valhalla/distilbart-mnli-12-3 model, we performed an experiment in which we substituted the GICS sector names with

richer descriptions obtained through ChatGPT. The objective was to enhance the model with a more specific understanding of the sectors. This approach yielded an F1 score of 0.61. Consistent with prior experiments, we employed the complete dataset's descriptions as input rather than dividing the dataset into training and testing sets.

We then proceed by using the One-vs-Rest classification. Following the One-vs-Rest approach described earlier, we conducted an experiment involving the same ChatGPT-generated dataset of company descriptions. This experiment entails partitioning the dataset into training and testing sets using an 80–20 split, resulting in the CG-Train and CG-Test datasets, respectively. The 80–20 split led to the utilization of 176 sentences from the ChatGPT-generated dataset for training (CG-Train dataset), with the remaining 44 sentences allocated for testing purposes (CG-Test dataset). We employed the all-mpnet-base-v2 model to embed the dataset elements. The model achieved an F1 score of 0.52, indicating its poor performance when trained on small datasets. Moreover, we evaluated the performance of this model on the WRDS dataset (W-Full), which was used to assess most of the models in this study. The resulting F1 score of 0.46 further confirms the inadequate performance of this model.

We continue by assessing ChatGPT on 20% of the WRDS dataset that we kept uncleaned after applying K-Fold (W-Test). As indicated previously, we created a model using a train set that comprises 80% of the WRDS dataset cleaned using K-Fold. The remaining 20% of the dataset is not cleaned and is used as a test set. This model achieved an F1 score of 0.78 on the test set; moreover, the model was tested on the Kaggle dataset, achieving an F1 score of 0.85, as shown in Table 7. We performed an additional evaluation by comparing our model's performance with that of ChatGPT in classifying company descriptions using the OpenAI API. Specifically, we employed ChatGPT to make predictions on the W-Test dataset, which comprised the remaining 20% of the uncleaned portion from the WRDS dataset. It is important to acknowledge that ChatGPT possessed prior knowledge of the companies mentioned in the descriptions due to their names, which could have potentially influenced its predictions, introducing a source of bias. Despite this advantage, ChatGPT achieved an F1 score of 0.71, indicating a performance inferior to that of our initial model. This outcome showcases the efficiency of our model in accurately classifying company descriptions, even when compared against a language model such as ChatGPT.

Table 7. Company classification performance using One-vs-Rest and multi-class classifiers applied on a Kaggle dataset that includes company descriptions and their respective classifications, following the GICS taxonomy.

	Datasets		Macro Average			Weighted Average				
	Train	Test	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Support	Final F1 Score
One-vs-Rest classifier based on all-mpnet-base-v2	W-Train-C	Kaggle	0.90	0.84	0.84	0.87	0.85	0.84	158	0.85
Multi-class classifier based on RoBERTa-base	W-Train	Kaggle	0.93	0.85	0.85	0.90	0.87	0.86	158	0.87

Finally, we also evaluated the performance of ChatGPT in the context of the experiment, where we removed the names of the companies from the descriptions as well as the entries consisting only of company names. As indicated previously in Table 5, the use of all-mpnet-base-v2 embeddings for training a One-vs-Rest classifier achieved an F1 score of 0.78. To further assess the performance of our model in comparison to ChatGPT, we analyzed the descriptions that our model misclassified. By removing company names from these descriptions, we aimed to minimize potential bias stemming from ChatGPT's prior knowledge about specific companies. Subsequently, we employed ChatGPT to classify these modified descriptions, yielding a considerably lower F1 score of 0.22. This outcome is documented in Table 5 and suggests that ChatGPT encounters significant challenges in accurately classifying these descriptions.

5.5. Analysis by GICS Sectors

The classification reports given in Appendix B present an overview of the model performance across the GICS sectors.

As shown in Tables B1 and B2, zero-shot classification utilizing sector names enhanced with TF-IDF yields advantages compared to employing the original sector names. The weighted F1 score across the entire dataset is notably higher at 0.64 when utilizing enhanced sector names, surpassing the F1 score of 0.56 achieved with the original sector names. Additionally, the enhancement of sector names leads to an improvement in F1 scores across all sectors. In Table B2, the highest individual F1 scores are obtained for Health Care and Oil & Natural Gas with 0.84 and 0.81 F1 scores, followed by Banking & Lending and Raw Minerals & Mining with 0.77 and 0.75 F1 scores, respectively. The lowest F1 scores are observed for Food, Beverages and Household Products with 0.30 F1 score, Real Estate with 0.39 F1 score, and Industrials and Transportation with 0.39 F1 score.

Tables B3 and B4 illustrate improved F1 scores when employing a multi-class classifier or One-vs-Rest classification when compared to zero-shot classification with enhanced sector names. Notable improvements are observed, especially for those sectors where the zero-shot classifier yields the lowest F1 scores. For example, in the Food, Beverages and Household Products sector, utilizing enhanced sector names achieves an F1 score of 0.30, whereas employing the multi-class classifier and One-vs-Rest classification significantly improves scores to 0.63 and 0.71, respectively. Similarly, in the Real Estate sector, the transition from zero-shot classification (0.39 F1 score) to the multi-class classifier and OvR classification increases the F1 scores to 0.61 and 0.56, respectively. For Industrials, the multi-class and OvR classifiers outperform the zero-shot classifier, achieving F1 scores of 0.73 and 0.63, which are notably higher than the 0.39 F1 score obtained by the latter.

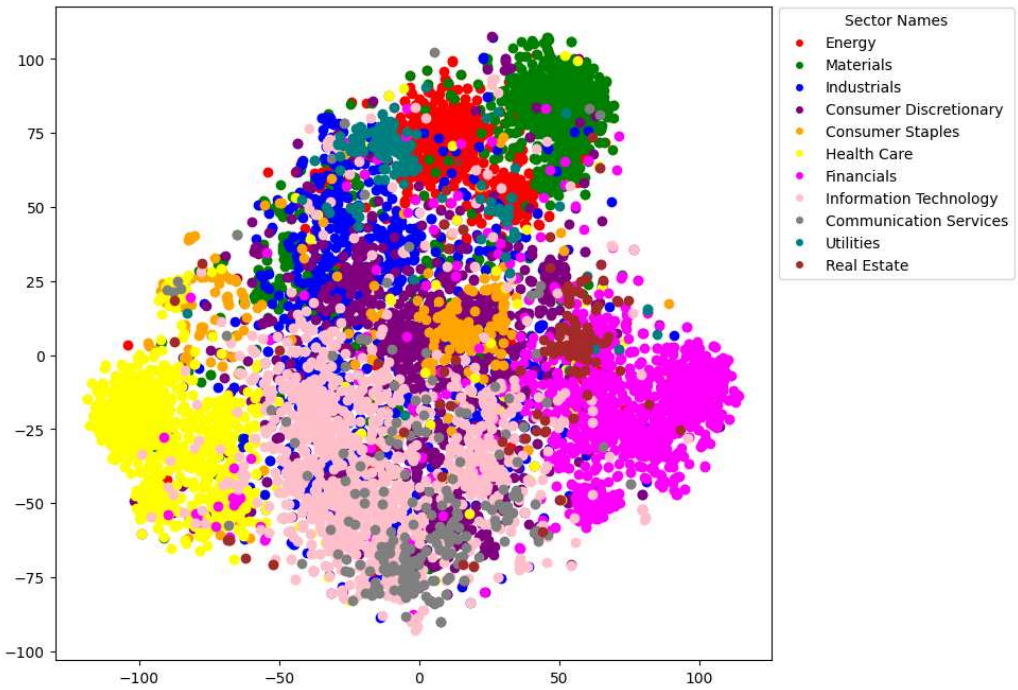
Both the multi-class and One-vs-Rest classifiers demonstrate the highest weighted F1 score of 0.80 across all models and performed experiments. However, there is no clear winner between the two when considering the analysis at the sector level. As seen in Table 8, both models exhibit comparable performance; the multi-class classifier achieves higher F1 scores across six sectors, while the One-vs-Rest classifier excels in five sectors. The multi-class classifier notably outperforms the One-vs-Rest classifier in the Communication Services and Consumer Discretionary sectors. Conversely, the One-vs-Rest classification shows a significant advantage in the Utilities and Health Care sectors. Across all other sectors, the F1 scores between the two models are comparable.

Table 8. Comparison between the multi-class classifier based on the RoBERTa-base model and the One-vs-Rest classifier regarding F1 scores across the GICS sectors.

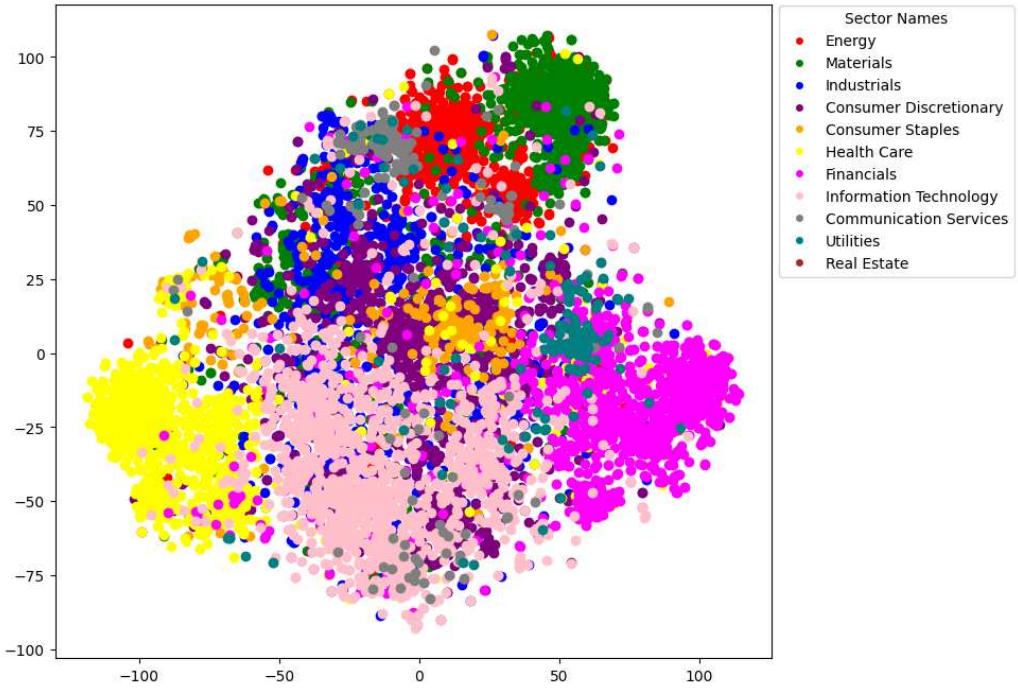
	F1 Scores	
GICS Sectors	**Multi-Class Classifier**	**One-vs-Rest Classifier**
Financials	0.89	**0.90**
Communication Services	**0.85**	0.53
Consumer Staples (Consumer Defensive)	0.63	**0.71**
Health Care	0.73	**0.89**
Industrials	**0.73**	0.63
Consumer Discretionary (Consumer Cyclical)	**0.87**	0.72
Energy	**0.90**	0.87
Materials	0.80	**0.86**
Real Estate	**0.61**	0.56
Information Technology	**0.81**	0.79
Utilities	0.60	**0.79**

Finally, as illustrated in Figure 3, we utilized t-SNE to visually represent the companies within the W-Test dataset and their distribution across GICS sectors. Specifically, Figure 3a represents the W-Test dataset as per the original GICS sectors, while Figure 3b shows the sector predictions made by the OvR model for the companies in the W-Test dataset. Both visualizations reveal a significant degree of overlap, indicating the model's efficiency in

correctly capturing the sector classifications. Figure 4 delineates inaccurately classified instances and presents them according to their original GICS sector affiliation.



(**a**) Based on the original GICS sectors



(**b**) Based on One-vs-Rest model predictions

Figure 3. Visual representation of the companies within the W-Test dataset and their distribution across GICS sectors using t-SNE: (**a**) shows a representation of the W-Test dataset based on the original GICS sectors, while (**b**) shows a representation of the sector predictions made by the One-vs-Rest model for the companies in the W-Test dataset.

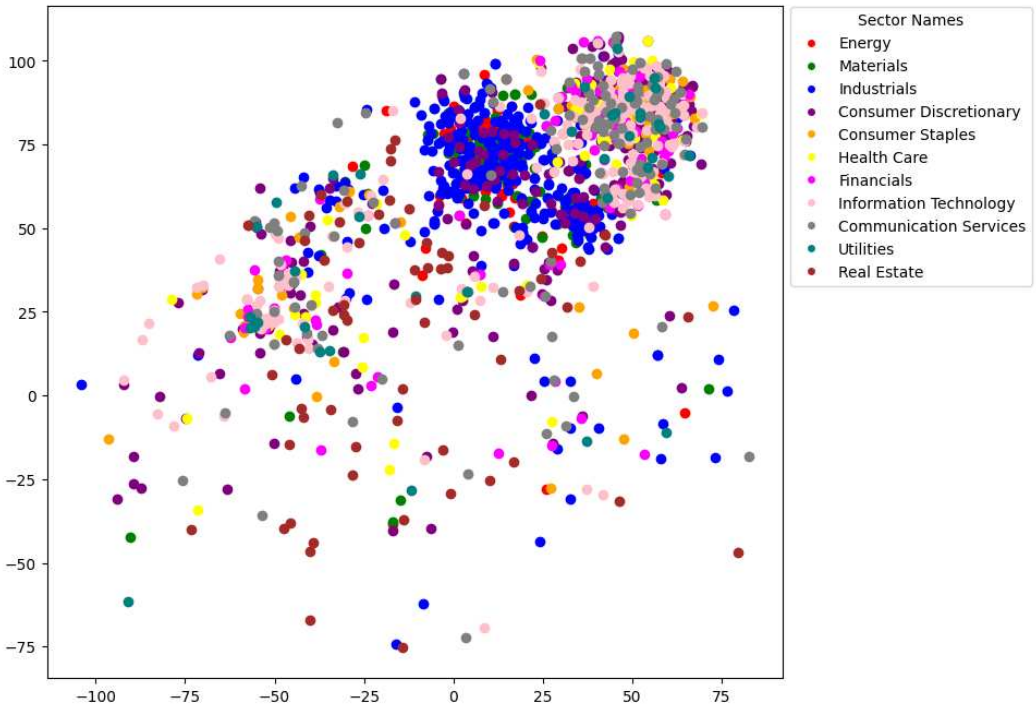


Figure 4. Visual representation of the instances in the W-Test dataset that are inaccurately classified by the One-vs-Rest model. The instances are presented according to their original GICS sector affiliation.

5.6. Discussion and Model Comparison

For zero-shot classification, we utilized the distilbart-mnli-12-3 model and achieved an overall F1 score of 0.56 with the original sector names from the GICS taxonomy. TF-IDF was employed to extract the top 30 most common words for each sector to obtain a more precise representation of the sector names. This technique increased the F1 score of the zero-shot classification to 0.64. Additionally, a multi-class classifier based on RoBERTa-base was applied to the GICS sectors, resulting in a superior F1 score of 0.80—the highest among all experiments conducted. The same F1 performance was attained with One-vs-Rest classification using the all-mpnet-base-v2 model. For One-vs-Rest classification, various techniques were attempted to enhance the F1 results, including changing the kernel of the support vector classifier (SVC), dimensionality reduction with principal component analysis (PCA) and autoencoder architecture, and model training on the WRDS dataset cleaned from company names with K-Fold. However, none of these approaches yielded an F1 score higher than 0.80. We also explored using a state-of-the-art contextual sentence transformer, namely hkunlp/instructor-large. Even with the use of this sophisticated model, the highest F1 score achieved was 0.79.

We leveraged ChatGPT to produce a dataset comprising company descriptions for assessing its influence on classification performance. The replacement of GICS sector names with descriptions generated by ChatGPT resulted in an F1 score of 0.61 for zero-shot classification on the WRDS dataset. Notably, this outcome is inferior to the performance achieved through zero-shot classification using the original GICS sector names. We then proceeded to divide the ChatGPT-generated dataset into training and testing sets utilizing an 80–20 split. However, the One-vs-Rest classification performed poorly in this scenario, obtaining an F1 score of 0.52. Subsequent evaluation of this model on the WRDS dataset yielded an even lower F1 score of 0.46, thereby reinforcing the evident shortcomings of this model. Employing ChatGPT in combination with OpenAI's API, we made predictions on a test set representing 20% of the WRDS dataset, which had been kept unaltered after the application of K-Fold. Despite ChatGPT's inherent advantage of possessing prior

knowledge about the companies, ChatGPT achieved an F1 score of 0.71, indicating a performance inferior to that of the initial model.

The approach in this study centers around the utilization of zero-shot, multi-class, and One-vs-Rest classifiers based on state-of-the-art language models. Additionally, we incorporate a novel direction by leveraging a ChatGPT-based company classification method. While previous studies have explored similar avenues, it is important to note that they employ diverse datasets. The absence of a universally acknowledged "gold standard" benchmark for evaluating company classification leads to difficulties related to the need for a comprehensive exploration of various datasets. Certain prior studies employ proprietary or country-specific datasets and, in some cases, classifications that are not aligned with the established GICS standard. Our study utilizes a range of models, and importantly, we leverage the WRDS dataset. We consider the incorporation of WRDS important, as it is widely recognized by academic researchers, corporate professionals, and financial analysts as a premier data management platform for business-related projects. This choice enhances the robustness and completeness of our study.

Using pretrained models without additional fine-tuning, such as those relying on zero-shot classification, can be advantageous in use cases where the speed of inference is an essential factor. Once the pretraining phase concludes, these models can be readily used in a wide range of downstream tasks, including company classification. However, as is evident in Table 5, zero-shot classifiers may yield suboptimal results compared to other models. In particular, multi-class and One-vs-Rest classifiers outperform zero-shot classification in all classification metrics, with a notable difference in terms of accuracy and F1 scores. This positions them as optimal choices for production systems that prioritize the quality of results over speed. Additionally, employing ChatGPT for company classification is a viable alternative in scenarios lacking dedicated datasets of company descriptions. However, the drawback of this approach lies in the inferior classification results. In essence, the selection of the model should be aligned with the specific system requirements, while also ensuring the right balance between accuracy and speed.

6. Conclusions

In this paper, we have explored the potential of various NLP-based models for company classification using the Wharton Research Data Services (WRDS) dataset. Our study aims to address the limitations of traditional classification standards, such as the Standard Industrial Classification (SIC), the North American Industry Classification System (NAICS), the Fama French (FF) system, and the Global Industry Classification Standard (GICS). These standards suffer from several important drawbacks. They are based on time-consuming, effort-intensive, and vendor-specific assignments by domain experts, leading to issues with accuracy, cost, lack of standardization, and timely updates to address the dynamic changes in the company landscape.

Addressing these issues requires a move towards automated, standardized, and continuously updated classification approaches that are efficient and cost-effective and also consider the evolving nature of businesses and industries. Thus, we have investigated the application of machine learning (ML) and natural language processing (NLP) methods in this domain. We performed a comparative analysis with experiments involving deep learning techniques, such as zero-shot learning, multi-class, One-vs-Rest classification, and ChatGPT-aided classification on the WRDS dataset. We have evaluated the performance of these techniques using standard classification metrics, such as precision, recall, and F1 score. The results of our experiments demonstrated the potential of the studied techniques for company classification. We observed that the NLP-based models achieve promising performance, indicating that these models can effectively automate the process of company classification, reducing costs, complexity, and manual labor.

One potential limitation of the NLP-based approaches lies in the definitions established by the standard used for company classification. Outdated definitions may not align with the current business landscape, impacting the relevance of the model predictions.

Additionally, changes in company descriptions or the emergence of new companies can pose challenges to the dataset coverage and accuracy. Thus, it is important to recognize the necessity for timely updates of the standards or datasets to maintain the effectiveness of production systems used for automatic company classification.

Furthermore, since ChatGPT has gained considerable popularity, it is worth mentioning that it exhibits potential limitations with speed and inherent bias. ChatGPT is relatively slower compared to other methods as it requires time to perform inference and generate the results. Additionally, it is inherently biased as it possesses prior knowledge about the companies, i.e., their names and descriptions. Despite this advantage, ChatGPT exhibited inferior classification results. This highlights the need for a nuanced consideration of the trade-offs associated with the use of ChatGPT and other NLP-based methods for company classification.

One promising avenue for future research involves conducting experiments with an increased dataset size or exploring the use of a larger, more performant model than RoBERTa. Improving the dataset quality is a challenging endeavor, often requiring effort-intensive data curation efforts. Additionally, investigating the potential benefits of additional fine-tuning can serve as a valuable direction for future work, as fine-tuning could enhance the model capabilities and performance on the company classification task.

The results show that the NLP-based methods hold the potential for automating the task of company classification, which can benefit various industries, including finance, marketing, and business intelligence, by providing a more efficient and cost-effective way of categorizing companies. It can also help in identifying emerging trends and patterns in the business world, which can be valuable for decision-making processes. While our study has focused on the WRDS dataset, future research could explore the application of these techniques on larger and more diverse datasets to validate their effectiveness further.

Author Contributions: Conceptualization, I.M., M.R. and D.T.; methodology, D.T., I.M. and M.R.; software, A.J., M.R., V.S. and I.M.; validation, M.R., A.J., D.T. and I.M.; investigation, M.R.; data curation, I.M., E.P., V.S. and A.J.; writing—original draft preparation, M.R.; writing—review and editing, M.R. and D.T.; visualization, M.R. and A.J.; supervision, D.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code is openly accessible on GitHub at https://github.com/nubs4dayz/company-classification-research (accessed on 15 January 2024). The WRDS dataset can be obtained from the WRDS research platform. Information about the remaining datasets is documented in the paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Randomly partitioned GICS industry groups. A division of GICS industries into three distinct lists, each comprising 23 sectors, to facilitate the zero-shot classification using the valhalla/distilbart-mnli-12-3 model. This random partitioning ensures diverse representation for robust model evaluation. The industries in each list are sorted in alphabetical order.

Part 1	Part 2	Part 3
Auto Components	Aerospace & Defense	Airlines
Beverages	Air Freight & Logistics	Banks
Capital Markets	Automobiles	Biotechnology
Construction & Engineering	Containers & Packaging	Building Products
Construction Materials	Distributors	Chemicals
Diversified Consumer Services	Diversified Financial Services	Commercial Services & Supplies
Diversified Telecommunication Services	Electrical Equipment	Communications Equipment
Energy Equipment & Services	Electronic Equipment, Instruments & Components	Consumer Finance
Entertainment	Equity Real Estate Investment Trusts (REITs)	Electric Utilities
Gas Utilities	Food Products	Food & Staples Retailing
Health Care Providers & Services	Health Care Technology	Health Care Equipment & Supplies
Household Products	Interactive Media & Services	Hotels, Restaurants & Leisure
IT Services	Life Sciences Tools & Services	Household Durables
Independent Power and Renewable Electricity Producers	Mortgage Real Estate Investment Trusts (REITs)	Industrial Conglomerates
Leisure Products	Multi-Utilities	Insurance
Machinery	Oil, Gas & Consumable Fuels	Internet & Direct Marketing Retail
Marine	Personal Products	Media
Multiline Retail	Road & Rail	Metals & Mining
Paper & Forest Products	Software	Real Estate Management & Development
Pharmaceuticals	Specialty Retail	Semiconductors & Semiconductor Equipment
Professional Services	Textiles, Apparel & Luxury Goods	Tobacco
Technology Hardware, Storage & Peripherals	Trading Companies & Distributors	Transportation Infrastructure
Thrifts & Mortgage Finance	Wireless Telecommunication Services	Water Utilities

Appendix B

Table B1. Classification report for the valhalla/distilbart-mnli-12-3 model on the WRDS dataset with original sector names.

	Precision	Recall	F1-Score	Support
Financials	0.68	0.61	0.64	5363
Communication Services	0.32	0.63	0.42	1285
Consumer Staples (Consumer Defensive)	0.20	0.01	0.02	1433
Health Care	0.83	0.84	0.83	4565
Industrials	0.42	0.20	0.27	3934
Consumer Discretionary (Consumer Cyclical)	0.41	0.46	0.43	4662
Energy	0.56	0.91	0.69	2822
Materials	0.54	0.65	0.59	3833
Real Estate	0.29	0.89	0.44	509
Information Technology	0.71	0.54	0.61	5192
Utilities	0.44	0.25	0.32	740
accuracy			0.56	34,338
macro avg	0.49	0.54	0.48	34,338
weighted avg	0.57	0.56	0.55	34,338

Table B2. Classification report for the valhalla/distilbart-mnli-12-3 model on the WRDS dataset with enhanced sector names.

	Precision	Recall	F1-Score	Support
Banking and Lending	0.84	0.71	0.77	5363
Communications, Telecommunications, Networking, Media and Entertainment	0.39	0.62	0.48	1285
Food, Beverages and Household Products	0.51	0.21	0.30	1433
Health Care	0.80	0.89	0.84	4565
Industrials and Transportation	0.57	0.29	0.39	3934
Non-Essential Goods, Retail and E-Commerce	0.44	0.54	0.48	4662
Oil, Natural Gas, Consumable Fuels and Petroleum	0.86	0.76	0.81	2822
Raw Materials, Mining, Minerals and Metals (Gold, Silver and Copper)	0.86	0.67	0.75	3833
Real Estate Properties	0.25	0.90	0.39	509
Software, Technology and Systems	0.61	0.66	0.63	5192
Utilities, Energy Distribution and Renewable Energy	0.47	0.80	0.59	740
accuracy			0.64	34,338
macro avg	0.60	0.64	0.58	34,338
weighted avg	0.67	0.64	0.64	34,338

Table B3. Classification report for the RoBERTa-base model on the WRDS dataset with GICS sector names.

	Precision	Recall	F1-Score	Support
Financials	0.88	0.89	0.89	545
Communication Services	0.85	0.86	0.85	793
Consumer Staples (Consumer Defensive)	0.65	0.62	0.63	769
Health Care	0.75	0.71	0.73	960
Industrials	0.71	0.75	0.73	260
Consumer Discretionary (Consumer Cyclical)	0.85	0.91	0.87	878
Energy	0.89	0.91	0.90	1058
Materials	0.80	0.80	0.80	1106
Real Estate	0.61	0.62	0.61	236
Information Technology	0.77	0.84	0.81	150
Utilities	0.77	0.50	0.60	113
accuracy			0.80	6868
macro avg	0.77	0.76	0.77	6868
weighted avg	0.80	0.80	0.80	6868

Table B4. Classification report for the approach using One-vs-Rest classification.

	Precision	Recall	F1-Score	Support
Financials	0.88	0.91	0.90	1058
Communication Services	0.62	0.46	0.53	236
Consumer Staples (Consumer Defensive)	0.70	0.73	0.71	260
Health Care	0.86	0.93	0.89	878
Industrials	0.69	0.59	0.63	769
Consumer Discretionary (Consumer Cyclical)	0.72	0.72	0.72	960
Energy	0.85	0.88	0.87	545
Materials	0.85	0.86	0.86	793
Real Estate	0.75	0.44	0.56	113
Information Technology	0.76	0.83	0.79	1106
Utilities	0.85	0.74	0.79	150
accuracy			0.80	6868
macro avg	0.78	0.74	0.75	6868
weighted avg	0.79	0.80	0.79	6868

Table B5. Classification report for the approach using autoencoder for dimensionality reduction.

	Precision	Recall	F1-Score	Support
Financials	0.58	0.48	0.53	131
Communication Services	0.88	0.83	0.85	183
Consumer Staples (Consumer Defensive)	0.88	0.91	0.89	214
Health Care	0.85	0.94	0.89	171
Industrials	0.72	0.65	0.68	48
Consumer Discretionary (Consumer Cyclical)	0.82	0.90	0.86	104
Energy	0.53	0.36	0.43	50
Materials	0.69	0.71	0.70	196
Real Estate	0.75	0.73	0.74	33
Information Technology	0.75	0.83	0.78	224
Utilities	0.73	0.40	0.52	20
accuracy			0.78	1374
macro avg	0.74	0.70	0.72	1374
weighted avg	0.77	0.78	0.77	1374

Table B6. Classification report for the approach using One-vs-Rest with K-Fold applied on 80% of the WRDS dataset.

	Precision	Recall	F1-Score	Support
Financials	0.86	0.91	0.89	1058
Communication Services	0.71	0.36	0.48	236
Consumer Staples (Consumer Defensive)	0.70	0.71	0.70	260
Health Care	0.86	0.93	0.89	878
Industrials	0.68	0.56	0.61	769
Consumer Discretionary (Consumer Cyclical)	0.70	0.70	0.70	960
Energy	0.84	0.87	0.86	545
Materials	0.83	0.85	0.84	793
Real Estate	0.82	0.33	0.47	113
Information Technology	0.73	0.84	0.78	1106
Utilities	0.80	0.69	0.74	150
accuracy			0.78	6868
macro avg	0.78	0.70	0.72	6868
weighted avg	0.78	0.78	0.77	6868

Table B7. Classification report for the approach using One-vs-Rest with the instructor-large model.

	Precision	Recall	F1-Score	Support
Financials	0.86	0.92	0.89	1058
Communication Services	0.69	0.45	0.55	236
Consumer Staples (Consumer Defensive)	0.72	0.74	0.73	0.73
Health Care	0.84	0.93	0.88	878
Industrials	0.68	0.57	0.62	769
Consumer Discretionary (Consumer Cyclical)	0.74	0.73	0.73	960
Energy	0.84	0.85	0.84	545
Materials	0.84	0.86	0.85	793
Real Estate	0.82	0.29	0.43	113
Information Technology	0.76	0.84	0.80	1106
Utilities	0.79	0.73	0.76	150
accuracy			0.79	6868
macro avg	0.78	0.72	0.73	6868
weighted avg	0.79	0.79	0.78	6868

References

1. Ozbayoglu, A.M.; Gudelek, M.U.; Sezer, O.B. Deep learning for financial applications: A survey. *Appl. Soft Comput.* **2020**, *93*, 106384. [CrossRef]
2. Goodell, J.W.; Kumar, S.; Lim, W.M.; Pattnaik, D. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *J. Behav. Exp. Financ.* **2021**, *32*, 100577. [CrossRef]
3. Kumar, S.; Sharma, D.; Rao, S.; Lim, W.M.; Mangla, S.K. Past, present, and future of sustainable finance: Insights from big data analytics through machine learning of scholarly research. *Ann. Oper. Res.* **2022**, 1–44.
4. Kraus, M.; Feuerriegel, S.; Oztekin, A. Deep learning in business analytics and operations research: Models, applications and managerial implications. *Eur. J. Oper. Res.* **2020**, *281*, 628–641. [CrossRef]
5. Delen, D.; Ram, S. Research challenges and opportunities in business analytics. *J. Bus. Anal.* **2018**, *1*, 2–12. [CrossRef]
6. Ajah, I.A.; Nweke, H.F. Big data and business analytics: Trends, platforms, success factors and applications. *Big Data Cogn. Comput.* **2019**, *3*, 32. [CrossRef]
7. Zhang, J.Z.; Srivastava, P.R.; Sharma, D.; Eachempati, P. Big data analytics and machine learning: A retrospective overview and bibliometric analysis. *Expert Syst. Appl.* **2021**, *184*, 115561. [CrossRef]
8. Lin, W.Y.; Hu, Y.H.; Tsai, C.F. Machine learning in financial crisis prediction: A survey. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2011**, *42*, 421–436.
9. Chen, N.; Ribeiro, B.; Chen, A. Financial credit risk assessment: A recent review. *Artif. Intell. Rev.* **2016**, *45*, 1–23. [CrossRef]
10. Bhatore, S.; Mohan, L.; Reddy, Y.R. Machine learning techniques for credit risk evaluation: A systematic literature review. *J. Bank. Financ. Technol.* **2020**, *4*, 111–138. [CrossRef]
11. Nassirtoussi, A.K.; Aghabozorgi, S.; Wah, T.Y.; Ngo, D.C.L. Text mining for market prediction: A systematic review. *Expert Syst. Appl.* **2014**, *41*, 7653–7670. [CrossRef]
12. Nti, I.K.; Adekoya, A.F.; Weyori, B.A. A systematic review of fundamental and technical analysis of stock market predictions. *Artif. Intell. Rev.* **2020**, *53*, 3007–3057. [CrossRef]
13. Kumbure, M.M.; Lohrmann, C.; Luukka, P.; Porras, J. Machine learning techniques and data for stock market forecasting: A literature review. *Expert Syst. Appl.* **2022**, *197*, 116659. [CrossRef]
14. Sohangir, S.; Wang, D.; Pomeranets, A.; Khoshgoftaar, T.M. Big Data: Deep Learning for financial sentiment analysis. *J. Big Data* **2018**, *5*, 1–25. [CrossRef]
15. Araci, D. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv* **2019**, arXiv:1908.10063.
16. Mishev, K.; Gjorgjevikj, A.; Vodenska, I.; Chitkushev, L.T.; Trajanov, D. Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access* **2020**, *8*, 131662–131682. [CrossRef]
17. Rizinski, M.; Peshov, H.; Mishev, K.; Jovanovik, M.; Trajanov, D. Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (XLex). *IEEE Access* **2024**, *12*, 7170–7198. [CrossRef]
18. Bhojraj, S.; Lee, C.M.; Oler, D.K. What's my line? A comparison of industry classification schemes for capital market research. *J. Account. Res.* **2003**, *41*, 745–774. [CrossRef]
19. Lyocsa, S.; Vyrost, T. Industry Classification: Review, Hurdles and Methodologies: Hurdles and Methodologies (30 September 2009). 2009. Available online: https://ssrn.com/abstract=1480563 (accessed on 15 January 2024).
20. Chan, L.K.; Lakonishok, J.; Swaminathan, B. Industry classifications and return comovement. *Financ. Anal. J.* **2007**, *63*, 56–70. [CrossRef]
21. Wood, S.; Muthyala, R.; Jin, Y.; Qin, Y.; Rukadikar, N.; Rai, A.; Gao, H. Automated industry classification with deep learning. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; pp. 122–129.
22. Porter, M.E.; Strategy, C. *Techniques for Analyzing Industries and Competitors*; The Free Press USA: New York, NY, USA, 1980.
23. Phillips, R.L.; Ormsby, R. Industry classification schemes: An analysis and review. *J. Bus. Financ. Librariansh.* **2016**, *21*, 1–25. [CrossRef]
24. Yang, H.; Lee, H.J.; Cho, S.; Cho, E. Automatic classification of securities using hierarchical clustering of the 10-Ks. In Proceedings of the 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016; pp. 3936–3943.
25. Lamby, M.; Isemann, D. Classifying companies by industry using word embeddings. In Proceedings of the International Conference on Applications of Natural Language to Information Systems, Paris, France, 13–15 June 2018; Springer: Cham, Switzerland, 2018; pp. 377–388.
26. Fama, E.F.; French, K.R. Industry costs of equity. *J. Financ. Econ.* **1997**, *43*, 153–193. [CrossRef]
27. Kile, C.O.; Phillips, M.E. Using industry classification codes to sample high-technology firms: Analysis and recommendations. *J. Account. Audit. Financ.* **2009**, *24*, 35–58. [CrossRef]
28. Hrazdil, K.; Zhang, R. The importance of industry classification in estimating concentration ratios. *Econ. Lett.* **2012**, *114*, 224–227. [CrossRef]
29. Boni, L.; Womack, K.L. Analysts, industries, and price momentum. *J. Financ. Quant. Anal.* **2006**, *41*, 85–109. [CrossRef]
30. Hrazdil, K.; Trottier, K.; Zhang, R. A comparison of industry classification schemes: A large sample study. *Econ. Lett.* **2013**, *118*, 77–80. [CrossRef]

31. Slavov, S.; Tagarev, A.; Tulechki, N.; Boytcheva, S. Company Industry Classification with Neural and Attention-Based Learning Models. In Proceedings of the 2019 Big Data, Knowledge and Control Systems Engineering (BdKCSE), Sofia, Bulgaria, 21–22 November 2019; pp. 1–7.
32. Kahle, K.M.; Walkling, R.A. The impact of industry classifications on financial research. *J. Financ. Quant. Anal.* **1996**, *31*, 309–335. [CrossRef]
33. Katselas, D.; Sidhu, B.K.; Yu, C. Know your industry: The implications of using static GICS classifications in financial research. *Account. Financ.* **2019**, *59*, 1131–1162. [CrossRef]
34. Tagarev, A.; Tulechki, N.; Boytcheva, S. Comparison of Machine Learning Approaches for Industry Classification Based on Textual Descriptions of Companies. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), Varna, Bulgaria, 2–4 September 2019; pp. 1169–1175.
35. He, J.; Chen, K. Exploring Machine Learning Techniques for Text-Based Industry Classification. 2020. Available online: https://ssrn.com/abstract=3640205 (accessed on 15 January 2024).
36. Wu, D.; Wang, Q.; Olson, D.L. Industry classification based on supply chain network information using Graph Neural Networks. *Appl. Soft Comput.* **2023**, *132*, 109849. [CrossRef]
37. Ito, T.; Camacho-Collados, J.; Sakaji, H.; Schockaert, S. Learning company embeddings from annual reports for fine-grained industry characterization. In Proceedings of the Second Workshop on Financial Technology and Natural Language Processing, Kyoto, Japan, 5 January 2020; pp. 27–33.
38. Wang, S.; Pan, Y.; Xu, Z.; Hu, B.; Wang, X. Enriching BERT with Knowledge Graph Embedding for Industry Classification. In Proceedings of the International Conference on Neural Information Processing, Sanur, Bali, Indonesia, 8–12 December 2021; Springer: Cham, Switzerland, 2021; pp. 709–717.
39. Dolphin, R.; Smyth, B.; Dong, R. A Machine Learning Approach to Industry Classification in Financial Markets. In Proceedings of the Irish Conference on Artificial Intelligence and Cognitive Science, Munster, Ireland, 8–9 December 2022; Springer: Cham, Switzerland, 2022; pp. 81–94.
40. Zhao, X.; Fang, X.; He, J.; Huang, L. Exploiting Expert Knowledge for Assigning Firms to Industries: A Novel Deep Learning Method. *arXiv* **2022**, arXiv:2209.05943.
41. Husmann, S.; Shivarova, A.; Steinert, R. Company classification using machine learning. *Expert Syst. Appl.* **2022**, *195*, 116598. [CrossRef]
42. Kim, D.; Kang, H.G.; Bae, K.; Jeon, S. An artificial intelligence-enabled industry classification and its interpretation. *Internet Res.* **2021**, *32*, 406–424. [CrossRef]
43. Bernstein, A.; Clearwater, S.; Provost, F. The relational vector-space model and industry classification. In Proceedings of the Learning Statistical Models from Relational Data Workshop at the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI), Acapulco, Mexico, 9–15 August 2003.
44. Drury, B.; Almeida, J.J. Identification, extraction and population of collective named entities from business news. In Proceedings of the Entity 2010—Workshop on Resources and Evaluation for Entity Resolution and Entity Management, Valletta, Malta, 22 May 2010.
45. Gerling, C. Company2Vec—German Company Embeddings based on Corporate Websites. *arXiv* **2023**, arXiv:2307.09332.
46. Vamvourellis, D.; Toth, M.; Bhagat, S.; Desai, D.; Mehta, D.; Pasquali, S. Company Similarity using Large Language Models. *arXiv* **2023**, arXiv:2308.08031.
47. de Carvalho, A.C.; Freitas, A.A. A tutorial on multi-label classification techniques. In *Foundations of Computational Intelligence Volume 5: Function Approximation and Classification*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 177–195.
48. Khan, S.S.; Madden, M.G. One-class classification: Taxonomy of study and review of techniques. *Knowl. Eng. Rev.* **2014**, *29*, 345–374. [CrossRef]
49. Mirończuk, M.M.; Protasiewicz, J. A recent overview of the state-of-the-art elements of text classification. *Expert Syst. Appl.* **2018**, *106*, 36–54. [CrossRef]
50. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. *Information* **2019**, *10*, 150. [CrossRef]
51. Tanha, J.; Abdi, Y.; Samadi, N.; Razzaghi, N.; Asadpour, M. Boosting methods for multi-class imbalanced data classification: An experimental review. *J. Big Data* **2020**, *7*, 70. [CrossRef]
52. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA; pp. 38–45.
53. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
54. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems, Proceedings of the NIPS 2017, Long Beach, CA, USA, 4–9 December 2017*; Curran Associates Inc.: Red Hook, NY, USA; Volume 30.
55. Pushp, P.K.; Srivastava, M.M. Train once, test anywhere: Zero-shot learning for text classification. *arXiv* **2017**, arXiv:1712.05972.
56. Rizinski, M.; Jankov, A.; Sankaradas, V.; Pinsky, E.; Miskovski, I.; Trajanov, D. Company classification using zero-shot learning. *arXiv* **2023**, arXiv:2305.01028.

57. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
58. Su, H.; Shi, W.; Kasai, J.; Wang, Y.; Hu, Y.; Ostendorf, M.; Yih, W.t.; Smith, N.A.; Zettlemoyer, L.; Yu, T. One Embedder, Any Task: Instruction-Finetuned Text Embeddings. *arXiv* **2022**, arXiv:2212.09741.