Essays on Sentiment Analysis in Finance

ICMA Centre

Henley Business School

Thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

Yi Zhu

July 2023

Declaration of Original Authorship

I hereby declare that this doctoral thesis, titled “Essays on Sentiment Analysis in Finance” submitted to the University of Reading, is entirely my original work, conducted under the guidance of my supervisors, Professor Andrew Urquhart, Dr. Tony Moore, and Professor Andreas G. F. Hoepner. I confirm that any external sources of information used in this thesis, including published or unpublished works, have been appropriately cited and referenced.

Yi Zhu

Acknowledgements

First and foremost, I am immensely grateful to my internal supervisors Professor Andrew Urquhart and Dr.Tony Moore, whose invaluable guidance and support have been instrumental in shaping my research and personal growth. Their expertise and insightful feedback have been invaluable throughout my studies. I am also truly grateful for the mentorship from my external supervisor Professor Andreas Hoepner, his sharp mind and willingness to challenge conventional thinking have inspired me to think critically, innovate, and push the boundaries of my technical skills.

I would like to express my appreciation to the ICMA Centre for their financial support and for providing an excellent research environment. I also thank Professor Chris Brooks and Dr.Konstantina Kappou for their time and effort in assessing my work during my PhD confirmation. Their constructive feedback and insightful suggestions have immensely contributed to refining my research. I am also grateful to Dr.Lisa Schopohl, with whom I collaborated. Her devotion, precision, and strong work ethic have been motivating me to strive for excellence in my research.

I extend my deepest appreciation to Dr.Qian Li, a remarkable female researcher whose dedication and passion have been a constant source of inspiration for me. Her belief in my potential propelled me to embark on this transformative journey of discovery. Special gratitude goes to Dr.Linyi Yang, whose passion for research and expertise in Natural Language Processing has deeply influenced my own work.

I would also like to acknowledge the support and companionship of my friends and colleagues XuanFeng Zhou, Meryem Bagriacik, Yu Zhang, Zih-Chun Huang and Jian Chen. Their encouragement, intellectual discussions, and shared experiences

have made this journey both fulfilling and enjoyable. To my dearest friends, Yihang Liao, Chengkai Jin, Pengfei Li, Ruoyang Wei, and Linshao Zhang. Your love, faith, and understanding have been a constant source of strength throughout my PhD journey. Your presence in my life has made the challenges more bearable and the triumphs more meaningful.

Last but certainly not least, I am filled with profound gratitude for the love of my life, my best friend Rhys Woodfield. Thank you for accompanying me from my undergraduate years to this very moment, offering unconditional love, support, and encouragement. You have brought immeasurable joy, strength, and inspiration to my life. Thank you for being a part of this wonderful chapter in my life.

To my parents,
you are my heroes.

Abstract

This thesis explores sentiment analysis in Glassdoor employee reviews, focusing on both English and multilingual contexts. By applying Natural Language Processing (NLP) techniques, we provide a comprehensive review of sentiment analysis in finance, its impact on financial outcomes, and the challenges associated with multilingual sentiment classification.

First, our research investigates the practical deployment and evaluation of various NLP models, ranging from lexicon-based approaches to machine learning models, and to state-of-the-art pre-trained language models. By comparing the performance of various sentiment analysis methods, we demonstrate the superiority of advanced models that consider contextual information. These models can substantially enhance sentiment analysis accuracy when compared to traditional dictionary-based approaches.

Second, our exploration of multilingual sentiment analysis reveals the impact of translation on sentiment classification. We observe the influence of translation on sentiment misclassification rates, with text attributes playing a more significant role than the quality of translation itself. This suggests that even if the translation quality is high, the sentiment expression might be lost during the translation process, thereby driving the sentiment misclassification rate on translated texts. Additionally, our findings highlight the benefits of zero-shot transfer, demonstrating the effectiveness of fine-tuning multilingual language models when labelled multilingual data is limited.

Third, in examining the correlation between employee satisfaction and stock returns, we reveal the predictive power of sentiment measures derived from employee reviews. In our study, we demonstrate that sentiment measures, particularly when derived from BERT, a highly advanced and widely acclaimed language model, effectively predict significant increases in stock returns over both short and long-term periods. BERT, which stands for Bidirectional Encoder Representations from Transformers, is a cutting-edge natural language processing model known for its exceptional contextual understanding of text. Our research also sheds light on the combined influence of employee satisfaction and employee-related costs on stock returns within specific industries. By considering these factors together, we expand the understanding of the complexities underlying the relationship between employee sentiment and financial performance.

This thesis makes a significant contribution to the existing literature. It stands out as the first comprehensive study to undertake a rigorous comparison of 31 sentiment analysis methods, employing the extensive dataset of Glassdoor employee reviews. Moreover, this thesis delves into unexplored territory by examining multilingual sentiment analysis within the context of finance. This novel exploration unveils the challenges and implications associated with sentiment classification in a linguistically diverse landscape. This thesis is also distinctive in its pioneering application of BERT to study the correlation between employee satisfaction and stock returns. The findings from this thesis establish a strong basis for future studies in the areas of sentiment analysis, employee satisfaction, and their impact on finance. The practical significance of our work extends to investors and organisations, enabling them to make informed, data-driven decisions that promote employee well-being and enhance corporate performance.

Contents

List of Figures

List of Tables

Glossary

- **ALBERT**: A Lite BERT - A variation of the **BERT** (Bidirectional Encoder Representations from Transformers) model designed to reduce the number of parameters while maintaining performance in natural language processing tasks.

- **ANMT**: Automatic Neural Machine Translation - A technique that uses neural networks to automatically translate text or speech from one language to another.

- **BERT**: Bidirectional Encoder Representations from Transformers - A pre-trained natural language processing model that learns contextualised word embeddings by training on large corpora of text in both forward and backward directions.

- **BLEU**: Bilingual Evaluation Understudy - A metric used to evaluate the quality of machine translation by comparing the machine-generated translation to one or more human-generated reference translations.

- **BM**: Book-to-Market - A financial ratio that measures the book value of a company's assets relative to the market value of its outstanding shares.

- **BOW**: Bag-of-Words - A simple and commonly used technique in natural language processing that represents text as an unordered collection of words, ignoring grammar and word order.

- **BPE**: Byte-pair Encoding - A subword tokenization technique used in natural language processing to represent words as sequences of subword units based on their frequency of occurrence in a corpus.

- **CBOW**: Continuous Bag of Words Model - A word embedding model in natural language processing that learns to predict a word from its context words, focusing on the continuous representation of words.

- **CNNs**: Convolutional Neural Networks - Deep learning models commonly used in computer vision tasks, such as image classification and object detection.

- **COGS**: Cost of Goods Sold - A financial metric representing the direct costs associated with the production of goods or services that a company sells during a specific period.

- **CRSP**: Center for Research in Securities Prices - A research centre that maintains historical data on stock and bond markets, widely used for financial research and analysis.

- **DistilmBERT**: Distilled mBERT - A compact version of the **BERT** model, designed for efficient natural language processing while maintaining good performance.

- **DL**: Deep Learning - A subfield of machine learning that uses artificial neural networks to model and solve complex problems.

- **DT**: Decision Trees - A machine learning algorithm used for classification and regression tasks, represented as a tree structure of decisions and outcomes.

- **ELMo**: Embeddings from Language Models - Word embeddings generated by pre-trained language models, capturing contextual information in text.

- **ERNIE**: Enhanced Language Representation with Informative Entities - A language representation model designed to improve the quality of word embeddings by incorporating entity information.

- **ES**: Employee Sentiment - An assessment of the emotional and cognitive responses of employees within an organisation.

- **ESG**: Environmental, Social, and Corporate Governance - A framework used to evaluate and measure the sustainability and ethical impact of a company or investment based on its environmental, social, and governance practices.

- **FFC**: Fama-French-Carhart - A multi-factor financial model used to explain and predict stock returns, based on factors like market risk, size, value, and momentum.

- **FN**: False Negative - A classification error in which a true positive is incorrectly classified as negative.

- **FP**: False Positive - A classification error in which a true negative is incorrectly classified as positive.

- **GI**: General Inquirer - A lexicon and computer program for content analysis that identifies words and phrases with particular meanings or sentiments.

- **GICS**: Global Industry Classification Standard - A standardised system for categorising companies into industry groups and sectors.

- **GloVe**: Global Vectors for Word Representation - An unsupervised learning algorithm for obtaining vector representations of words, often used for word embedding in natural language processing tasks.

- **GPT**: Generative Pre-trained Transformer - A family of natural language processing models that use a transformer architecture and are pre-trained on large text corpora to generate human-like text.

- **GRU**: Gated Recurrent Unit - A type of recurrent neural network architecture that is designed to address the vanishing gradient problem and improve the training of sequential data models.

- **HIV4**: Harvard IV-4 Dictionary - A dictionary used for text analysis that provides sentiment scores and linguistic categories for words and phrases.

- **LDA**: Latent Dirichlet Allocation - A generative statistical model used for topic modelling, which helps discover topics within a collection of documents.

- **LIWC**: Linguistic Inquiry and Word Count - A text analysis software program used to analyze the linguistic and psychological content of text, providing insights into language usage and emotional tone.

- **LM**: Loughran-McDonald Dictionary - A specialised dictionary used in financial and textual analysis to classify words and phrases into categories related to financial sentiment.

- **LR**: Logistic Regression - A statistical method used for binary classification and estimating the probability of a binary outcome.

- **LSA**: Latent Semantic Analysis - A technique in natural language processing that analyses relationships between words in a large text corpus to uncover underlying semantic structures.

- **LSTM**: Long Short Term Memory - A type of recurrent neural network architecture designed to capture long-term dependencies in sequential data.

- **MaxEnt**: Maximum Entropy - A probabilistic modelling approach that aims to find the probability distribution with maximum entropy, given certain constraints and prior information.

- **mBERT**: Multilingual BERT - A version of the BERT model that is pretrained on text from multiple languages and can handle multilingual natural language processing tasks.

- **MCC**: Matthews Correlation Coefficient - A measure of the quality of binary and multi-class classifications that takes into account true positives, true negatives, false positives, and false negatives.

- **ML**: Machine Learning - A field of artificial intelligence that focuses on the development of algorithms and models that enable computers to learn from and make predictions or decisions based on data.

- **MLM**: Masked Language Model - A type of language model used in natural language processing where certain words or tokens in a sentence are masked, and the model predicts those masked tokens.

- **MQM**: Multidimensional Quality Metrics - A set of metrics and criteria used to assess the quality of translation through various aspects, such as fluency, adequacy, and terminology.

- **NB**: Naive Bayes - A probabilistic algorithm used for classification and text analysis, particularly for tasks like spam detection and document categorisation.

- **NLP**: Natural Language Processing - A branch of artificial intelligence that focuses on the interaction between computers and human language, including tasks like text analysis and language understanding.

- **NSP**: Next Sentence Prediction - A pre-training task used in some natural language processing models, where the model learns to predict whether a given pair of sentences are consecutive in a text.

- **OOV**: Out-of-vocabulary - Refers to words or tokens that do not appear in the vocabulary or training data of a language model and are, therefore, difficult to handle in natural language processing tasks.

- **POS**: Part-of-speech - The grammatical category to which a word belongs, indicating its function within a sentence (e.g., noun, verb, adjective).

- **R&D**: Expenses, Research and Development - The costs associated with the activities aimed at creating or improving products, processes, or services in the context of innovation.

- **RC**: Restructuring Costs - Expenses incurred by an organisation in the process of reorganising its operations or making significant changes to its structure.

- **RF**: Random Forests - An ensemble machine learning method that combines the predictions of multiple decision trees to improve classification and regression tasks.

- **RNNs**: Recurrent Neural Networks - A class of neural networks designed for processing sequential data, with the ability to maintain a form of memory to handle sequences of arbitrary length.

- **ROA**: Return on Assets - A financial ratio that measures a company's ability to generate earnings from its total assets.

- **RoBERTa**: Robustly Optimised BERT Pretraining Approach - A variation of the BERT model, optimised for improved performance in natural language understanding tasks.

- **ROE**: Return on Equity - A financial ratio that measures a company's profitability relative to its shareholders' equity.

Chapter 1

Introduction

This chapter will introduce the background of sentiment analysis in finance. It aims to explore the various sentiment analysis methods that are popular in the current literature. It will discuss the motivations that drive this thesis, highlighting the current challenges faced within the field and outlining our proposed solutions to address them. Moreover, this chapter will provide an overview and structure of the thesis, effectively summarising the contributions we aim to make.

1.1 Sentiment Analysis in Finance

Sentiment analysis, a branch of NLP, involves extracting subjective information from text data and identifying the emotional state behind it. It covers techniques such as text classification, natural language understanding, and machine learning to categorise text as positive, negative, or neutral. The process often involves analysing large volumes of data from sources such as news articles, social media posts, earnings calls, and financial reports. Each of these sources provides unique insights for market analysis, investment decision-making, risk management, and understanding consumer and employee behaviour.

Sentiment analysis of financial news such as the Wall Street Journal and Dow Jones News can provide timely information on market events, economic indicators, company updates, and geopolitical developments. Positive news sentiment may indicate favourable market conditions, while negative sentiment may signify potential risks or market downturns (Tetlock et al. 2008, Kothari et al. 2009, Engelberg et al. 2012, Huang et al. 2014, Sousa et al. 2019). On the other hand, investigating sentiment of earnings calls and financial reports is crucial for understanding a company's financial performance, future prospects, management outlook, and potential risk factors (Li 2010, Feldman et al. 2010, Rogers et al. 2011, Loughran & McDonald 2011, Tsai & Wang 2017, Yang et al. 2020, Jaggi et al. 2021, Frankel et al. 2022, Huang et al. 2023).

Social media platforms have emerged as valuable sources of real-time information and a reflection of public opinion. Monitoring employee sentiment on social media platforms can significantly contribute to enhancing employee engagement and job satisfaction, leading to sustainable growth for organisations. It has been argued when companies prioritise employee satisfaction, they are more likely to have a productive, motivated, and committed workforce. This, in turn, can positively impact the company's financial performance and lead to higher stock returns (Edmans 2011, Melián-González et al. 2015, Huang et al. 2015, Symitsi et al. 2018, Stamolampros et al. 2019, Corritore et al. 2020, Green et al. 2019). Satisfied employees are more likely to go the extra mile, contribute innovative ideas, and foster a positive work culture, which ultimately translates into improved business outcomes and shareholder value (Hales et al. 2018, Wolter et al. 2019, Huang et al. 2020).

In finance, popular sentiment analysis methods include lexicon-based approaches, machine learning techniques, deep learning models and aspect-based sentiment analysis. Lexicon-based methods rely on pre-defined sentiment dictionaries, while machine learning approaches train models on labelled data. Deep learning and pre-trained language models, such as RNNs, CNNs, BERT and FinBERT can capture complex relationships in textual data. Aspect-based sentiment analysis focuses on extracting sentiment towards specific aspects or entities. The choice of method depends on specific requirements and data characteristics. We will discuss these methods in more detail, along with their implications for finance in the later part of this thesis.

1.2 Motivation of This Thesis

In today's data-driven world, understanding and leveraging sentiment analysis in finance is becoming increasingly important to both businesses and researchers. However, with the multitude of sentiment analysis methods and choices available, there is a pressing need to establish a rigorous framework to guide practitioners and researchers in effectively analysing sentiment. Our research aims to fill this gap by providing a comprehensive framework that enables the practical deployment and evaluation of sentiment analysis methods in the financial domain.

By empirically evaluating a wide range of sentiment analysis methods using Glassdoor employee reviews, we aim to identify the most effective approaches and metrics for assessing sentiment in this context. Our empirical evaluation is facilitated by the Glassdoor employee reviews as they are self-labelled and accompanied by ratings. This unique combination of textual comments and corresponding ratings allows us to evaluate sentiment analysis methods more fairly and comprehensively because we can use the overall rating as a “ground truth” against which to measure the accuracy of the different methods. Therefore, by establishing this framework, we seek to equip practitioners and researchers with reliable tools and methodologies that enable them to make informed decisions based on accurate financial sentiment analysis.

In an increasingly interconnected world, where global markets demand a deep understanding of non-English language markets and perspectives, multilingual sentiment analysis has become a crucial aspect for organisations operating in diverse regions. However, the challenges of comprehending sentiment across different languages remain relatively unexplored within the field of finance and international business. Recognising this gap, our research takes on the challenge of investigating the impact of machine translation on multilingual sentiment analysis using Glassdoor reviews in multiple languages.

We highlight the significant increase in sentiment misclassification rates for translated texts compared to the original texts and illustrate the complexities involved in accurately capturing sentiment across languages. To address these challenges, our research delves into the potential of zero-shot transfer learning, allowing us to make sentiment predictions in languages that were not part of the model's original training data. Zero-shot transfer learning is a machine learning technique in which a model is trained on a specific task, such as sentiment analysis in a particular language, and then applied to related tasks in other languages without any additional training. This means that our model can generalise its understanding of sentiment across multiple languages, enabling us to make accurate sentiment predictions even in languages that were not explicitly included in the initial training process. These findings contribute to overcoming the scarcity of labelled data in diverse languages and pave the way for effective multilingual sentiment analysis in finance.

In addition, the relationship between employee satisfaction and a firm's financial performance has long been of interest to researchers. However, previous studies in this area have primarily relied on using overall ratings as a measure of employee sentiment. In our research, we innovate by examining the text comments provided by employees in Glassdoor reviews, allowing us to gain more nuanced insights into their sentiments and experiences. Furthermore, we are pioneers in applying BERT to analyse the sentiment expressed in the text comments. By using BERT, which has proven to be highly effective in capturing contextual information and semantic understanding, we are able to extract more accurate sentiment signals from employee reviews. This innovative approach enables us to uncover previously unrecognised patterns and relationships between employee satisfaction and stock returns.

1.3 Chapter Overview and Contributions

Following this introduction, Chapters 2 to 4 present the main body of this thesis, where we discuss sentiment analysis methods, multilingualism and financial implications. Table 1.1 provides a systematic overview by chapter, summarising its primary focus, findings and contributions.

In Chapter 2, we establish a rigorous framework for sentiment analysis applications in finance, contributing to the practical deployment and evaluation of NLP models. Our study encompasses a wide range of approaches, from lexicon-based methods to machine learning and robust BERT models. We are the first to empirically evaluate 31 different sentiment analysis methods using a sample of 20,000 Glassdoor employee reviews. We employ various metrics across seven experimental settings, providing a comprehensive analysis of their performance.

Our findings demonstrate that BERT and machine learning models outperform lexicon-based approaches, particularly when the task becomes more complex. This contribution to the literature is essential as it documents the effectiveness of different sentiment analysis methods in the finance domain. This chapter also significantly adds to the existing literature by moving beyond the traditional focus on star ratings. Our research highlights the importance of considering sentiment in text comments for capturing additional nuance. By documenting the empirical evaluation of various methods and their performance in a finance-specific context, we offer a novel contribution to sentiment analysis research.

In Chapter 3, we explore the impact of automatic neural machine translation (ANMT) on multilingual sentiment analysis. This research establishes us as the first to delve into the intersection of translation, multilingualism and finance. Our empirical investigation focuses on analysing Glassdoor reviews in Portuguese, French, Spanish, and German, as well as their translations into English. This unique focus on the financial domain sets our research apart from previous studies in multilingual sentiment analysis. The results in this Chapter reveal a significant increase in sentiment misclassification rates for the translated texts compared to the original texts. However, we find that the quality of translation has minimal influence on the misclassification rates. Instead, attributes such as prediction probabilities, language, sentiment, and readability of the text play a more substantial role. This outcome suggests that some meaningful information could be lost or altered during translation. It is essential for researchers and practitioners in finance to exercise caution when analysing multilingual texts.

Additionally, we demonstrate the benefits of zero-shot transfer, where knowledge is transferred across languages to enable predictions in languages not included in the model training. We show that fine-tuning multilingual language models directly on English text and subsequently making predictions on multilingual texts can be highly effective, especially in situations where labelled multilingual data is scarce. This finding presents an innovative approach for conducting sentiment analysis across different languages, even in the absence of extensive multilingual training data. Furthermore, our results indicate that the success of zero-shot transfer is affected by the syntactic similarity between foreign languages and English. Languages that share greater syntactic similarity with English tend to exhibit better performance in this transfer learning setup. This observation further contributes to the understanding of the knowledge transformation in multilingual sentiment analysis and provides valuable guidance for future researchers.

Chapter 4 examines the relationship between employee satisfaction and stock returns using Glassdoor employee reviews. Our study measures employee sentiment through three approaches: the overall star rating of the reviews and sentiment predictions derived from BERT and Loughran-McDonald (LM) dictionary applied to text comments. To the best of our knowledge, our work is the first to use BERT to assess employee sentiment and empirically evaluate it in correlation with stock returns. According to BERT and the overall star ratings, portfolios with medium to high sentiment exhibit significant positive alphas. However, LM suggests that the low sentiment portfolio performs better, indicating a contrasting relationship. We further investigate the outcomes by sorting portfolios based on the difference between the sentiment measures. Our analysis indicates that BERT offers more advantages than LM when considering high employee sentiment portfolios. Additionally, the assessment of text using either BERT or LM appears to be more empirically indicative than relying solely on the overall star rating. This finding adds to the existing literature by strengthening the understanding that text-based sentiment measures offer a richer and more nuanced perspective on employee satisfaction compared to simple numerical ratings.

Moreover, our study highlights the predictive power of positive sentiment in forecasting significant increases in stock returns across both short and long-term periods when employing BERT. However, LM and the overall star ratings lack substantial evidence to support their effectiveness in predicting stock returns. Our findings also discover the combined effect of employee satisfaction and the firm's employee-related costs on stock returns in specific industries. This emphasises the importance of considering both employee satisfaction and associated costs when examining the relationship between employee sentiment and stock returns in certain sectors.

Lastly, Chapter 5 serves as the concluding chapter of the thesis, summarising the key findings and implications of the research. It also provides a discussion of the limitations of the study and suggests potential directions for future research in sentiment analysis, employee satisfaction, and their impact on various domains. By exploring these directions, researchers can further deepen their understanding of sentiment analysis techniques, improve the accuracy of predictions, and uncover new insights into the relationship between employee sentiment and outcomes in different organisational contexts.

Table 1.1: Chapter Overview

	Chapter 2	**Chapter 3**	**Chapter 4**
Title	Sentiment Analysis Methods: Survey and Evaluation	Multilingual Sentiment Analysis with Glassdoor Employee Reviews	BERT Employee Sentiment and Stock Returns
Research Objective	Establish a rigorous framework for sentiment analysis applications in finance and evaluate the performance of 31 different sentiment analysis methods.	Overcome language barriers for global communication by conducting a novel exploration at the intersection of translation, multilingualism, and finance.	Examine the relationship between employee satisfaction and stock returns and compare the empirical outcomes of sentiment analysis methods.
Related Work	Huang et al. (2023), Bochkay et al. (2023)	Pires et al. (2019), Poncelas et al. (2020)	Green et al. (2019), Symitsi et al. (2021)
Dataset	20,000 Glassdoor employee reviews in English.	31,024 Glassdoor employee reviews in Portuguese, French, Spanish and German; 21,833 Glassdoor employee reviews in English.	1,352,736 Glassdoor employee reviews in English for 617 unique constituents of the S&P 500 index from January 2008 to March 2021.
Sentiment Analysis Methods	Harvard IV-4 dictionary (HIV4), Loughran-McDonald (LM) dictionary, SentiWordNet, and VADER, Logistic Regression (LR), Naive Bayes (NB), Support Vector Machines (SVM), Decision Tree (DT), Random Forests (RF) and Extreme Gradient Boosting (XGB), BERT and FinBERT	Multilingual BERT (mBERT), DistilmBERT, Cross-Lingual Language Model-RoBERTa (XLM-R)	Overall star rating, LM, BERT
Findings	BERT models outperform other methods in binary and three-class sentiment classification, but they require higher computing power. LR, SVM, and XBG are reliable machine learning classifiers. Lexicon-based approaches are weaker but useful in limited-resource scenarios.	Translated texts show higher sentiment misclassification rates compared to originals, irrespective of translation quality. Models with zero-shot transfer ability offer effective predictions in multilingual sentiment analysis, even for languages with limited training data.	BERT is effective in predicting stock returns and it shows advantages over LM for high employee sentiment portfolios, but in general text-based sentiment measures (BERT and LM) provide a richer perspective on employee satisfaction than numerical ratings.
Contributions	It contributes to the practical deployment and evaluation of NLP models in financial sentiment analysis, enabling future research to select informed models based on data resources, computational power, and research goals.	First work on multilingual sentiment analysis in finance, emphasising direct text analysis to prevent information loss in translation. Offers practical solutions for researchers with limited multilingual training data.	First study to use BERT for employee sentiment assessment and examine its relationship with stock returns. Strengthens understanding that text-based sentiment measures offer deeper insights into employee satisfaction beyond star ratings.

Chapter 2

Sentiment Analysis Methods: Survey and Evaluation

2.1 Introduction

The rise of digital platforms means that an unprecedented amount of textual data from news and social media is now easily accessible to academics and market participants. While the information within the textual data is abundant and potentially valuable to gain insights into public opinions and predict future trends, at the same time, it can be noisy and the sheer volume of material overwhelming. Therefore, researchers need to select the most efficient tools to identify, extract and analyse the most relevant information (Kobayashi et al. 2018, Pandey & Pandey 2019, Abbasi et al. 2019, Hickman et al. 2022, Bochkay et al. 2023). In this paper, we focus on an emerging topic in finance, sentiment analysis, a task that allows researchers to process the massive information flow by computationally identifying the emotions behind textual data.

As one of the most popular datasets used in finance research, Glassdoor stores over 70 million employee reviews, in both structured (numerical ratings) and unstructured (text comments) formats, relating to more than 600,000 companies worldwide. However, most empirical analysis focuses more on the numerical star rating (from one to five) provided by the reviewers as a measure of sentiment rather than considering the content of any text comments. Using the overall star ratings, researchers have demonstrated that employee satisfaction can be associated with ROA, Tobin's Q and operating margin to predict corporate performance (Melián-González et al. 2015, Huang et al. 2015, Symitsi et al. 2018, Stamolampros et al. 2019, Corritore et al. 2020). In addition, changes in employee satisfaction and flexibility reflected in their star ratings can also impact on stock returns (Huang et al. 2015, Au et al. 2021), customer contact (Wolter et al. 2019), and business outlooks such as future operating performance (Huang et al. 2020) and corporate disclosure (Hales et al. 2018).

Previous studies using Glassdoor text comments have employed text mining techniques such as topic modelling (Schmiedel et al. 2019, Symitsi et al. 2021), dictionary-based text analysis programmes such as DICTION, Linguistic Inquiry and Word Count (LIWC) and WordNet (Symitsi et al. 2018, Stamolampros et al. 2019, Corritore et al. 2020), and data-mining software like IBM Watson (Dabirian et al. 2017, 2019) to extract factors that affect employee satisfaction including organisational structure (Huang et al. 2015, Creek et al. 2019), culture (Robertson et al. 2019, Canning et al. 2020), financials (Jing et al. 2019) and policies (Storer & Reich 2021). Moreover, Tambe et al. (2020) use cluster analysis of text from Glassdoor reviews to claim information technology workers prefer to work for companies that use emerging technologies in part because they value technology and learning on the job. Campbell & Shang (2022) use an inverse regression approach to derive importance weights for words and demonstrate that attributes related to corporate misconduct are widespread among employee comments. Sull et al. (2022) identify 172 topics frequently mentioned in Glassdoor employee reviews and study the sentiment related to each topic, finding that cultural toxicity drives up employee turnover.

We find the methods used in the textual analysis of employee evaluations are still relatively homogeneous. While Bochkay et al. (2023)'s work provides a guideline for implementing Natural Language Processing (NLP) models in finance, our work is specialised in sentiment analysis and presents the first rigorous and comprehensive comparison of 31 different sentiment analysis methods using the text from 20,000 Glassdoor employee reviews. The methods we adopt in this paper can be divided into three categories: the lexicon-based approaches (i.e., Harvard IV-4 dictionary (HIV4), Loughran-McDonald (LM) dictionary, SentiWordNet, and VADER); the machine learning approaches (i.e., Logistic Regression (LR), Naive Bayes (NB), Support Vector Machines (SVM), Decision Tree (DT), Random Forests (RF) and Extreme Gradient Boosting (XGB)); and the pre-trained language models (i.e., Bidirectional Encoder Representations from Transformers (BERT) and its extensions). By comparing these methods, we demonstrate how NLP models can be applied to sentiment analysis, discuss the trade-offs between models, and how to select a suitable metric to evaluate the performance of the models.

Our work also expands on this literature by conducting a comprehensive review of sentiment analysis methods previously applied in finance studies and developing a framework that summarises current sentiment analysis methods based on the fundamental characteristics of texts and the types of NLP tools used for processing them. A recent study by Huang et al. (2023) found that less accurate sentiment measures tend to underestimate the economic impact of the sentiment of analyst reports and, as a result, have less explanatory power compared to the models with higher accuracy. Their paper utilises manually labelled analysts' reports as their sample for testing sentiment predictions. This has several limitations. Firstly, their sample size is smaller at only 10,000 sentences. Secondly, there may be subjectivity bias from the researcher labelling the sentence. In contrast, our approach using Glassdoor avoids the need for manual labelling, allowing for larger samples and reducing the potential for researcher bias in labelling. Additionally, the sentences in Huang et al. (2023) are not evenly distributed among the sentiment categories, and they do not use the Matthews Correlation Coefficient (MCC) as their metric. This could potentially lead to issues with the other metrics used in their analysis. Furthermore, Huang et al. (2023) only considers the sentiment of individual sentences, whereas our approach takes into account the sentiment of the whole review, which consists of at least three sentences (pros, cons, and advice).

We design seven different experiments using both the text and star ratings from Glassdoor reviews to investigate the extent to which the sentiment identified from the text matches the numerical rating. Our preferred evaluation metric is the MCC score, but we also apply four other standard metrics (Precision, Recall, F1-Score, and Accuracy) to ensure the robustness of our results. Our findings clearly show that the BERT models consistently deliver the best performance in all experiments, followed by the machine learning approaches with word embeddings and lastly the lexicon-based approaches. At the same time, we can also observe that the performance of all different approaches declines as the ambiguity of the text increases. For instance, all models perform best in Experiment 1, which simply tests whether they identify text comments as positive or negative, and worst in Experiment 7, which introduces a neutral sentiment class as well as positive or negative. This demonstrates that researchers must use caution when employing automated textual sentiment analysis in particularly complex or ambiguous cases. In addition, Xu et al. (2021) have suggested that contextual descriptions can actually provide a better indication of users' perceptions of quality than their numerical ratings. We also find some discrepancies between a review's sentiment as indicated by its overall star rating and by the text comments. Numerical ratings are good indicators of the overall sentiment but it is likely that valuable additional information can be extracted from the associated text comments.

The structure of this paper is as follows. After this introduction, Section 2.2 introduces the framework for the sentiment analysis methods most used in finance studies and reviews the related literature in detail. We distinguish between conceptual approaches involving human judgement, hybrid approaches based on the text contents, and empirical approaches to assign sentiment such as reaction studies. The hybrid approaches can be divided into deductive, mostly lexicon-based which we discuss first, and inductive approaches (machine learning or pre-trained language models) that use word embeddings and machine learning classifiers. Section 2.3 first describes and compares the word embeddings. These are the processes by which text is transformed into vectors for automated sentiment classification. We also discuss several popular machine learning classifiers that are used in combination with the above word embeddings for sentiment classification in Section 2.4. This is followed by the pre-trained language models (the BERT-related models) in Section 2.5. After setting out the background and key features of the different sentiment analysis methods, we move on to our empirical assessment of their performance. We introduce the Glassdoor data and explain the experimental setup as well as the evaluation metrics and result discussion in Section 2.6. Finally, Section 2.7 summarises our key findings and lays out several possible directions for further research.

2.2 Sentiment Analysis Framework

There are a wide variety of sentiment analysis techniques that are based on different approaches and have different levels of complexity. This can make it difficult for researchers to identify the best approach to use in their research. Based on a systematic review of the current literature in finance and partly considering the amount of human intervention with the textual data, we have categorised these techniques into three broad categories shown in Figure 2.1: 1) *Pure Conceptual Approach*; 2) *Hybrid Approach*; 3) *Pure Empirical Approach.* Further details of the literature we reviewed to develop this framework are outlined in Appendix A.1.

[INSERT Figure 2.1 ABOUT HERE]

2.2.1 Pure Conceptual Approach

In the Pure Conceptual Approach, the sentiment of a text is solely determined by human perception. The perception of this sentiment can be formed during or assigned after the text is generated, and by the creator of the text or a later reader. In most cases, a researcher has to rely on manual annotation to assign sentiment to a text. This process is usually applied to existing text that has not already been labelled for its sentiment. Manual annotation is frequently used to provide a training sample for sentiment analysis models. When dealing with a large unlabelled corpus of text, for example, it is common to label a sample from this corpus manually, and use this sample to train a model as a classifier so that it can be applied to the rest of the corpus. Using manually annotated text, Antweiler & Frank (2004) find the tone of stock messages on the internet help predict market volatility, and the effect on stock returns is statistically significant but economically small, Li (2010) suggests the tone of forward-looking financial statements is positively associated with future earnings, Huang et al. (2014) discover investors react more strongly to negative texts than to positive ones, and Abbasi et al. (2019) show that user-generated content on social media can detect adverse events in advance, and false-positive rates are further reduced after including negative sentiment polarity in the models. In addition, this approach is also used to create standard and high-quality datasets for the NLP community to test new models.

One fundamental concern with this approach is that it is inherently subjective and different readers might interpret the same text differently, particularly in more ambiguous cases. To combat this, in the Financial Phrase Bank (Malo et al. 2014), a dataset consisting of around 5,000 sentences from LexisNexis categorised by sentiment, a piece of text is usually evaluated by multiple annotators to reduce human factor biases and concluded by the strength of majority agreement. Araci (2019), Yang et al. (2020), Huang et al. (2023) used this dataset to extend the BERT model with finance-specific domain knowledge, known as FinBERT. Another issue is the time-consuming and resource-intensive process of manually annotating unlabelled text, particularly if this is being carried out by multiple different annotators to reduce bias.

Some sources, however, have already been labelled with their sentiment. For instance, during the Glassdoor review process, the reviewer first assigns a numerical star rating to the company and then provides text comments highlighting pros and cons with an option to provide further information such as “advice to management”. Another example is Stocktwits, which includes a feature for users to tag their messages as either “bullish” or “bearish” as a way of expressing their sentiment in addition to the message’s textual content. In such self-reporting systems, an indication of the sentiment is provided by the original creator of the content. In this paper, we will use the self-reported star ratings in Glassdoor reviews as our “ground truth” for the intended sentiment of the review and test how accurately the different sentiment analysis methods can identify this sentiment from the associated text comments. Our methodology will be explained in more detail below but, for now, we may note that it has two key advantages: first, it removes the possibility that a subjective human annotator might assign a different sentiment to that intended by the creator; and second, it reduces the resources needed for manual annotation and enables a larger sample to be used.

2.2.2 Hybrid Approach

The Hybrid Approach consists of two sub-categories: the Deductive Encoding of Meaning and the Inductive Encoding of Meaning. The former uses a pre-existing set of rules to deduce the sentiment of the text and is associated with lexicon-based methods of sentiment analysis, while the latter aims at analysing the text to develop a set of rules that can assign a sentiment to the text and is linked to machine learning and transformer models. This could be seen as a hybrid approach as it requires a certain level of text processing but not the same degree of manual annotation as the Pure Conceptual Approach.

Deductive Encoding of Meaning

The Deductive Encoding of Meaning applies an established framework to deduce the meaning of a text. In the field of sentiment analysis, this method uses pre-defined rules to score the sentiment of words in a document and generates an aggregated value for the polarity between positive and negative words. In the introduction, we listed several examples of rule-based programmes and software such as DICTION, LIWC and IBM Watson. Here we focus on the lexicon-based approaches that rely on dictionaries or word lists. Lexicon-based sentiment extraction is one of the most popular approaches used on financial text, not only for its relatively straightforward application but also because it does not require labelled text, and most financial sources are not labelled.

One of the earliest lexicons, General Inquirer (GI), was developed at Harvard University in the 1960s for automated content analysis and attempted to tag words across 182 categories with dictionaries (Stone & Hunt 1968). The Harvard IV-4 dictionary (HIV4) of GI assigns a positive and a negative semantic dimension to words and is often used for sentiment analysis. Tetlock (2007) use HIV4 to detect sentiment in news articles and find that high values of media pessimism are associated with downward pressure on market prices. Later, they also found that the sentiment of text can be used to predict individual firms' accounting earnings and stock returns (Tetlock et al. 2008). Engelberg et al. (2012) suggests that negative news will increase the negative relation between short sales and future returns. Using the same approach, some research has discovered that a negative tone in the text of 10-Ks and 10-Qs can increase the cost of capital and return volatility (Kothari et al. 2009). Feldman et al. (2010) identify that changes in sentiment in the management discussion and analysis (MD&A) section of a firm's SEC filing are a significant predictor of short-term market reaction. Yekini et al. (2016) find that the tone of narratives published by UK companies in annual reports is related to market reaction, and should not be considered only as impression management tools but also as a way of disseminating price-sensitive information.

The creation of a dictionary also benefits from domain-specific knowledge. While the HIV4 dictionary has been proven efficient on general texts, the Loughran-McDonald (LM) dictionary offers more specific insights into financial texts. Loughran & McDonald (2011) examined a large sample of 10-K filings between 1994 and 2008 and concluded that nearly three-quarters of negative words according to the HIV4 dictionary appeared to be *non-negative* in a finance context, and so they created the LM dictionary by evaluating the co-occurrence and frequency of words in the corpus. Both HIV4 and LM dictionaries interpret polarity employing a Lydia system (Godbole et al. 2007).

$$Polarity = \frac{(Pos - Neg)}{(Pos + Neg)} \tag{2.1}$$

Pos and Neg are word counts for the words in positive and negative sets. $Polarity > 0$ implies a positive sentiment whereas $Polarity < 0$ suggests a negative sentiment for the text. Rogers et al. (2011) show firms that have more optimistic statements in their earnings announcements suffer from higher litigation risk. Price et al. (2012) suggests both positive and negative tones of conference calls are significantly associated with abnormal returns and trading volume. Similarly, Tsai & Wang (2017) report that finance-specific sentiment lexicon has a strong correlation to financial risks, the greater the amount of finance-specific sentiment, the higher the risk. From a different perspective, Feuerriegel & Gordon (2019) finds that high-dimensionality text input from financial news can lead to over-fitting in machine learning models used to predict macroeconomic indicators, therefore, the LM dictionary can be a better solution.

Another lexicon-based sentiment analysis approach is SentiWordNet (Esuli & Sebastiani 2006, Baccianella et al. 2010), an extension of WordNet which was introduced by (Miller 1995) to expand opinion mining by evaluating words' semantic relations based on synsets. A synset is a group of synonyms and antonyms that share the same context, and each synset expresses a distinct concept but is interlinked by their conceptual relations. SentiWordNet assigns a numerical score between 0 and 1 for positivity, negativity or objectivity to each synset. Any one of these three scores can be 0 as long as they sum up to one because the scores suggest a proportional agreement of the sentiment depending on the context. The polarity is calculated as the difference between positive and negative scores of each synset s.

$$Polarity = Pos(s) - Neg(s) \tag{2.2}$$

This can raise a problem in some cases where synsets have no positive or negative polarity score. Moreover, the scores are generated through a semi-supervised step followed by a random walk step, therefore causing a lot of noise in the SentiWordNet lexicon compared to the human-validated dictionaries.

In 2014, VADER (Valence Aware Dictionary for Sentiment Reasoning) was created to overcome some of the challenges that SentiWordNet had encountered with social media textual data (Hutto & Gilbert 2014). VADER interfaces with SentiWordNet and adapts the difference between positive and negative scores of each synset as valence to differentiate a word's sentiment intensity. For instance, the words "good" and "excellent" both express positive sentiments, but "excellent" is more positive thus VADER will give it a higher sentiment rating. In addition to modifying the degree of intensity, VADER also incorporates the impact of punctuation, capitalisation, contrastive conjunction and negation. VADER was developed by obtaining a large number of ratings for words in the existing lexicons using crowdsourced resources from Amazon's Mechanical Turk. After assigning a rating to each word, VADER aggregates the ratings of all words in a text and produces a positive, a neutral, and a negative score normalised between -1 and 1 to indicate the proportion of text that falls in these categories. VADER has shown impressive results in many studies, and it also has the advantage of supporting emoticons and acronyms for sentiment analysis.

Despite a large amount of contextual interpretation and human-factored decision-making required in the development of these dictionaries, the word lists within the dictionaries can be relatively limited as well, which means that the dictionaries are not able to assign a sentiment score to words that are outside of the pre-defined lists.

Inductive Encoding of Meaning

On the other hand, the Inductive Encoding of Meaning develops an analytical mechanism derived from the data itself. This can be done using machine learning or pre-trained language models.

Van der Heijden (2022) examines the use of machine learning in accounting research and applies it to predict a firm's industry sector using publicly available financial statement data. The results show that machine learning algorithms can accurately predict industry sectors, and can be valuable in accounting domains where prediction is the main focus. Commonly, texts need to be represented numerically before being inserted into training, which can be done through bag-of-words (BOW) or term frequency-inverse document frequency (TF-IDF), these word embeddings will be examined in more detail in the following Section 2.3. Previous studies have reported that these text representations combined with machine learning models, such as NB, LR, SVM, DT, RF and Maximum Entropy (MaxEnt), performed very well on financial news and tweets (Antweiler & Frank 2004, Hagenau et al. 2013, Nassirtoussi et al. 2014, Abbasi et al. 2019, Renault 2020, Frankel et al. 2022). These machine learning classifiers will be examined in more detail in Section 2.4 below. Li (2010) and Huang et al. (2014) have compared the NB algorithm with dictionary-based approaches to classify the sentiment of analyst reports and have shown NB algorithm achieves better performance. In contrast, Sohangir et al. (2018) concluded that VADER outperforms LR, NB and SVM in extracting sentiment from financial social media. In our evaluation of these methods, we included more extensive analysis to test which approach performs better, we find the performance of machine learning relies on the choice of word embedding and hyperparameters.

Most recently, the introduction of the BERT model (Devlin et al. 2018), a state-of-the-art language representation model in NLP, has increased the accuracy of sentiment analysis to a new level. This approach lies between deductive encoding and machine learning because it inherits the features from pre-training but also relies on fine-tuning sample data for downstream tasks. Sousa et al. (2019) have conducted experiments using financial news to show that BERT (82.5% in accuracy) outperformed NB and SVM (69% in accuracy). Further studies using different textual data have also reached the same conclusion that BERT achieved the best accuracy compared to other lexicon-based and machine learning approaches (González-Carvajal & Garrido-Merchán 2020, Mishev et al. 2020, Zhao et al. 2021, Stevenson et al. 2021). Leippold (2023) highlights the vulnerability of dictionaries in financial sentiment analysis to adversarial attacks by Generative Pre-trained Transformer 3 (GPT-3) (Brown et al. 2020) and suggests that advanced techniques like BERT, which employ context-aware approaches, are shown to be more resilient.

In addition, studies have proposed different versions of FinBERT to gain more domain knowledge in finance (Araci 2019, Yang et al. 2020, Liu et al. 2021, Huang et al. 2023). Another study, Bingler et al. (2022) introduce ClimateBert which specialises in climate-related texts. Although, as Kriebel & Stitz (2021) point out, both simple and complex deep learning architectures can yield comparable results, it is not always the case that more complex models outperform in financial sentiment analysis. Ni et al. (2023) discover that fine-tuning FinBERT for financial sentiment classification achieves higher accuracy than prompting GPT-3 for zero-shot and few-shot tasks. As these are the most complicated methods, we examine them further in Section 2.5 below.

2.2.3 Pure Empirical Approach

The last class of sentiment analysis is the Pure Empirical Approach. This does not rely on human judgement or deducing/inducing sentiment from the textual data itself. Instead, the Pure Empirical Approach "let the market" decide the sentiment reflected in the text. For instance, in reaction studies the sentiment of financial reports, news and tweets is determined by observing changes in stock prices (Hagenau et al. 2013, Jaggi et al. 2021, Frankel et al. 2022). Alternatively, it has been proposed that default records could be used as a measure of the sentiment of credit reports. Kriebel & Stitz (2021) claim that textual information can improve credit default predictions. Conversely, Stevenson et al. (2021) argue that textual loan information produces relatively accurate predictions alone, however, it does not offer additional performance improvement when used with structured data. There are limitations to this approach. It is not always possible to make a clear link between a text and a market reaction; it would be presuming too much to attribute an increase (decrease) in the market value of a large firm to a single positive (negative) employee review on Glassdoor. It also limits the analytical usefulness of the identified sentiment. If a text has been identified as having positive sentiment based on a subsequent increase in firm value, then it would be tautologous to use that sentiment as an explanatory factor for firm performance. The Pure Empirical approach could be useful to assign sentiment labels to a sample of text that could be used to train a machine learning model and then applied to a larger corpus. We do not use this approach in the paper since we can use the user-assigned star rating in Glassdoor reviews as a proxy for the intended sentiment.

2.3 Word Embeddings

We will explore various approaches in our analysis, previously we discussed lexicon-based approaches, in the following sections we will explain machine learning and BERT-related approaches in more detail. However, in order to use the more advanced methods for sentiment classification, the first and most important thing is to turn unstructured text data into a structured format. Word embeddings are techniques designed to represent a text by a string of numbers. The numbers are called vectors and this process of this transformation is known as text vectorization.

2.3.1 Bag-of-Words

The bag-of-words (BOW) model is the simplest feature representation method that vectorizes text. As its name suggests, the model forms a vocabulary of all the unique words in a document and counts the occurrence of each word. However, the model ignores the order of words in a document, such that contextual information will be lost without knowing where the words occurred. In addition, the size of the vocabulary will be very large if new words occur in a new document. Increased vector lengths raise an issue of a sparse matrix where most elements in the vectors are zeros.

2.3.2 Term Frequency-Inverse Document Frequency

Whilst the BOW model is simple to use, it is biased towards words with high occurrence. High-frequency words such as “and” and “the” may not contain relevant or meaningful information about the documents. Term Frequency-Inverse Document Frequency (TF-IDF) overcome this problem by re-scaling the word frequency across all documents and providing a weight for each word. The weighting is achieved by multiplying two matrices: a term frequency matrix $tf(t, d)$ measuring the word frequency in a document, and an inverse document frequency matrix $idf(t, D)$ capturing the distinctiveness of a word across the entire corpus. As a result, words containing important information will be assigned with a relatively high TF-IDF weight.

$$tf(t, d) = \frac{f_d(t)}{max_{w \,\epsilon\, d}\, f_d(w)} \tag{2.3}$$

$$idf(t, D) = log(\frac{|D|}{|\{d \,\epsilon\, D : t \,\epsilon\, d\}|}) \tag{2.4}$$

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \tag{2.5}$$

$f_d(t)$ is the frequency of term t appears in the document d; $max_{w\,\epsilon\,d}\, f_d(w)$ is total number of terms in the document; D is the corpus of documents; $\{d\,\epsilon\,D : t\,\epsilon\,d\}$ is the number of documents with term t.

2.3.3 Word2Vec

A fascinating improvement made on word embeddings was when Mikolov et al. (2013) from Google introduced the Word2Vec model that takes advantage of the neural network structure. In the previous BOW and TF-IDF models, vectors used to represent text are only numbers without any semantic meanings, and the vector space is proportional depending on the vocabulary size. The Word2Vec model generates a unique vector for each word in the document, and instead of storing all the information of the vocabulary, the idea is to design a model that groups the vectors of similar words in the vector space. Word2Vec has two architectures to achieve this, a continuous bag of words model (CBOW) and a continuous Skip-Gram model, as shown in Figure 2.2.

[INSERT Figure 2.2 ABOUT HERE]

The CBOW model aims to predict the centre word by its surrounding words with a fixed window size, while the Skip-Gram model does exactly the opposite, it is trained to predict the probability of the surrounding neighbour words given a centre word. Since the Skip-Gram model takes single words of input, it is less sensitive to frequent words and it has a better representation of rare words, whereas CBOW tends to overfit high-frequency words. However, Skip-Gram is computationally more expensive because it needs to predict several outputs for the context words whilst CBOW is only trained to predict one centre word at a time. In general, the Word2Vec model has the advantage of maintaining the semantic relations of different words in a document with little requirement of computer memory, and these encoded properties can be easily extended for sentiment analysis. One of the disadvantages of Word2Vec is that it handles out-of-vocabulary (OOV) words poorly because it assigns random vector representations for them, and the information retrieved is only locally optimal because a word representation relies on its neighbours.

2.3.4 Global Vectors for Word Representation

GloVe is short for Global Vectors for Word Representation, created by researchers from Stanford University (Pennington et al. 2014), and is an unsupervised learning algorithm that derives word embeddings through global word-word co-occurrence statistics. It combines a global matrix factorisation (e.g. latent semantic analysis (LSA)) and a local context window (e.g. Skip-Gram) in the training process. GloVe and Word2Vec are both unsupervised methods for word embeddings fitted up to 300-dimensional word vectors. Similar to Word2Vec, GloVe uses contextual information for text representation but the main difference between them comes from their structures of generating word vectors. Word2Vec is a predictive model that focuses on local context information of words, whereas GloVe is a count-based model which incorporates global statistics through dimensionality reduction on the co-occurrence counts matrix to obtain word vectors. Using different training methods, the Word2Vec embeddings provided by the Python library Gensim are pre-trained on Google News dataset for 3 million words and phrases [1], while the GloVe embeddings introduced by Stanford are pre-trained on Wikipedia, Common Crawl, and Twitter with varying sizes [2]. In various NLP tasks such as analogy, word similarity and named entity recognition (NER), GloVe has demonstrated an outstanding performance compared to other models even with a small text corpus. The pre-trained word vectors of GloVe for large corpora are publicly available for researchers to

[1] https://radimrehurek.com/gensim/models/word2vec.html

[2] https://nlp.stanford.edu/projects/glove/

access.

2.4 Machine Learning Classifiers

Sentiment analysis is essentially a classification task, where a classifier builds and trains on extracted features from the text. We have discussed the word embeddings for feature extraction in the previous section, here we consider several supervised machine learning classifiers that are widely used for sentiment analysis and have proven to be powerful in many studies. We are interested in comparing their performance in combination with different feature representations and word embedding techniques for sentiment analysis on our Glassdoor employee reviews.

2.4.1 Logistic Regression (LR):

LR is one of the most popular algorithms for its simplicity to implement and interpret. It operates with a hypothesis and a sigmoid function to determine the probability of an output sentiment. The range of the sigmoid function is between 0 and 1, with a threshold of 0.5, the sigmoid function transforms any value into a range from 0 to 1. Hence any probabilities that are greater than the threshold will lead to the class value of 1 (e.g. positive sentiment) otherwise 0 (e.g. negative sentiment). Logistic regression is usually used for binary classification, it assumes the features and classes are independent of each other and predicts a binomial probability. In multi-class sentiment analysis, standard LR can be modified by predicting a multinomial probability for the input features and changing the loss function from the Least Squared Error to Cross Entropy.

2.4.2 Naive Bayes (NB):

In the sentiment analysis, NB is also referred to as Multinomial Naive Bayes classifier (Rish et al. 2001). It uses the Bayes Theorem that predicts the probabilities of sentiment class by using the joint probabilities of words and classes. With the assumption of feature independence, given the features $X = x_i$, and the sentiment classes $Y = y_i$.

$$P(Y|X) = \frac{P(Y) \cdot P(X|Y)}{P(X)} \tag{2.6}$$

Where $P(Y)$ is the prior probability and $P(X|Y)$ is the likelihood probability. It is fast computing compared to other classifiers, requires little training data and performs well in multi-class classification tasks. The main limitation of NB is that it assumes all the attributes are mutually independent implying no link between one word to another, which is rarely the case in real life.

2.4.3 Support Vector Machines (SVM):

SVM performs sentiment classification by assigning a hyperplane that best separates the positive and negative classes (Cortes & Vapnik 1995). The optimal hyperplane is found by maximising the distance between the hyperplane and each class. The data points used to define the hyperplane are referred to as the support vectors. In a binary sentiment classification, the classes are two-dimensional and the hyperplane can be as simple as a line, however, it will become more complex and non-linear if the dimension increases. Thanks to this attribute, SVM allows for higher accuracy when multi-dimensional features occur.

2.4.4 Decision Trees (DT):

Proposed by Quinlan (1986), this algorithm is built in the structure of a tree with decision nodes, branches and leaves as its name suggests. Each decision node represents a filtering criterion, and the branch shows whether the criteria are met then leads to an outcome of a class label on a leaf node. Starting from the root node, predictions on the leaves are made by continuously going through the nodes on the tree until the decision nodes can not be split further. DT algorithm works well on both numerical and categorical data and requires little data preparation. However, due to their hierarchical structure, decision trees often over-fits the data and suffers from high variance issue where any small changes can cause instability of the predictions.

2.4.5 Random Forests (RF):

RF was first introduced by Breiman (2001) and functions by aggregating the predictions of a collection of DT. It adopts a type of ensemble method called bagging. With bagging, each decision tree is trained by a random subset of features independently, and their predictions are then averaged for an eventual prediction. The training can be efficiently done in parallel. Since RF selects the features randomly, it processes the inputs in a more generalised way without depending highly on any specific set of features and therefore usually outperforms DT.

2.4.6 Extreme Gradient Boosting (XGB):

Similar to RF, gradient boosting is an ensemble learner and a set of DTs as well. Different from bagging, boosting is another type of tree ensemble method. Instead of building each tree independently, gradient boosting builds one tree at a time and combines the predictions along the training process by constantly correcting the errors from the prior models. XGB is an implementation of the gradient boosting framework introduced by Chen & Guestrin (2016), and since its release, it has been one of the leading algorithms in Kaggle competitions and popular in sentiment analysis.

We will use all combinations of these word embeddings and machine learning classifiers in our study.

2.5 Pre-trained Language Models

Pre-trained language models are large neural networks that are used in a wide variety of NLP tasks. They operate under a "pre-train" to "fine-tune" paradigm: models are first pre-trained over a large scale of text and then fine-tuned on a specific downstream task.

The motivation behind pre-training is to understand the contextual meaning of a word, and the word embeddings discussed previously ignore the order of words during their training. As a result, the vector of a word is static and always the same across any sentence or document even though the same word might have different meanings in a different context. By contrast, contextual embedding such as BERT (Devlin et al. 2018) relies on a Transformer architecture that takes into account the sequence of all words and their positions in a sentence from both left-to-right and right-to-left contexts simultaneously.

2.5.1 Transformer

The Transformer is a state-of-the-art deep learning model in NLP for processing sequential data such as language modelling and machine translation (Vaswani et al. 2017). It adopts a self-attention mechanism that enables the model to transform one sequence to another with an encoder and a decoder and allows for more parallelization during training. Intuitively, attention is a function that relates different positions of the input sequence to compute a representation of that sequence. From each attention unit, the Transformer learns a set of Query, Key, and Value weight matrices denoted by W_Q, W_K, W_V respectively during the model training. The first step in calculating self-attention is to use the word embedding matrix X to multiply the weight matrices to produce three vectors; a query vector: $Q = XW_Q$, a key vector $K = XW_K$, and a value vector $V = XW_V$. The second step is to calculate the self-attention weights by taking the dot product of the query vector and the key vector, then dividing it by the square root of the dimension of key vector $\sqrt{d_k}$ for stabilising gradients. Thirdly, the attention weights are passed through a softmax to get normalised scores which are then multiplied by the value vectors by a compatibility function. Lastly, the output of the self-attention layer is computed as the sum of the weighted value vectors.

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{2.7}$$

Running through an attention mechanism several times in parallel, the multi-head attention in the Transformer model operates by mapping 8 pairs of queries, keys, and values. After concatenating and linearly transforming the independent attention outputs, the multi-head attention enables the model to jointly attend to information between words at various positions.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{2.8}$$

$$head_1 = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2.9}$$

where W_i^Q , W_i^K ,W_i^V and W_i^O are learnable parameter matrices.

Figure 2.3 shows a demonstration of the structure of the Transformer. The encoder (the left half of Figure 2.3) builds a continuous word embedding for sequential representation and feeds into the decoder (the right half of Figure 2.3) to generate an output sequence. The output embedding comes from the embedding layer which is shared with the input embedding, and the positional encoding is to make sure the model learns the order of the sequence. The encoder and decoder are composed of a stack of six identical layers, and each layer mainly consists of a feed-forward network and a multi-head attention module.

[INSERT Figure 2.3 ABOUT HERE]

In a feed-forward neural network, the sequential information is passed on in one direction, from the input layer through hidden layers and to the output layer. It is different from Recurrent Neural Networks (RNNs) where the information cycles through a loop, and every decision is made based on a current and a previous input. Without communications or inferences with the previous computation in each position, the Transformer model successfully reduces the training time. A sequence is processed as a whole instead of word by word, therefore, the Transformer model is not affected by long dependencies. In the decoder, masking is applied to the first sub-layer to ensure the predictions for a position are only dependent on the known outputs before such a position. This is then connected with the outputs from the encoder and eventually passes through a softmax layer to generate a prediction of the output sequence.

2.5.2 Bidirectional Encoder Representations from Transformers

Building on the foundation of the Transformer architecture, recently many groundbreaking models for text representation have been proposed. BERT stands for Bidirectional Encoder Representations from Transformers, it was introduced by Devlin et al. (2018) from Google to pre-train deep bidirectional representations from the unlabelled text by joint conditioning on both left and right context in all layers. In the BERT model, several Transformer encoders are stacked on top of each other. Recall the Transformer architecture, the encoder takes an input sequence and the decoder outputs the predicted sequence word by word, since the goal here with BERT is to generate features for downstream NLP tasks, the decoder part of the Transformer is discarded. The Transformer model in BERT uses bidirectional self-attention that can read the input sequence from both left-to-right and right-to-left of a word. Such bidirectionality offers an understanding of word relations based on the whole context which makes it one of the core competencies of BERT compared to other language representation models.

BERT is pre-trained on the BooksCorpus (800M words) and the English Wikipedia (2,500M words) using two unsupervised tasks simultaneously shown in Figure 2.4 a: Masked Language Model (MLM) and Next Sentence Prediction (NSP). In MLM, 15% of the words in the input sequence are randomly replaced by a [MASK] token for BERT to predict given the context while the NSP performs a binary classification to predict whether or not two sentences follow each other given a [SEP] token as a separator between sentences. BERT uses WordPiece (Wu et al. 2016*a*) for tokenization, it has the coverage of 30,000 most frequent and common combinations of words in the vocabulary and any OOV words are broken down into subwords with "##" symbol. For instance, the word "embedding" would be tokenized as "em", "##bed", "##ding". Two versions of the pre-trained BERT model are made publicly available [3], the base version has 12 encoder layers, 12 attention heads, and 110M parameters whereas the large version has 24 encoder layers, 16 attention heads, and 340M parameters.

[INSERT Figure 2.4 ABOUT HERE]

The pre-trained contextual word representations and parameters can be applied to downstream tasks after they have been fine-tuned on task-specific labelled data (Figure 2.4 b). Fine-tuning is a common step where a classification layer is added to the pre-trained language model with all parameters jointly updated, it takes less time and data to develop compared to the pre-training process. The reason behind it is to utilise the pre-trained language model to recognise classes they were not originally trained on. Figure 2.4 (b) shows an example of the fine-tuning procedure for classification tasks such as sentiment analysis. Words from an input sentence are tokenized and passed into the BERT model to produce word embeddings, the special [CLS] token in front of the input sentence treated as a pooled representation of the input is then fed to an output layer for classification. During this process,

[3] https://github.com/google-research/bert

all parameters are fine-tuned. The architectures of pre-training and fine-tuning are the same except for the output layer, from which the class label probabilities are computed with a standard softmax. With minimal adjustments on the pre-trained language model [4], BERT has obtained new state-of-the-art accuracy on various NLP tasks including sentiment analysis.

2.6 Experiment

Our empirical study aims to test which of the above Hybrid approaches (lexicon-based, machine learning or BERT) performs best at identifying the sentiment of Glassdoor employee reviews across seven experiments.

2.6.1 Data

Our data comes from *Glassdoor.com*, one of the most extensive job and recruiting platforms. Since its launch in 2008, Glassdoor has collected more than 70 million anonymous reviews by employees. Glassdoor offers reviews in different categories: companies, salaries, interviews etc. In this paper, we focus on the core company reviews. Figure 2.5 demonstrates an anonymous company review from Glassdoor. On starting a review, the user is first asked to identify the company from a drop-down list. Then they are asked to "rate your experience of this company" between one star and five stars. No further guidance is provided on the interpretation of the task or the scale. We take this overall star rating as a proxy for the reviewer's sentiment towards the company. Referring to the classification above, it is a conceptual and self-reported sentiment label. Reviewers also have to provide text comments with a minimum of five words on *pros* and *cons* for the company. There is an optional text

[4] The Transformer models only require a fine-tuning step on the pre-trained language model with a classification layer instead of training a whole separate classifier when performing sentiment analysis. The word embeddings we discussed previously, on the other hand, only create feature representation of the text hence they need the help of a separate classifier to deliver the task.

field to provide *advice to management.* To maximise the text available for sentiment analysis, our sample only includes reviews that provide advice to management. The density histogram and word cloud for these three columns are shown in Appendix A.2 The majority of the reviews are in short phrases or incomplete sentences. There are additional fields, including six sub-ratings from one star to five stars for *Overall, Work-life Balance, Culture and Values, Career Opportunities, Compensation and Benefits*, *Diversity and Inclusion*, and *Senior Leadership*, and "thumbs up or down" questions on the rating of the CEO, whether the reviewer would recommend the company to a friend, and their view of the business outlook for the next six month. All of these are optional and we do not include them in our analysis.

[INSERT Figure 2.5 ABOUT HERE]

In our analysis, we web-crawled 20,000 employee reviews from S&P 500 companies. Based on the majority of sentiment analysis studies in the NLP field, we believe our sample size has a balanced trade-off between the baseline representation and computational cost. The textual data is cleaned and pre-processed using Python's NLTK package to reduce noise for sentiment analysis. We start with tokenisation where the entire text review is split into individual words. Then we remove stopwords such as "a", "and", "or", and "the", which occur frequently in the reviews but are not useful for the analysis. Punctuation is also removed, this step is followed by stemming which is removing the suffix from a word and reducing it to its root form. Lastly, we perform parts of speech (POS) tagging to label words with grammatical descriptions, such as nouns, adjectives and verbs etc. such that words are incorporated with context.

2.6.2 Experimental Setup

To perform sentiment analysis, the lexicon-based approaches count the number of positive and negative words in the text and assign the corresponding sentiment, whereas the machine learning and pre-trained language models function in a supervised manner and require labelled data during the training process. Many sentiment analyses using review-based social media data, such as Yelp restaurant and hotel reviews, Amazon product reviews, and IMDB movie reviews commonly label the sentiment of text by the overall rating. We use both the text and star ratings from Glassdoor reviews and develop a series of seven experiments to test the performance of different sentiment analysis methods across a variety of tasks. The experiments will identify the best-performing sentiment analysis approach overall but also allow us to see how the performance of the different methods changes as the ambiguity and complexity of the tasks increase. The aim is to inform the choice of method by future researchers depending on their research aims and resources. Table 2.1 shows a summary of our experiments including the text used in an experiment, and how it is labelled. Positive sentiment is indicated by $+$, and $-$ represents negative sentiment.

Experiment 1 takes advantage of the fact that Glassdoor employee reviews separate positive and negative comments into *pros* and *cons* fields. The first and simplest task is, therefore, to see if the sentiment analysis method can correctly identify the text from the *pros* field as representing positive sentiment and the text from the *cons* field as representing negative sentiment. This provides a baseline test of performance in a simple binary sentiment classification. Note that all the reviews contain text for *pros* and *cons* as these were compulsory fields.

The following experiments (2-7) take the overall star rating for the review as the intended sentiment (as the self-reported and subjective assessment of the reviewer) and test how accurately the different methods can predict this sentiment based on the text comments. The experiments differ in terms of the text used and how the sentiment is labelled. Experiment 2 only uses the text from the *pros* and *cons* fields while Experiments 3-7 add the text from *advice to management*. While including *advice to management* will increase the size of the text corpus available for analysis and so, we might assume, its accuracy, it is also possible that “advice” might be less straightforward and more difficult to interpret than *pros* and *cons*. Since our data collection was limited to those reviews that did complete the optional *advice to management* field, this also includes the full sample.

Experiments 3-7 are designed to test how well the different sentiment analysis methods perform at varying levels of ambiguity and complexity. Experiments 2 and 3 are limited to reviews with an overall rating of one star or five stars - which are likely to represent the strongest expressions of sentiment - and a binary classification as either positive or negative. We expect the sentiment analysis methods to be more accurate when dealing with these cases than in the more ambiguous situations in the subsequent experiments. Note that these experiments have a smaller sample size because of the restriction to one- and five-star reviews as shown in Figure 2.6.

[INSERT Figure 2.6 ABOUT HERE]

Experiment 4 widens the definition of positive and negative sentiment by including one and two-star reviews as representing negative sentiment and four- and five-star reviews as representing positive sentiment. We expect that the inclusion of two- and four-star reviews will tend to increase the ambiguity of the sentiments expressed and therefore reduce the performance of the sentiment analysis methods. As a result, the sample size is larger than Experiments 2 and 3 but smaller than Experiments 1 and 5-7.

Finally, Experiments 5-7 incorporate three-star reviews, which are likely to be the most difficult to classify in terms of sentiment. Experiments 5 and 6 retain a binary sentiment classification. In Experiment 5, three-star reviews are classified as positive sentiment (so one and two stars = negative and three, four and five stars = positive). In Experiment 6, three-star reviews are classified as a negative sentiment (so one, two and three stars = negative and four and five stars = positive). Experiment 7 introduces a third category of “neutral” for three-star reviews (so one and two stars = negative; three stars = neutral; four and five stars = positive). Our expectation is that the performance of the models will decline as the degree of ambiguity and complexity increases with the introduction of a multi-class rather than binary sentiment classification.

[INSERT Table 2.1 ABOUT HERE]

Figure 2.6 illustrates the class distribution of the experiments, the portion of the positive class is mostly larger than the negative ones. To handle the class imbalance issue, our data is split in a stratified fashion such that the proportions of each class are approximately the same in the training and testing sets. Under each one of the seven experimental settings, we are also interested in the performance of BERT models in a binary and a multi-class sentiment analysis compared to the lexicon-based approaches as well as the machine learning techniques combined with BOW, TF-IDF, Word2Vec and GloVe word embeddings. The hyper-parameters in machine learning methods are optimised through grid search. The description of these models is listed in Appendix A.3. In addition to the vanilla BERT models[5], we also include FinBERT[6] (Yang et al. 2020) because the reviews may contain financial-contextual information, such as salary, compensation, benefits, work-life balance, and the financial stability of a company.

During fine-tuning of the pre-trained language models, we use AdamW optimiser and select the learning rate of 2e-5, batch size of 32 and training epoch of five. Figure 2.7 demonstrates the pipeline of the sentiment analysis.

[INSERT Figure 2.7 ABOUT HERE]

2.6.3 Evaluation Metrics

There are many evaluation metrics for measuring classification performance. We chose the Matthews correlation coefficient (MCC) as the main evaluation metric

[5] In this paper, we adopt two versions of the pre-trained BERT models: BERT-base-case `https://huggingface.co/bert-base-cased` and BERT-base-uncased `https://huggingface.co/bert-base-uncased`.

[6] `https://huggingface.co/yiyanghkust/finbert-tone`

for the sentiment analysis methods comparison. A primary reason for the choice of MCC is because it excels in dealing with imbalanced class issues in the classification. As indicated in Figure 2.6, the proportion of sentiment class in Experiments 1, 2, 3 and 6 are relatively close, however, the class distribution is more imbalanced in Experiments 4, 5 and 7 where the differences are more than 20%. We also report other classic evaluation metrics as a robustness check: Precision, Recall, F-1 score and Accuracy.

Given the performance measurements of a confusion matrix: true positive (TP), false negative (FN), true negative (TN), and false positive (FP), these evaluation metrics are interpreted and computed as follows in binary sentiment classification.

Matthews correlation coefficient (MCC): MCC was introduced by Matthews (1975) to measure the quality of machine learning classifications. In statistics, it is referred to as the phi coefficient or mean square contingency coefficient. It is designed to assess all four of the confusion matrix matrices: TP, FN, TN, and FP regardless of the size of each class, therefore, it handles imbalanced data very well.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(FN + TN)(FP + TN)(TP + FN)}} \tag{2.10}$$

Accuracy: It is the most intuitive measure by looking at the overall performance of a model. It tells a ratio of correctly predicted observations to the total observations. However, it works best if the classes are balanced. For example, if the data contains 10% of negative instances and a classifier always assigns the positive label, the overall accuracy would still reach 90% since it would correctly predict 90% instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.11}$$

Precision: It is known as the positive predictive value. Precision calculates the

TP cases out of the total predicted positives, and it is a good measure for detecting FP.

$$Precision = \frac{TP}{TP + FP} \tag{2.12}$$

Recall: Unlike precision which ignores all but positive instances and predictions, recall tells the coverage of the actual positive sample. It is also called sensitivity or true positive rate, which indicates the ratio of positive instances that are correctly detected by the classifier.

$$Recall = \frac{TP}{TP + FN} \tag{2.13}$$

F-1 Score: It is a weighted average of Precision and Recall. It is a good measure to use when seeking a balance between Precision and Recall, and the value goes up only when both Precision and Recall are high. F-1 Score is usually preferred to accuracy if there is a large uneven class distribution.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{2.14}$$

Although F-1 Score can handle class imbalance issues to an extent and is arguably easier to interpret than MCC, Boughorbel et al. (2017), Yao & Shepperd (2020), Chicco & Jurman (2020) and Chicco et al. (2021) prove F-1 score to be more biased and misleading than MCC because it ignores the count of TN and highly influenced by positive observations. In contrast, MCC generates a more balanced assessment by taking into account all four items in the confusion matrix. Therefore, in a case where the majority class is negative, for instance, Experiment 6 in Figure 2.6, MCC is a fairer option.

The range of MCC is between -1 and 1, a high score is achievable only if the prediction performs well on the majority of data instances from every class, independently of the class ratios in the overall sample. The value of Accuracy, Precision, Recall, and F-1 score ranges from 0 to 1, with a higher score indicating better performance. In a multi-class scenario, these metrics can be computed by one-vs-all comparisons through macro averaging which turns multi-class predictions into multiple sets of binary predictions and then averages the corresponding metric for each of the binary cases.

2.6.4 Result Discussion

The performance of the different sentiment analysis methods in the seven experiments is shown in Tables 2.2. Across the experiments, the average MCC of lexicon-based, machine learning and BERT-related approaches are 0.224, 0.507 and 0.674, respectively. According to the mean MCC, we observe that the performance of machine learning approaches was double and BERT-related approaches were triple that of the lexicon-based approaches. In addition, the MCC of BERT-related approaches was also 50% higher than machine learning ones.

[INSERT Table 2.2 ABOUT HERE]

We discuss the results in more detail below. However, to provide an accessible overview of the performance of the different sentiment analysis methods, Figure 2.8 presents them in rank order based on our preferred MCC metric. Models on the far right have the highest MCC score on average and hence are ranked the highest.

[INSERT Figure 2.8 ABOUT HERE]

The first aim of this paper is to identify the best-performing method for sentiment analysis. It is clear that the BERT models are the best performing across all experiments, regardless of the association with various star ratings or whether we use binary and three-class sentiment analysis. As shown in Figure 2.8, looking at the maximum mean MCCs from the three types of sentiment classifiers, BERT models have the highest mean MCC of 0.91. The best-performing machine learning approach has the highest mean MCC of 0.77. There is considerable variation in the performance of different combinations of word embeddings and classifiers across the different experiments. However, all the machine learning approaches perform worse than the BERT and better than the lexicon-based approaches (highest mean MCC = 0.55), which are ranked the lowest on average.

To consider the BERT models in more detail, normally the choice of BERT-base-case or BERT-base-uncased is dependent on whether the letter case is sensitive to the task. From our observation of the raw text, the reviews are usually short phrases and incomplete sentences in lowercase. As a result, in Experiments 1 and 2, uncased BERT has a slight advantage over cased. FinBERT is a BERT model pre-trained on financial communication text, it works very well on employee reviews containing financial terms. After a closer examination of the BERT models' misclassified cases, we find the comment "None that I can think of" or sarcastic expressions frequently appeared in *pros* or *cons*. It is difficult to assign the sentiment to these cases without human intervention, but we notice that FinBERT handles these cases slightly better than BERT. This may say something about the level of sarcasm or cynicism in financial language. Nevertheless, we find after including the additional text from *advice to management* in Experiments 3 to 6, the performance of FinBERT declines noticeably, probably due to the fact that the *advice to management* contains fewer finance-related words. For this reason, BERT should be preferred when using non-financial texts.

Turning to Machine Learning approaches, Figure 2.9 reports the average of MCC grouped by word embeddings and machine learning classifiers. This aggregated matrix identifies that while RF can be used as a baseline method, NB and DT have less predicting power compared to LR, SVM and XGB. Word2Vec and GloVe do not show a clear advantage over the simple feature representations such as BOW and TF-IDF even though theoretically the task should benefit from the semantic meaning of word embeddings such as Word2Vec and GloVe. Mishev et al. (2020) also find that TF-IDF is the best text representation method among BOW, Word2Vec and GloVe testing on the Financial Phrase Bank (Malo et al. 2014) and SemEval2017-Task5 dataset (Cortis et al. 2017). A possible explanation is that pre-trained word embeddings are usually high-dimensional, and machine learning methods struggle to model all of the information included.

[INSERT Figure 2.9 ABOUT HERE]

Despite their relatively poor performance in our experiments, lexicon-based approaches may still have value, particularly when the speed of computing is an issue or when dealing with unlabelled text. VADER is the best-performing lexicon-based approach, the LM dictionary designed for financial texts performs better than the HIV4 dictionary and SentiWordNet, which matches the results from previous studies. Before we perform the grid search to optimise hyper-parameters in machine learning approaches, we find VADER outperforms all DT in Experiments 2, 5 and 6 (results of DT with untuned hyper-parameters are displayed in Appendix A.4). This suggests that, although the dictionary approaches are weak, they have the potential to improve if the dictionaries can be tuned subject to the task as well.

In addition to the comparison of sentiment analysis methods, we are also interested in the impact of ambiguity of labelling and the complexity of classification on performance. Therefore, we aggregate the performance of the same type of model. Figure 2.10 displays the change of mean MCC when varying how the star ratings are used to provide sentiment labels.

[INSERT Figure 2.10 ABOUT HERE]

Experiment 1 serves as a baseline case. In Experiment 1, the sentiment is identified based on the text from the *pros* and *cons* fields separately. In effect, could the sentiment analysis method correctly identify the text from *pros* as expressing positive sentiment and that from *cons* as negative? This is the most straightforward task and all methods performed their best. The BERT models have a mean MCC of 0.91 in predicting that text from the pros field represented positive sentiment and text from the cons field negative sentiment. As noted above, some sarcastic responses may account for some of the errors. The machine learning method and the lexicon-based approach both achieve their highest MCC of 0.70 and 0.44, respectively.

The next set of experiments is more complex. In Experiment 2, we test whether the methods could identify one-star reviews as negative and five-star reviews as positive based on the combined text from the *pros* and *cons* fields. We observe a relative decrease of 9% - 34% in mean MCC for all types of approaches from Experiment 1 to Experiment 2. The reduced mean MCC reveals that using the overall rating as the intended sentiment can potentially introduce more noise to the models [7]. Experiment 3 includes the text from *advice to management*. As the amount of text increases, the BERT models and the machine learning models benefit from the additional training data and therefore, their mean MCC increase relatively by 1% and 9%, respectively. By contrast, longer texts on average reduce the MCC of the lexicon-based approach by relatively 3%.

As was pointed out in the experiment setups, we are curious about the impact of increasing the ambiguity of the sentiment by including first two and four-star ratings, and then three-star overall ratings. Experiments 4 to 7 use the texts from *pros*, *cons* and *advice to management* but change how the star ratings are used to label the review sentiment as positive or negative. In Experiment 4, we include one and two-star ratings as negative and four and five-star ratings as positive. Adding two and four-star ratings should have included more ambiguous and difficult-to-classify cases. Indeed, we see a sizeable drop in mean MCC for all types of methods, a drop that continues after we introduced the three-star ratings - the most ambiguous category in terms of sentiment. From Experiments 3 to 5, the mean MCC dropped relatively around 40% for all types of methods, The first point is that when more ambiguous labels are included, performance drops. Second, ratings in the middle ground are

[7] It is unlikely that the main factor in the decrease in mean MCC in Experiments 2 and 3 compared to Experiment 1 is due to the reduced sample size after restraining the star ratings because the mean MCC continued to decrease when the sample size was increased in the following experiments.

more ambiguous representations of textual sentiment than ratings at the extremes (one or five stars).

Experiments 5 and 6 treat three-star ratings as positive and negative respectively. The models have slightly higher MCC on average when three-star reviews are considered as being positive as compared to negative. One interpretation could be that the underlying sentiment of three-star ratings could be seen as grudgingly positive or at least not as negative. The other explanation is that there is a discrepancy between a review's sentiment as indicated by its overall star rating and by the contents of text comments. Experiment 7 is the most complex of all, introducing a third sentiment class (neutral) as well as positive and negative. The mean MCC of BERT models drops from 0.528 in Experiment 5 to 0.436 in Experiment 7, the machine learning models from 0.384 to 0.326, and lexicon-based approaches from 0.173 to only 0.024. Given this decline in performance across the board, researchers should be cautious when applying any sentiment analysis method to more ambiguously-labelled data or to produce multi-class sentiment classifications.

We look at the accuracy of models in predicting the user-assigned sentiment from their text comments but this ignores the possibility that there is additional information that could be extracted by looking at the contents of the text fields themselves. The preceding analysis assumes that the user-generated star rating is the best guide to the intended sentiment of the review, in order to test the accuracy of different sentiment analysis methods in predicting this sentiment from the associated text comments. However, it is possible that, in some cases, the text comments could contain valuable additional information that could be extracted. Therefore, solely focusing on numerical ratings as measures of sentiment risks missing significant information that could be gleaned from the text, potentially providing more insight and context to deepen the research findings.

2.7 Conclusion

In summary, this paper offers a comprehensive assessment of 31 frequently-used sentiment analysis methods used by finance academics, from the lexicon-based approach to machine learning classifiers with word embeddings, and then to the powerful NLP model BERT. These different approaches were classified within a framework that consists of the Pure Conceptual Approach, the Hybrid Approach, and the Pure Empirical Approach for detecting the sentiment of textual data. The paper focuses on the Hybrid approach, which is divided into methods employing a Deductive Encoding of Meaning (e.g. lexicon-based approaches) and an Inductive Encoding of Meaning (e.g. machine learning and pre-trained language models).

Our work appears to be the first study to evaluate these sentiment analysis methods using the text comments within Glassdoor employee reviews. Our empirical study has two main purposes. First, we compare the relative performance of the different sentiment analysis methods to help future researchers to identify the optimal method for their work. The results conclusively prove that BERT models produce the best outcomes in both binary and three-class sentiment classification. However, their deployment requires a higher level of computing power. The machine learning methods produce solid predictions, and their performance is better when the model has better tolerance towards high-dimensional feature representations. We show that LR, SVM, and XBG are reliable classifiers, and the choice of word embedding is as important. Among them, our results indicate that word embeddings such as Word2Vec and GloVe do not show a clear advantage over BOW and TF-IDF due to their high dimensionality.

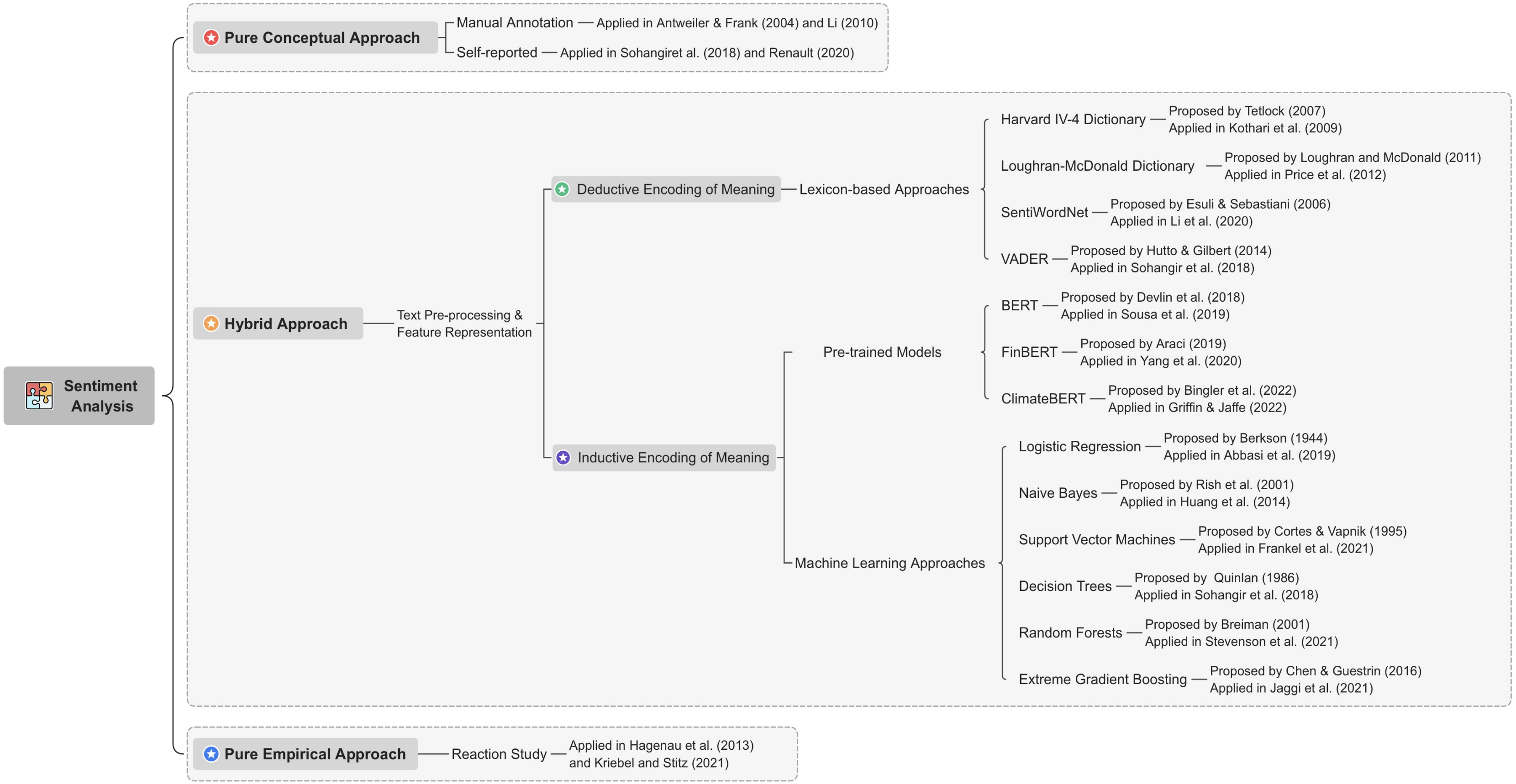
In collaboration with machine learning models, future research could consider exploring their usage along with more advanced classifiers such as RNNs, Convolutional Neural Networks (CNNs), and Long Short Term Memory (LSTM). Lexicon-based approaches are the weakest performers but may still have some use where computing power or expertise is at a premium and when it is not feasible to manually annotate a training sample. In some cases, VADER produces very close results to DT and NB. Our comparison of different models offers a reference for future research on sentiment analysis of financial texts in general. The BERT models are more powerful but are more computationally expensive at the same time. Depending on the specific task, researchers can choose a model suitable for them considering the complexity and accuracy trade-off. Given the increasing dominance of large language models in NLP, future research should further explore the role of AI in addressing the challenges of information overload.

We demonstrate the performance of all the sentiment analysis models declines as the complexity of the task and ambiguity of the labels increase. All methods perform their best in identifying the text from the *pros* field as representing positive sentiment and the text from the *cons* field as representing negative sentiment. However, when we take the overall star rating of the reviews as the expected sentiment, the performance of all methods decreases. This is especially the case for the three-star ratings, which contain the most ambiguous expressions of sentiment. Finally, we show that the complexity of the sentiment classification increases with the introduction of neutral sentiment and therefore, the performance of the model decreases.

The analysis of Glassdoor employee reviews using advanced NLP techniques holds significant importance in finance as it provides organisations with a more precise and efficient means of comprehending the sentiment, concerns, and feedback expressed by their employees. By gaining a deeper understanding of employee sentiment, organisations are empowered to take appropriate actions to enhance workplace culture, increase employee satisfaction, and ultimately improve overall firm performance.

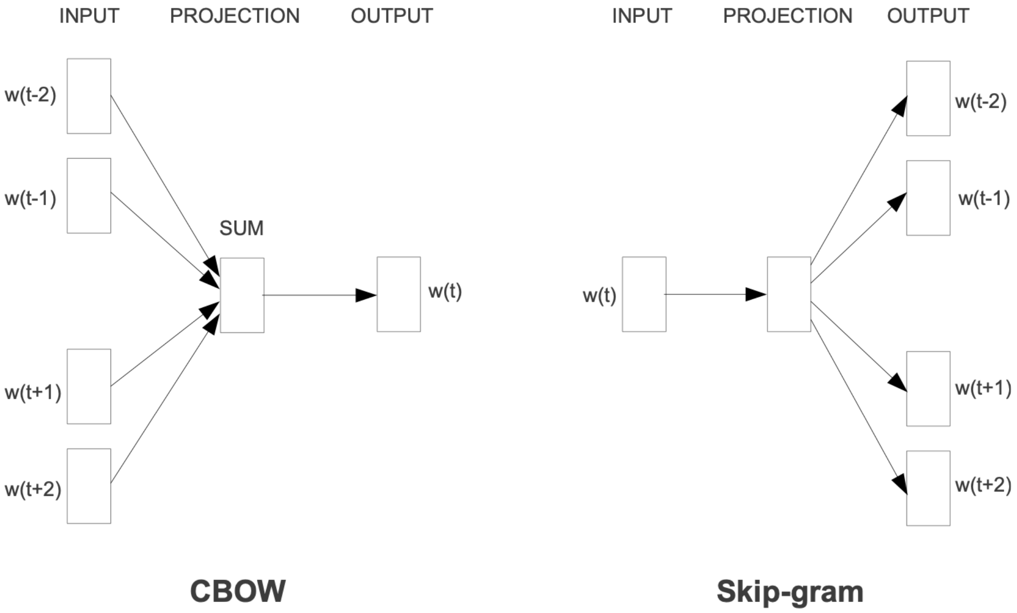
2.8 Figures and Tables for Chapter 2

Figure 2.1: Sentiment analysis framework


Sentiment Analysis
Pure Conceptual Approach
Manual Annotation — Applied in Antweiler & Frank (2004) and Li (2010)
Self-reported — Applied in Sohangiret al. (2018) and Renault (2020)
Hybrid Approach
Text Pre-processing & Feature Representation
Deductive Encoding of Meaning
Lexicon-based Approaches
Harvard IV-4 Dictionary — Proposed by Tetlock (2007) Applied in Kothari et al. (2009)
Loughran-McDonald Dictionary — Proposed by Loughran and McDonald (2011) Applied in Price et al. (2012)
SentiWordNet — Proposed by Esuli & Sebastiani (2006) Applied in Li et al. (2020)
VADER — Proposed by Hutto & Gilbert (2014) Applied in Sohangir et al. (2018)
Inductive Encoding of Meaning
Pre-trained Models
BERT — Proposed by Devlin et al. (2018) Applied in Sousa et al. (2019)
FinBERT — Proposed by Araci (2019) Applied in Yang et al. (2020)
ClimateBERT — Proposed by Bingler et al. (2022) Applied in Griffin & Jaffe (2022)
Machine Learning Approaches
Logistic Regression — Proposed by Berkson (1944) Applied in Abbasi et al. (2019)
Naive Bayes — Proposed by Rish et al. (2001) Applied in Huang et al. (2014)
Support Vector Machines — Proposed by Cortes & Vapnik (1995) Applied in Frankel et al. (2021)
Decision Trees — Proposed by Quinlan (1986) Applied in Sohangir et al. (2018)
Random Forests — Proposed by Breiman (2001) Applied in Stevenson et al. (2021)
Extreme Gradient Boosting — Proposed by Chen & Guestrin (2016) Applied in Jaggi et al. (2021)
Pure Empirical Approach
Reaction Study — Applied in Hagenau et al. (2013) and Kriebel and Stitz (2021)

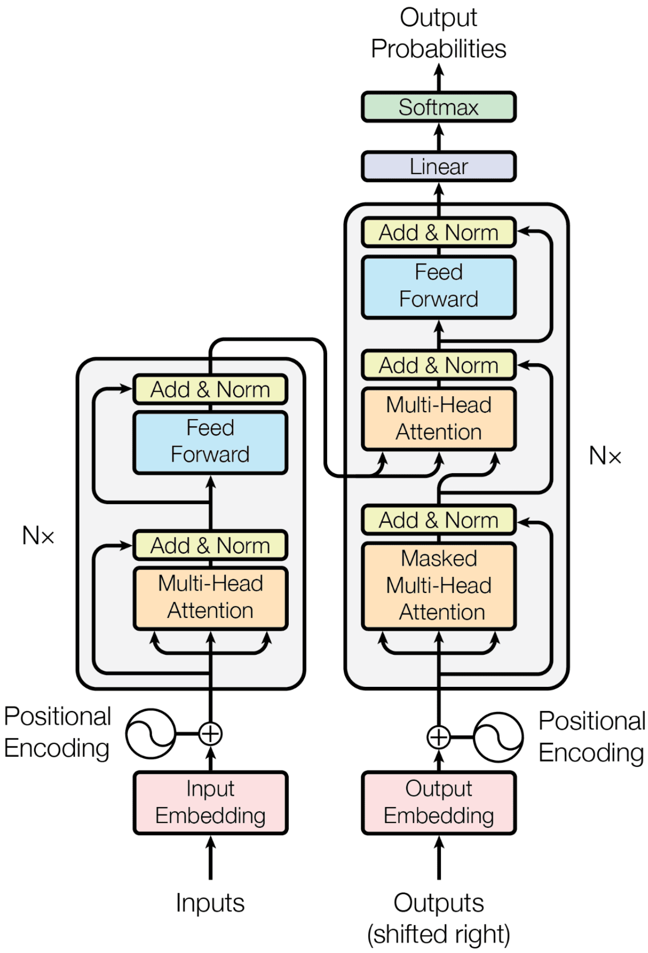
Notes: (1) Pre-trained language models refer to models that have been pre-trained on a large amount of text to gain semantic understandings of words, for instance, the BERT-related models. (2) Machine Learning Approach is a category that encodes features of a sample text that is a part of the corpus.

Figure 2.2: The Word2Vec architectures



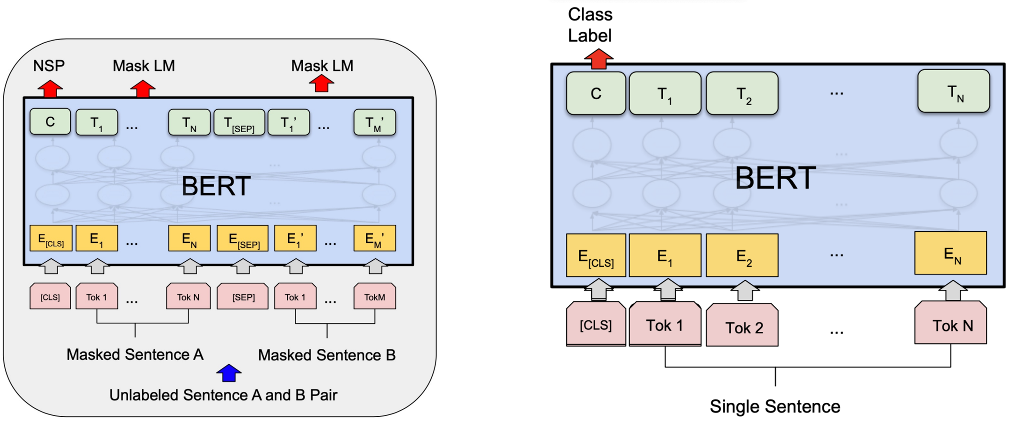
Notes: CBOW: A neural network that learns from the context words to predict the centre word. Skip-Gram: A neural network that learns from the centre word to predict its context words.

Figure 2.3: The Transformer model architecture



Notes: The left half of this figure is the encoder which builds a continuous word embedding for sequential representation. It is then fed into the decoder (the right half) to generate an output sequence. The output embedding comes from the embedding layer which is shared with the input embedding, and the positional encoding is to make sure the model learns the order of the sequence. The encoder and decoder are composed of a stack of six identical layers, and each layer mainly consists of a feed-forward network and a multi-head attention module.

Figure 2.4: Overall pre-training and fine-tuning procedures for BERT.

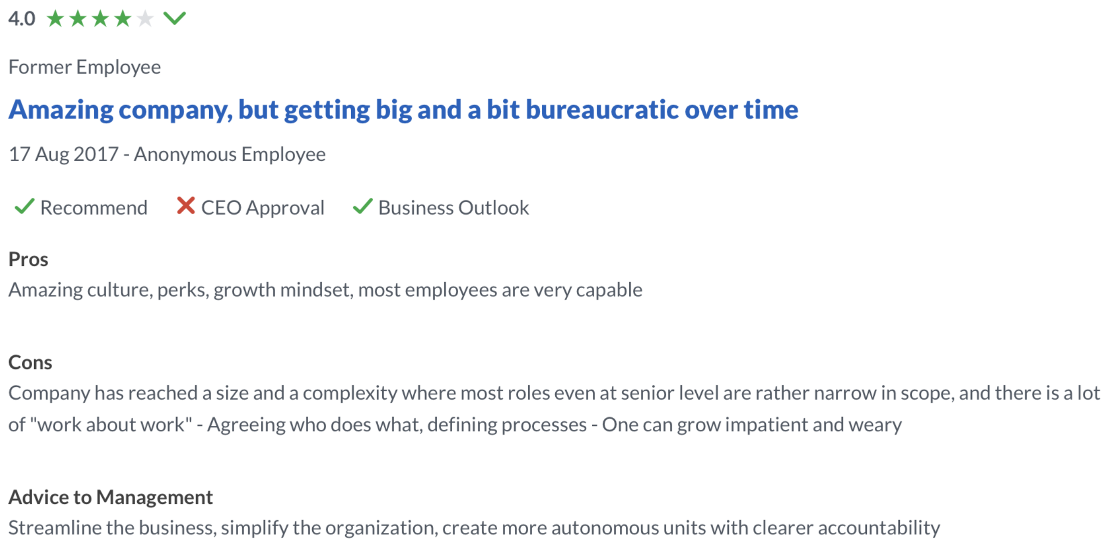


(a) Pre-training

(b) Fine-tuning

Notes: (a) shows the two tasks for BERT per-training: Masked Language Model (MLM) and Next Sentence Prediction (NSP). (b) shows how BERT is fine-tuned to a downstream task, this example is for sentiment analysis.

Figure 2.5: An anonymous employee review from *Glassdoor.com*



4.0 ★★★★☆ ✓

Former Employee

Amazing company, but getting big and a bit bureaucratic over time

17 Aug 2017 - Anonymous Employee

✓ Recommend ✗ CEO Approval ✓ Business Outlook

Pros

Amazing culture, perks, growth mindset, most employees are very capable

Cons

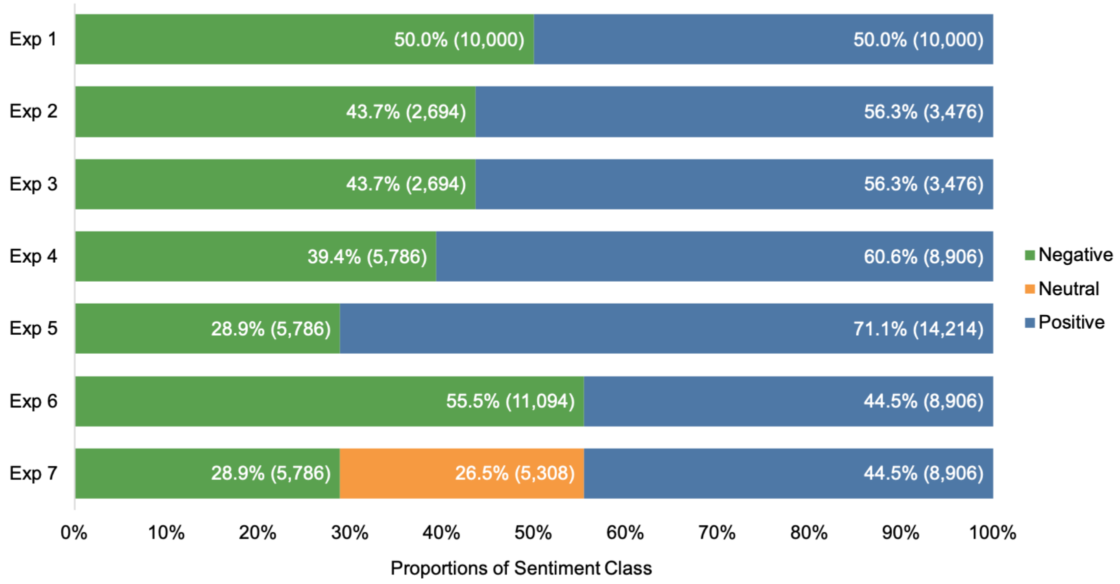
Company has reached a size and a complexity where most roles even at senior level are rather narrow in scope, and there is a lot of "work about work" - Agreeing who does what, defining processes - One can grow impatient and weary

Advice to Management

Streamline the business, simplify the organization, create more autonomous units with clearer accountability

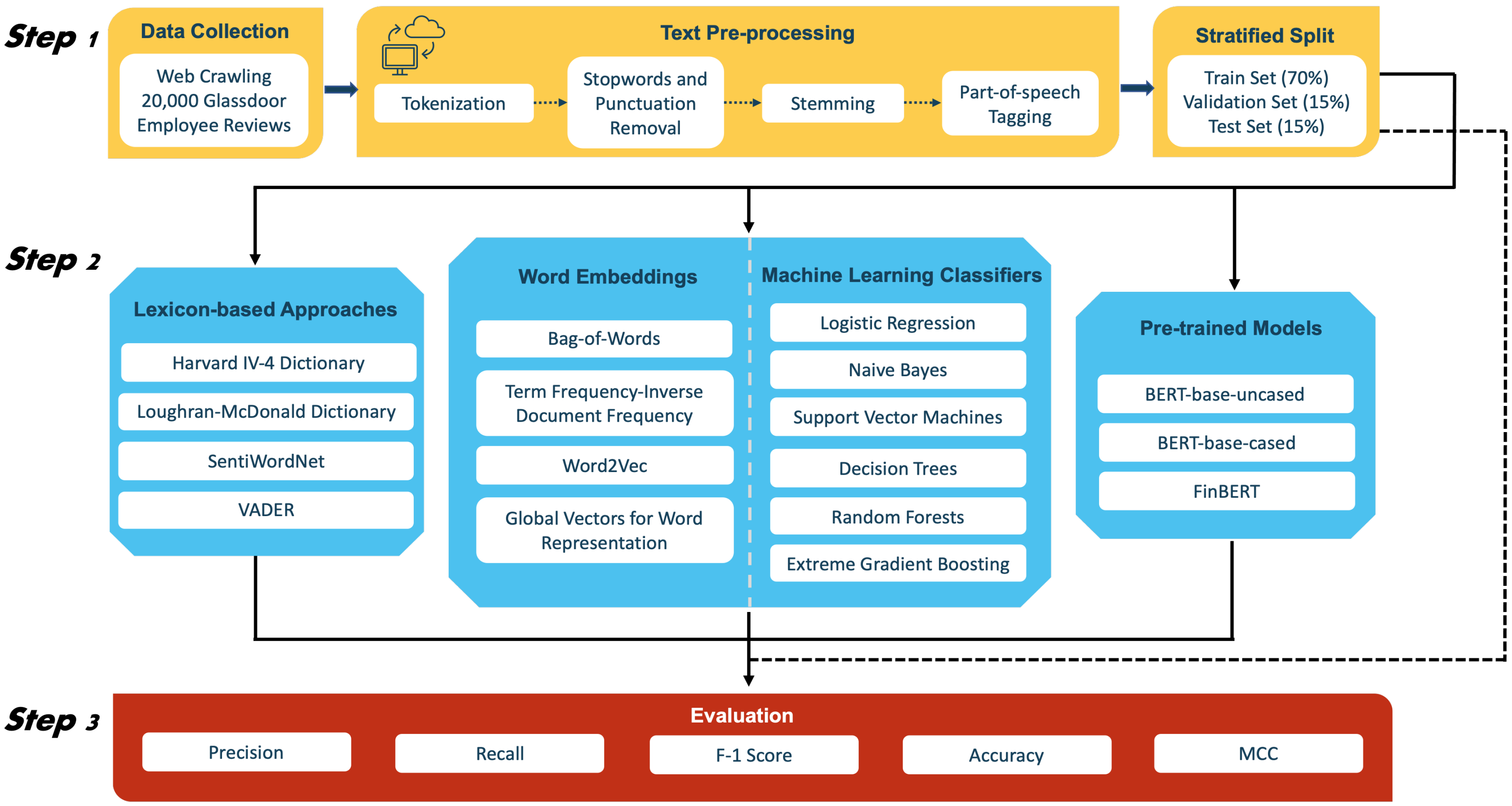
Notes: This is a screenshot of a piece of anonymous employee review from Glassdoor. The sub-ratings are optional and not explicitly linked to the overall rating. Reviewers have to leave text comments of at least five words for *pros* and *cons*, and have the option to provide advice to management.

Figure 2.6: Sentiment class proportions of experiments



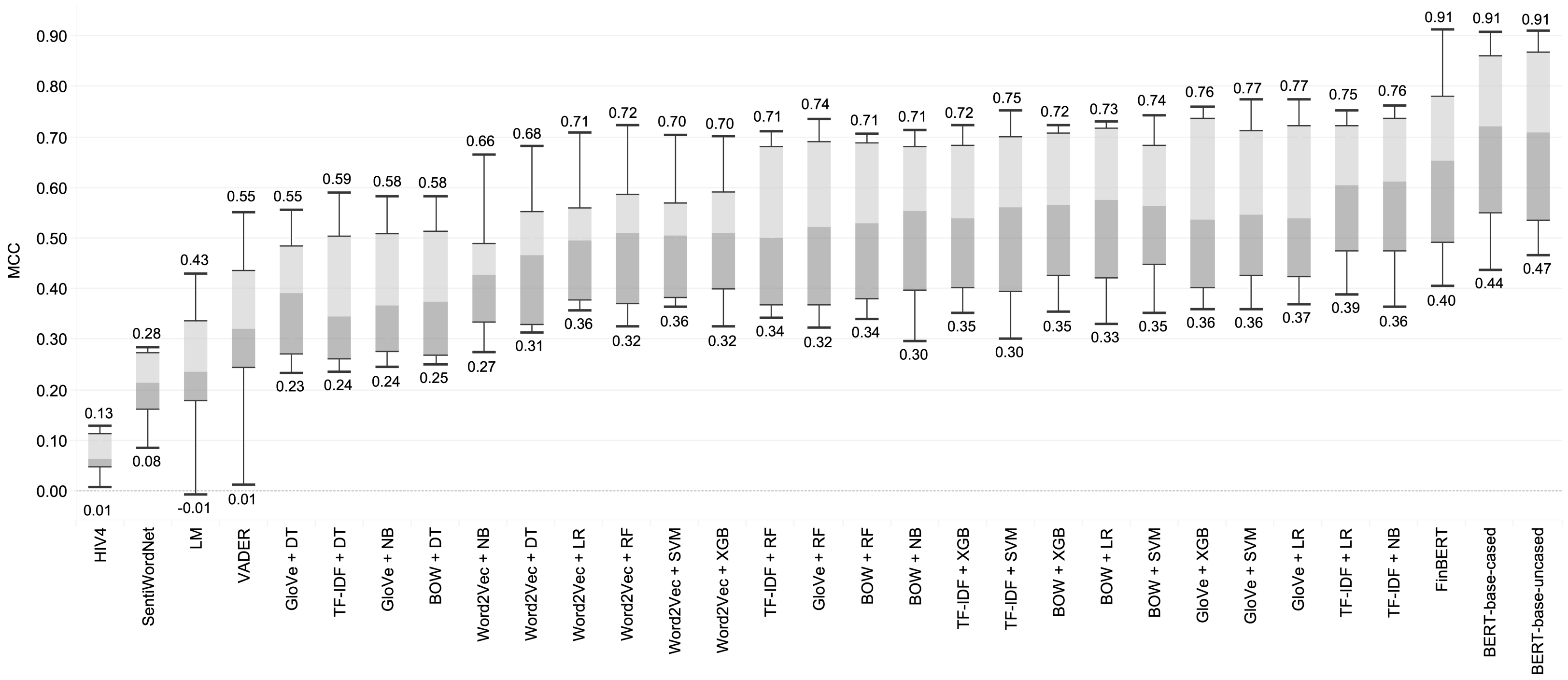
Notes: Our sample data consists of 20,000 reviews, This bar chart shows the proportion of sentiment class in each experiment, and the number of observations is included in the pretences. The proportion of sentiment class in Experiments 1, 2, 3 and 6 are relatively close, however, it is more imbalanced in Experiments 4, 5 and 7 where the differences are more than 20%.

Figure 2.7: Sentiment analysis experiments pipeline



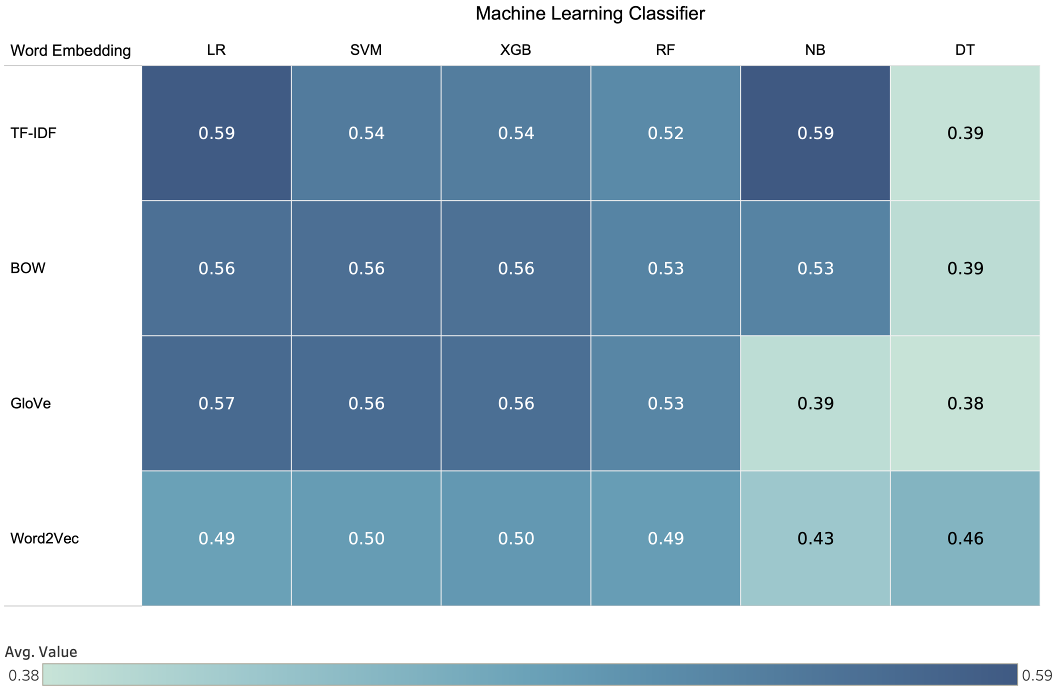
Notes: We first collect the data, pre-process the text and split them into the train, validation and test sets. Next, we fit the data into different sentiment classifiers and eventually evaluate them on the test dataset.

Figure 2.8: Performance of sentiment analysis methods



Notes: This box plot shows the performance of sentiment analysis methods in MCC. Each box is composed of the MCC scores of all experiments in which it is placed, the x-axis is sorted by the rank of their mean MCC. The whiskers are extended to values within 1.5 times the interquartile range. The values above and below the whiskers represent the highest and the lowest MCC of a method.

Figure 2.9: Performance of machine learning classifiers in relation to word embeddings

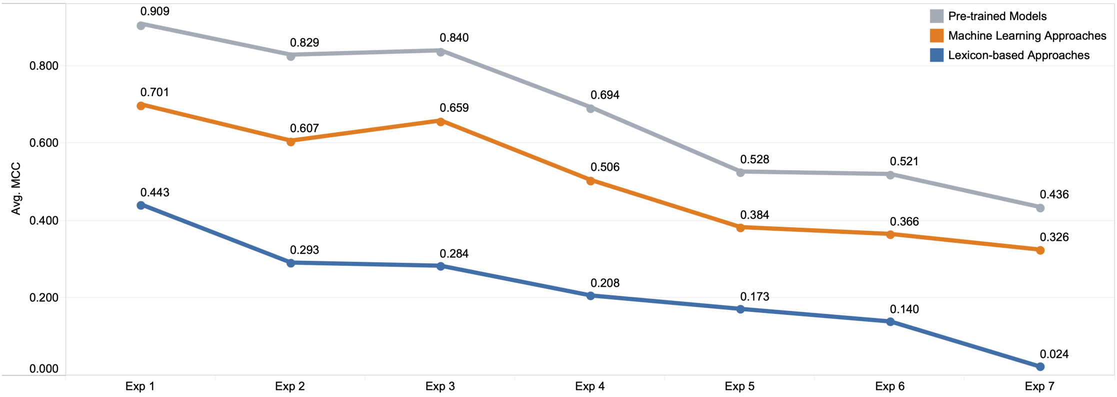


Machine Learning Classifier

Word Embedding	LR	SVM	XGB	RF	NB	DT
TF-IDF	0.59	0.54	0.54	0.52	0.59	0.39
BOW	0.56	0.56	0.56	0.53	0.53	0.39
GloVe	0.57	0.56	0.56	0.53	0.39	0.38
Word2Vec	0.49	0.50	0.50	0.49	0.43	0.46

Notes: This matrix shows the correlation between the machine learning classifiers and word embeddings in terms of their mean MCC.

Figure 2.10: Change of Avg. MCC per experiment (grouped by the types of sentiment analysis methods)



Notes: This figure shows the change of the mean MCC per experiment after grouping the models from each sentiment analysis method.

Table 2.1: Experimental Setup

	Text			**Label**				
	Pros	Cons	Advice to Management	Overall Rating: 1 Star	Overall Rating: 2 Star	Overall Rating: 3 Star	Overall Rating: 4 Star	Overall Rating: 5 Star
Exp 1	✓, +	✓, −						
Exp 2	✓	✓		−				+
Exp 3	✓.	✓	✓	−				+
Exp 4	✓	✓	✓	−	−		+	+
Exp 5	✓	✓	✓	−	−	+	+	+
Exp 6	✓	✓	✓	−	−	−	+	+
Exp 7	✓	✓	✓	−	−	Neutral	+	+

Notes: This table summarises the setup of our experiments. The left side of the table suggests the part of textual review used in an experiment, and the right side suggests how they are labelled. + means the corresponding overall rating is considered as positive whilst − means the corresponding overall rating is considered as negative.

Table 2.2: Full Sentiment Analysis Results

	Exp 1					Exp 2					Exp 3					Exp 4				
Method	Precision	Recall	F1-score	Accuracy	MCC	Precision	Recall	F1-score	Accuracy	MCC	Precision	Recall	F1-score	Accuracy	MCC	Precision	Recall	F1-score	Accuracy	MCC
Lexicon-based Approaches																				
LM	0.721	0.709	0.706	0.709	0.430	0.664	0.666	0.663	0.665	0.330	0.670	0.671	0.670	0.673	0.341	0.615	0.620	0.615	0.623	0.235
HIV4	0.694	0.648	0.626	0.648	0.339	0.594	0.544	0.495	0.584	0.129	0.579	0.530	0.464	0.573	0.096	0.551	0.519	0.466	0.601	0.063
SentiWordNet	0.738	0.715	0.707	0.715	0.453	0.636	0.623	0.622	0.639	0.258	0.651	0.633	0.631	0.652	0.284	0.614	0.599	0.600	0.637	0.213
VADER	**0.782**	**0.768**	**0.765**	**0.768**	**0.550**	**0.764**	**0.694**	**0.690**	**0.721**	**0.453**	**0.756**	**0.669**	**0.658**	**0.700**	**0.416**	**0.712**	**0.621**	**0.610**	**0.686**	**0.320**
Machine Learning Approaches																				
BOW + LR	0.861	0.861	0.861	0.861	0.722	0.859	0.851	0.853	0.857	0.709	0.868	0.861	0.864	0.867	0.730	0.798	0.778	0.784	0.800	0.575
BOW + NB	0.857	0.857	0.857	0.857	0.714	0.842	0.834	0.836	0.840	0.676	0.849	0.836	0.840	0.845	0.686	0.784	0.769	0.774	0.789	0.553
BOW + SVM	0.872	0.871	0.871	0.871	0.743	0.831	0.819	0.822	0.827	0.649	0.862	0.854	0.857	0.860	0.716	0.790	0.774	0.780	0.794	0.564
BOW + DT	0.797	0.786	0.784	0.786	0.582	0.749	0.739	0.741	0.749	0.488	0.780	0.757	0.760	0.771	0.536	0.701	0.673	0.677	0.709	0.373
BOW + RF	0.853	0.853	0.853	0.853	0.707	0.845	0.832	0.836	0.840	0.677	0.854	0.844	0.847	0.851	0.698	0.782	0.749	0.757	0.779	0.530
BOW + XGB	0.861	0.861	0.861	0.861	0.722	0.848	0.841	0.844	0.847	0.689	0.864	0.859	0.861	0.863	0.723	0.794	0.772	0.779	0.795	0.565
TF-IDF + LR	0.877	0.876	0.876	0.876	0.753	0.854	0.846	0.849	0.852	0.700	0.874	0.869	0.871	0.873	0.743	0.813	0.792	0.799	0.814	0.605
TF-IDF + NB	0.871	0.871	0.871	0.871	0.742	**0.866**	**0.861**	**0.863**	**0.865**	**0.727**	**0.883**	**0.880**	**0.881**	**0.883**	**0.763**	**0.817**	**0.796**	**0.803**	**0.817**	**0.612**
TF-IDF + SVM	0.877	0.877	0.877	0.877	0.753	0.844	0.824	0.829	0.835	0.668	0.872	0.859	0.863	0.867	0.731	0.818	0.747	0.758	0.789	0.560
TF-IDF + DT	0.800	0.789	0.787	0.789	0.590	0.758	0.742	0.745	0.754	0.500	0.760	0.747	0.749	0.757	0.506	0.679	0.666	0.669	0.694	0.345
TF-IDF + RF	0.855	0.855	0.855	0.855	0.711	0.842	0.817	0.822	0.829	0.659	0.858	0.842	0.846	0.851	0.700	0.775	0.726	0.735	0.764	0.499
TF-IDF + XGB	0.861	0.861	0.861	0.861	0.722	0.838	0.817	0.822	0.828	0.654	0.861	0.851	0.854	0.858	0.712	0.780	0.759	0.765	0.783	0.538
Word2Vec + LR	0.854	0.854	0.854	0.854	0.708	0.750	0.746	0.747	0.752	0.495	0.809	0.809	0.809	0.811	0.618	0.751	0.746	0.748	0.761	0.496
Word2Vec + NB	0.834	0.831	0.830	0.831	0.665	0.736	0.734	0.734	0.739	0.469	0.754	0.754	0.744	0.744	0.507	0.709	0.719	0.704	0.706	0.428
Word2Vec + SVM	0.853	0.852	0.852	0.852	0.705	0.752	0.752	0.752	0.755	0.504	0.813	0.815	0.814	0.816	0.628	0.755	0.753	0.754	0.765	0.507
Word2Vec + DT	0.841	0.841	0.841	0.841	0.682	0.752	0.747	0.749	0.754	0.499	0.801	0.804	0.801	0.803	0.604	0.735	0.730	0.732	0.746	0.465
Word2Vec + RF	0.861	0.861	0.861	0.861	0.722	0.756	0.753	0.754	0.759	0.509	0.826	0.825	0.825	0.828	0.651	0.762	0.755	0.758	0.771	0.517
Word2Vec + XGB	0.850	0.850	0.850	0.850	0.701	0.770	0.766	0.768	0.772	0.536	0.823	0.821	0.822	0.825	0.644	0.757	0.753	0.755	0.767	0.509
GloVe + LR	**0.888**	**0.887**	**0.887**	**0.887**	**0.775**	0.851	0.849	0.850	0.852	0.700	0.872	0.871	0.871	0.873	0.743	0.772	0.765	0.768	0.781	0.538
GloVe + NB	0.794	0.790	0.789	0.790	0.584	0.745	0.748	0.743	0.743	0.493	0.760	0.763	0.758	0.759	0.523	0.679	0.688	0.675	0.678	0.367
GloVe + SVM	0.887	0.886	0.886	0.886	0.773	0.850	0.850	0.850	0.852	0.701	0.860	0.861	0.860	0.862	0.721	0.775	0.770	0.772	0.784	0.545
GloVe + DT	0.778	0.777	0.777	0.777	0.555	0.734	0.733	0.734	0.738	0.468	0.748	0.750	0.749	0.751	0.498	0.697	0.693	0.695	0.711	0.390
GloVe + RF	0.869	0.867	0.867	0.867	0.736	0.849	0.833	0.837	0.842	0.682	0.857	0.839	0.844	0.849	0.696	0.778	0.746	0.754	0.776	0.523
GloVe + XGB	0.880	0.879	0.879	0.879	0.759	0.866	0.859	0.862	0.864	0.725	0.875	0.872	0.873	0.875	0.747	0.773	0.762	0.766	0.780	0.535
Pre-trained Language Models																				
BERT−base−cased	0.953	0.953	0.953	0.953	0.907	0.927	0.921	0.924	0.925	0.848	**0.935**	**0.933**	**0.934**	**0.935**	**0.868**	**0.864**	**0.856**	**0.860**	**0.867**	**0.720**
BERT−base−uncased	0.955	0.954	0.954	0.954	0.909	**0.937**	**0.932**	**0.934**	**0.935**	**0.868**	0.936	0.929	0.931	0.933	0.864	0.859	0.850	0.854	0.861	0.708
FinBERT	**0.956**	**0.956**	**0.956**	**0.956**	**0.912**	0.891	0.880	0.884	0.886	0.771	0.900	0.889	0.893	0.895	0.789	0.843	0.811	0.821	0.835	0.653

Table 2.2: Full Sentiment Analysis Results - Continued

	Exp 5					Exp 6					Exp 7				
Method	Precision	Recall	F1-score	Accuracy	MCC	Precision	Recall	F1-score	Accuracy	MCC	Precision	Recall	F1-score	Accuracy	MCC
Lexicon-based Approaches															
LM	0.586	0.603	0.577	0.603	0.188	0.583	0.584	0.578	0.578	0.167	0.137	0.188	0.150	0.207	-0.007
HIV4	0.546	0.519	0.495	0.684	0.060	0.525	0.509	0.393	0.465	0.031	0.144	0.242	0.125	0.260	0.006
SentiWordNet	0.585	0.586	0.585	0.656	0.171	0.583	0.569	0.535	0.546	0.152	0.234	0.384	0.276	0.319	0.085
VADER	**0.671**	**0.609**	**0.616**	**0.733**	**0.273**	**0.644**	**0.579**	**0.500**	**0.540**	**0.213**	**0.091**	**0.220**	**0.106**	**0.236**	**0.011**
Machine Learning Approaches															
BOW + LR	0.733	0.691	0.704	0.776	0.422	0.709	0.711	0.709	0.711	0.419	0.532	0.533	0.529	0.573	0.331
BOW + NB	0.720	0.687	0.698	0.768	0.406	0.692	0.695	0.692	0.693	0.387	0.508	0.513	0.508	0.549	0.297
BOW + SVM	0.751	0.719	0.731	0.791	0.468	0.712	0.714	0.708	0.708	0.426	0.546	0.544	0.540	0.587	0.352
BOW + DT	0.696	0.598	0.601	0.740	0.277	0.629	0.630	0.629	0.634	0.259	0.471	0.463	0.423	0.529	0.250
BOW + RF	0.748	0.626	0.637	0.763	0.354	0.706	0.695	0.697	0.707	0.401	0.530	0.513	0.466	0.577	0.340
BOW + XGB	0.741	0.692	0.707	0.780	0.430	0.709	0.710	0.709	0.712	0.418	0.544	0.538	0.525	0.589	0.353
TF-IDF + LR	**0.790**	**0.700**	**0.722**	**0.802**	**0.482**	0.736	0.730	0.732	0.738	0.466	**0.569**	**0.567**	**0.557**	**0.610**	**0.389**
TF-IDF + NB	0.767	0.714	0.731	0.798	0.478	0.735	0.734	0.734	0.738	0.469	0.552	0.554	0.547	0.593	0.363
TF-IDF + SVM	0.813	0.573	0.554	0.747	0.301	0.728	0.697	0.698	0.716	0.424	0.550	0.505	0.444	0.582	0.361
TF-IDF + DT	0.684	0.606	0.612	0.738	0.279	0.617	0.618	0.617	0.619	0.235	0.481	0.466	0.427	0.521	0.242
TF-IDF + RF	0.780	0.606	0.608	0.760	0.344	0.699	0.689	0.691	0.701	0.388	0.490	0.504	0.442	0.576	0.343
TF-IDF + XGB	0.731	0.687	0.701	0.774	0.416	0.693	0.693	0.693	0.696	0.385	0.542	0.540	0.528	0.588	0.352
Word2Vec + LR	0.727	0.654	0.669	0.765	0.374	0.693	0.686	0.687	0.696	0.378	0.535	0.533	0.495	0.588	0.358
Word2Vec + NB	0.659	0.691	0.650	0.669	0.348	0.659	0.655	0.656	0.664	0.314	0.502	0.506	0.493	0.513	0.274
Word2Vec + SVM	0.735	0.661	0.677	0.770	0.389	0.690	0.683	0.685	0.693	0.374	0.719	0.523	0.444	0.585	0.363
Word2Vec + DT	0.707	0.640	0.652	0.754	0.340	0.663	0.650	0.650	0.665	0.313	0.503	0.514	0.483	0.561	0.315
Word2Vec + RF	0.727	0.661	0.676	0.767	0.383	0.679	0.677	0.678	0.684	0.357	0.505	0.517	0.489	0.570	0.325
Word2Vec + XGB	0.735	0.694	0.708	0.778	0.427	0.686	0.684	0.685	0.690	0.370	0.519	0.526	0.515	0.569	0.324
GloVe + LR	0.747	0.698	0.713	0.785	0.442	0.705	0.699	0.700	0.708	0.404	0.542	0.548	0.525	0.596	0.369
GloVe + NB	0.637	0.666	0.627	0.647	0.302	0.628	0.618	0.617	0.633	0.245	0.476	0.487	0.469	0.497	0.244
GloVe + SVM	0.751	0.701	0.717	0.787	0.449	0.705	0.698	0.699	0.707	0.402	0.514	0.536	0.499	0.589	0.360
GloVe + DT	0.674	0.631	0.641	0.736	0.301	0.619	0.615	0.615	0.625	0.234	0.338	0.460	0.389	0.515	0.239
GloVe + RF	0.758	0.642	0.656	0.771	0.382	0.676	0.673	0.674	0.681	0.349	0.512	0.507	0.472	0.569	0.322
GloVe + XGB	0.734	0.690	0.704	0.777	0.422	0.690	0.687	0.688	0.694	0.377	0.543	0.547	0.534	0.591	0.359
Pre-trained Language Models															
BERT−base−cased	**0.767**	**0.786**	**0.775**	**0.807**	**0.553**	**0.771**	**0.774**	**0.770**	**0.771**	**0.545**	0.624	0.614	0.617	0.629	0.437
BERT−base−uncased	0.758	0.780	0.767	0.799	0.538	0.765	0.765	0.765	0.767	0.530	**0.643**	**0.632**	**0.635**	**0.650**	**0.466**
FinBERT	0.777	0.718	0.737	0.803	0.492	0.743	0.746	0.741	0.742	0.488	0.605	0.579	0.585	0.618	0.405

Notes: This table summarises the sentiment analysis results for all 7 experiments. The first 6 experiments are binary sentiment classifications, and the last one is a three-class sentiment classification introducing the neutral sentiment. The best result of each method category is in bold, and the best one of each experiment is underscored.

A Appendices for Chapter 2

A.1 Literature Review Summary

Paper	Text	Label	LX	WE	ML	PLM	Findings
Antweiler & Frank (2004)	Yahoo! Finance, Raging Bull message boards	Manually Annotated	-	BOW	NB, SVM	-	The effect of messages on stock returns is statistically significant but economically small. NB and SVM produce similar sentiment classification results.
Tetlock et al. (2008)	Wall Street Journal, Dow Jones News	Not Required	HIV4	-	-	-	Negative words in financial news stories forecast low firm earnings. Firms' stock prices reflect information embedded in negative terms with a slight delay. Negative words related to fundamentals have the most significant impact on earnings and return predictability.
Kothari et al. (2009)	Dow Jones News, Investext, Factiva, SEC EDGAR	Not Required	HIV4	-	-	-	Negative disclosures from business press sources drive up the cost of capital and return volatility.
Li (2010)	10-Ks and 10-Qs	Manually annotated	HIV4, LIWC, DICTION	BOW	NB	-	The tone of the forward-looking statements is positively associated with future earnings. The dictionary-based sentiment measures fail to predict future performance.

Paper	Text	Label	LX	WE	ML	PLM	Findings
Feldman et al. (2010)	10-Ks and 10-Qs	Not Required	HIV4, LM	-	-	-	The sentiment change in the MD&A section of the SEC filing is a significant predictor of short-term market reactions.
Rogers et al. (2011)	Corporate Disclosures	Not Required	DICTION, LM	-	-	-	Firms with more optimistic statements in their earnings announcements suffer from higher litigation risk.
Loughran & McDonald (2011)	10-Ks	Not Required	LM	-	-	-	Creation of an alternative negative word list that suits financial text better than Harvard IV-4.
Engelberg et al. (2012)	Dow Jones News	Not Required	HIV4, LM	-	-	-	The negative relation between short sales and future returns is more severe in negative news.
Hagenau et al. (2013)	German Adhoc, EuroAdhoc	Market Reaction	-	TF-IDF	SVM	-	Feature selection with bi-grams significantly improves sentiment classification accuracies of financial news, which help to improve stock price prediction.
Huang et al. (2014)	Investext	Manually Annotated	HIV4, LM, LIWC, DICTION	BOW	NB	-	Investors react more strongly to negative texts than to positive ones. NB is more effective in extracting opinions from analyst reports than dictionary-based approaches.
Tsai & Wang (2017)	10-Ks	Not Required	LM	-	-	-	Using regression and ranking methods, the experimental results show that soft information such as finance-specific sentiment lexicon strongly correlates to financial risks.
Sohangir et al. (2018)	StockTwits	Self-reported	VADER, WordNet	N-grams	LR, NB, SVM	-	VADER outperforms machine learning methods in extracting sentiment from financial social media.

Paper	Text	Label	LX	WE	ML	PLM	Findings
Sousa et al. (2019)	Financial News	Manually Annotated	-	BOW, TF-IDF	NB, SVM, TextCNN	BERT	Sentiment change in financial news achieved a 69% hit rate in predicting stock exchange variation. BERT has the highest accuracy of 82.5% among the sentiment analysis methods.
Araci (2019)	Financial Phrase Bank, FiQA	Manually Annotated	-	-	-	BERT, FinBERT	Further pre-training BERT on a subset of Reuters' TRC2 1.8M news articles, their vision of FinBERT improved BERT by 15% in accuracy.
Feuerriegel & Gordon (2019)	German Adhoc	Not Required	LM	-	-	-	High-dimensionality text input from financial news overfits machine learning models when predicting macroeconomic indicators; feature reduction can be achieved by mapping semantic categories onto latent structures.
Abbasi et al. (2019)	Twitter, Forum Channels	Manually Annotated		N-grams	RF, LR, SVM	-	User-generated content on social media can detect adverse events in advance, and false-positive rates are further reduced after including negative sentiment polarity in the models.
Renault (2020)	StockTwits	Self-reported	-	N-grams	MaxEnt, MLP, NB, RF, SVC	-	MaxEnt has the highest accuracy of 74.451% in sentiment analysis, followed by SVM with an accuracy of 74.292%. There is no substantial evidence that investors' opinion on social media helps predict significant capitalization stock returns daily.
Yang et al. (2020)	Financial Phrase Bank, AnalystTone, FiQA	Manually Annotated	-	-	-	BERT, FinBERT	Pre-training BERT-like on a sizeable financial corpus with 4.9 billion tokens (e.g. 10-Ks and 10-Qs, earnings call transcripts and analyst reports), FinBERT demonstrates superiority in handling financial sentiment classification tasks.

Paper	Text	Label	LX	WE	ML	PLM	Findings
Mishev et al. (2020)	Financial Phrase Bank, SemEval 2017 TASK 5	Manually Annotated	HIV4, LM	BOW, TF-IDF, Word2Vec, FastText, GloVe, ELMO	Attention, BiGRU, BiLSTM, CNN, Dense, LSTM, SVC, XGB	BERT, XLNet, XLM, FinBERT, DistilBERT, ALBERT, RoBERTa	The BERT models show better performances than machine learning methods, mainly because semantic meaning enriches text representation.
Stevenson et al. (2021)	Anonymous Credit Lender	Default Record	-	TF-IDF	LR, RF	BERT	BERT outperformed LR and RF in predicting small business loan default. Textual loan information produces relatively accurate predictions alone. However, it does not offer additional performance lift when used with structured data.
Jaggi et al. (2021)	Stocktwits	Market Reaction	-	BOW, TF-IDF	LR, NB, RF, XGB	BERT, FinBERT, ALBERT, FinALBERT	The best results are given by the BERT and NB across one year and two years of data for binary and 3-class sentiment classification.
Kriebel & Stitz (2021)	Lending Club	Default Record	-	TF-IDF, GloVe	CNN, RNN, CRNN, AE	BERT, RoBERTa	Textual information can improve credit default predictions. While machine learning models hold good results, BERT models have better performance in nearly all cases.
Frankel et al. (2022)	10-Ks, Earnings Call Transcripts	Market Reaction	HIV4, LM	N-grams	RF, sLDA, SVM	-	Machine learning methods are not only implementable and reliable measures of disclosure sentiment but also more potent than dictionary-based methods. RF is better at capturing disclosure sentiment than other machine learning methods.

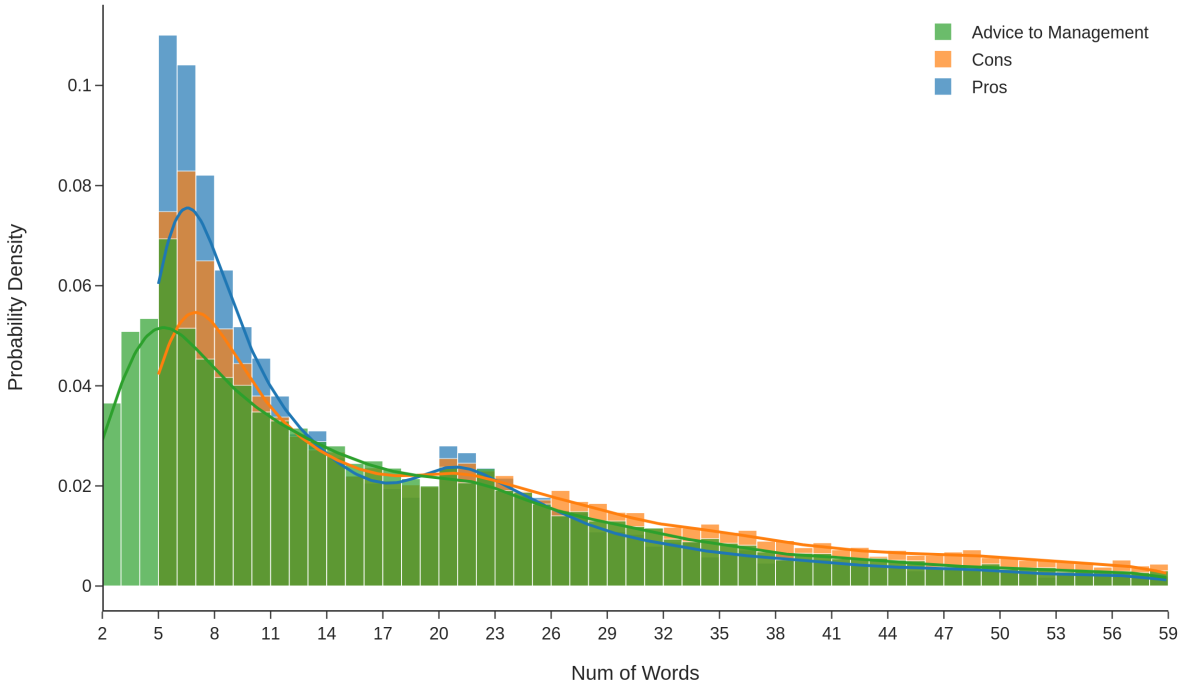
Paper	Text	Label	LX	WE	ML	PLM	Findings
Huang et al. (2023)	Analyst Reports	Manually Annotated	LM	BOW, N-grams	NB, SVM, RF, CNN, LSTM	FinBERT	FinBERT incorporates finance knowledge and can better summarise contextual information in financial texts.

Notes: LX: Lexicon-based Approach, WE: Word Embedding, ML: Machine Learning Approach, PLM: Pre-trained Language Model.

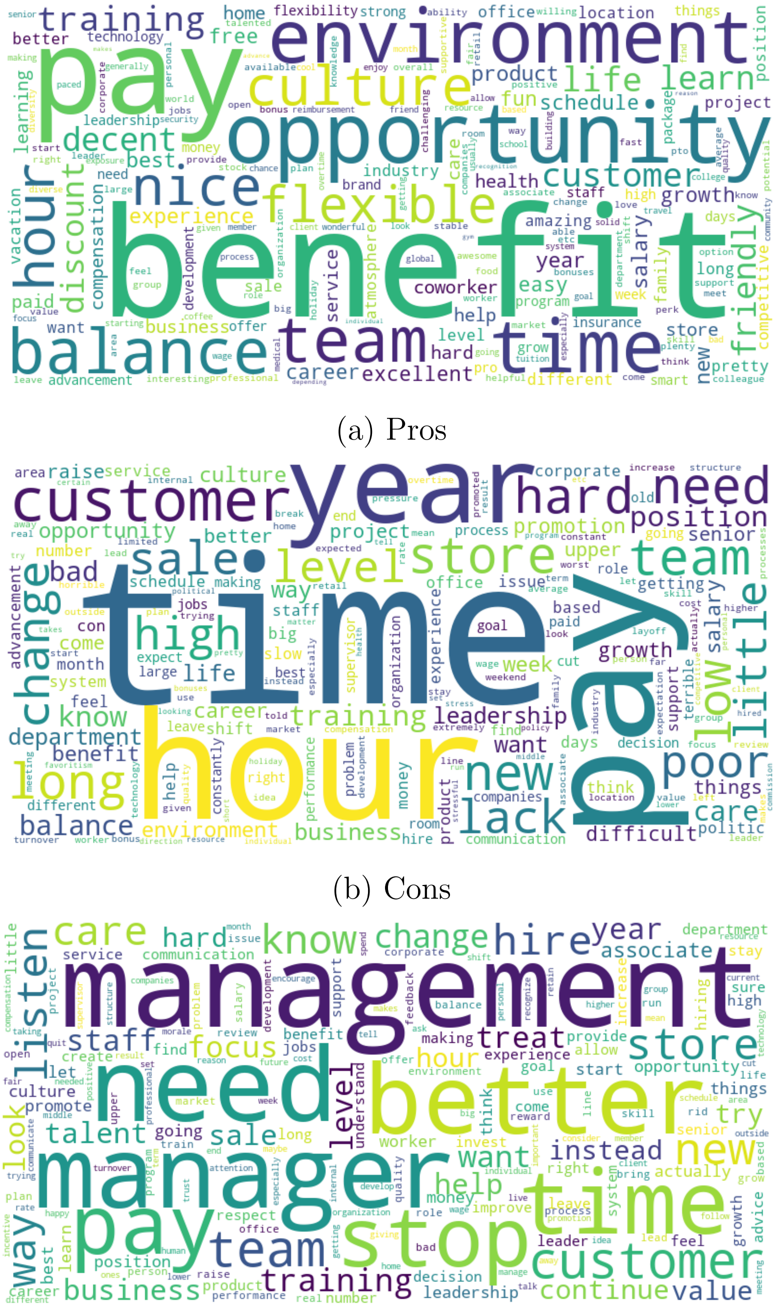
A.2 Glassdoor Employee Reviews

Reviewers have to provide text comments with a minimum of five words on *pros* and *cons* for the company. There is an optional text field to provide *advice to management*. From the distribution in Panel A, we can see the majority of reviews from *pros* and *cons* in our sample are between 5-8 words, but the probability of *pros* is higher than *cons* and *advice to management*. Comments with more than 23 words have a higher chance of coming from the *cons*. Panel B provides a visual representation of words in our sample data.

Panel A: Distribution of Review Lengths in *Pros*, *Cons* and *Advice to Management*.



Panel B: Word Cloud of *Pros*, *Cons* and *Advice to Management*.



(b) Cons

(c) Advice to Management

A.3 Models Description

Method	Model Description
Lexicon-based Approaches	
LM	Loughran-McDonald features + Lydia sentiment analysis system
HIV4	Harvard IV-4 dictionary features + Lydia sentiment analysis system
SentiWordNet	SentiWordNet features + Sentiment polarity
VADER	Compound valence score
Machine Learning Approaches	
BOW + LR	Count Vectorizer + Logistic Regression (C = 10)
BOW + NB	Count Vectorizer + Multinomial NB
BOW + SVM	Count Vectorizer + SVC (kernel ='linear', C = 1)
BOW + DT	Count Vectorizer + DT classifier (max_depth = 11)
BOW + RF	Count Vectorizer + RF classifier (n_estimators = 100)
BOW + XGB	Count Vectorizer + XGB classifier (learning_rate = 0.3, max_depth = 8)
TF-IDF + LR	TF-IDF Vectorizer + Logistic Regression (C = 10)
TF-IDF + NB	TF-IDF Vectorizer + Multinomial NB
TF-IDF + SVM	TF-IDF Vectorizer + SVC (kernel ='linear', C = 1)
TF-IDF + DT	TF-IDF Vectorizer + DT classifier (max_depth = 7)
TF-IDF + RF	TF-IDF Vectorizer + RF classifier (n_estimators = 100)
TF-IDF + XGB	TF-IDF Vectorizer + XGB classifier (learning_rate=0.3, max_depth = 10)
Word2Vec + LR	Word2Vec + Logistic Regression (C = 10)
Word2Vec + NB	Word2Vec + Multinomial NB
Word2Vec + SVM	Word2Vec + SVC (kernel ='linear', C = 1)
Word2Vec + DT	Word2Vec + DT classifier (max_depth = 4)
Word2Vec + RF	Word2Vec + RF classifier (n_estimators = 210)
Word2Vec + XGB	Word2Vec + XGB classifier (learning_rate = 0.3, max_depth = 10)
GloVe + LR	GloVe + Logistic Regression (C = 10)
GloVe + NB	GloVe + Multinomial NB
GloVe + SVM	GloVe + SVC (kernel ='linear', C = 1)
GloVe + DT	GloVe + DT classifier (max_depth = 6)
GloVe + RF	GloVe + RF classifier (n_estimators = 190)
GloVe + XGB	GloVe + XGB classifier (learning_rate = 0.3, max_depth = 6)
Pre-trained Language Models	
BERT−base−cased	110M parameters, pre-trained on cased Wikipedia and BookCorpus, learning rate = 2e-5, batch size = 32, epoch = 5
BERT−base−uncased	110M parameters, pre-trained on lower-cased Wikipedia and BookCorpus, text learning rate = 2e-5, batch size = 32, epoch = 5
FinBERT	110M parameters, futher pre-train BERT on Reuters TRC2 dataset (financial text), learning rate = 2e-5, batch size = 32, epoch = 5

A.4 Performance of Untuned DT

Decision trees have the lowest performance of all machine learning models, mainly because they are too easy to overfit the data. During training, a decision tree evaluates all possible splits and grows iteratively until it reaches the terminating nodes. Tuning the maximum depth in this case helps to reduce overfitting by limiting the expansion of the tree. Initially, we did not specify the value of maximum depth and found VADER outperform DT in most cases of the binary sentiment analysis, this proves that although the dictionary approaches are weak, they have the potential to improve if the dictionaries can be tuned subject to the task as well.

	Exp 1					Exp 2				
Method	Precision	Recall	F1-score	Accuracy	MCC	Precision	Recall	F1-score	Accuracy	MCC
VADER	0.782	0.768	0.765	0.768	0.550	0.764	0.694	0.690	0.721	0.453
BOW + DT	0.775	0.775	0.775	0.775	0.549	0.729	0.722	0.724	0.731	0.451
TF-IDF + DT	0.783	0.782	0.782	0.782	0.565	0.718	0.713	0.715	0.721	0.431
Word2Vec + DT	0.804	0.804	0.804	0.804	0.609	0.663	0.664	0.663	0.667	0.327
GloVe + DT	0.745	0.745	0.745	0.745	0.491	0.700	0.698	0.699	0.703	0.398

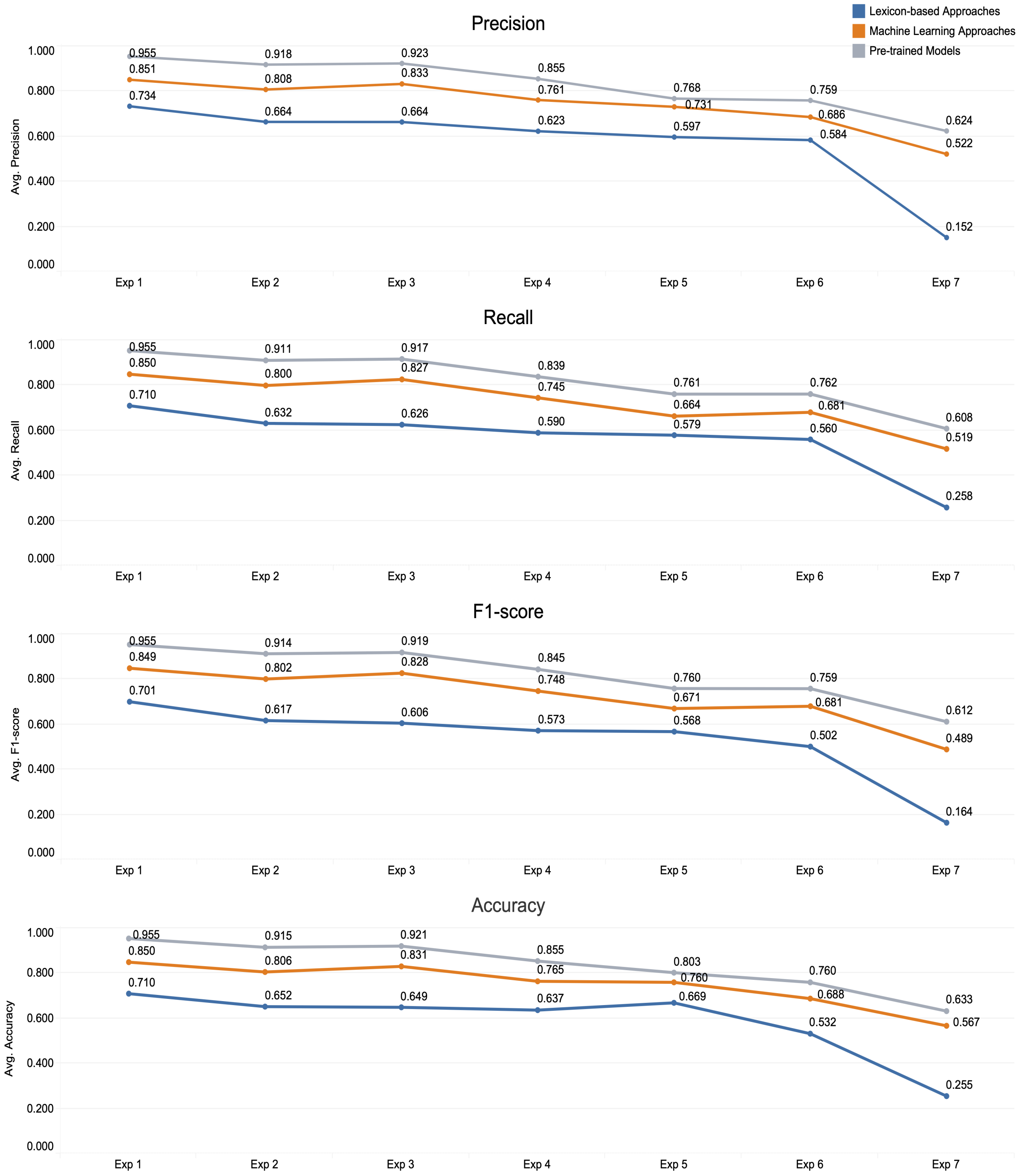
	Exp 3					Exp 4				
Method	Precision	Recall	F1-score	Accuracy	MCC	Precision	Recall	F1-score	Accuracy	MCC
VADER	0.756	0.669	0.658	0.700	0.416	0.712	0.621	0.610	0.686	0.320
BOW + DT	0.753	0.744	0.746	0.753	0.497	0.665	0.664	0.664	0.680	0.328
TF-IDF + DT	0.711	0.704	0.705	0.713	0.415	0.652	0.648	0.650	0.669	0.300
Word2Vec + DT	0.740	0.738	0.739	0.743	0.478	0.680	0.680	0.680	0.694	0.360
GloVe + DT	0.715	0.716	0.715	0.718	0.431	0.638	0.640	0.639	0.653	0.278

	Exp 5					Exp 6				
Method	Precision	Recall	F1-score	Accuracy	MCC	Precision	Recall	F1-score	Accuracy	MCC
VADER	0.671	0.609	0.616	0.733	0.273	0.644	0.579	0.500	0.540	0.213
BOW + DT	0.628	0.617	0.621	0.699	0.244	0.596	0.597	0.597	0.600	0.193
TF-IDF + DT	0.599	0.592	0.594	0.677	0.191	0.591	0.592	0.591	0.593	0.184
Word2Vec + DT	0.627	0.630	0.628	0.689	0.256	0.601	0.601	0.601	0.605	0.202
GloVe + DT	0.578	0.583	0.579	0.640	0.160	0.580	0.580	0.580	0.584	0.159

	Exp 7				
Method	Precision	Recall	F1-score	Accuracy	MCC
VADER	0.091	0.220	0.106	0.236	0.011
BOW + DT	0.446	0.443	0.443	0.472	0.178
TF-IDF + DT	0.440	0.440	0.440	0.464	0.170
Word2Vec + DT	0.443	0.441	0.442	0.461	0.172
GloVe + DT	0.425	0.424	0.425	0.441	0.141

A.5 Robustness Check

The figures below show the change in the mean values of the other evaluation metrics when text reviews are associated with star ratings over different intervals. The overall results are the same as for our main evaluation metric MCC, except that the mean recall for the BERT model and the mean recall and F1 scores for the machine learning approach is higher in Experiment 6 than in Experiment 5. This is because, in Experiment 5, the classes are skewed to positive sentiment whereas, in Experiment 6, the classes of positive and negative sentiment are more balanced. Researchers should cautiously choose the evaluation metrics when dealing with unbalanced data.


Lexicon-based Approaches
Machine Learning Approaches
Pre-trained Models
Precision
Avg. Precision
1.000
0.800
0.600
0.400
0.200
0.000
0.955
0.918
0.923
0.855
0.768
0.759
0.624
0.851
0.808
0.833
0.761
0.731
0.686
0.522
0.734
0.664
0.664
0.623
0.597
0.584
0.152
Exp 1
Exp 2
Exp 3
Exp 4
Exp 5
Exp 6
Exp 7
Recall
Avg. Recall
1.000
0.800
0.600
0.400
0.200
0.000
0.955
0.911
0.917
0.839
0.761
0.762
0.608
0.850
0.800
0.827
0.745
0.664
0.681
0.519
0.710
0.632
0.626
0.590
0.579
0.560
0.258
Exp 1
Exp 2
Exp 3
Exp 4
Exp 5
Exp 6
Exp 7
F1-score
Avg. F1-score
1.000
0.800
0.600
0.400
0.200
0.000
0.955
0.914
0.919
0.845
0.760
0.759
0.612
0.849
0.802
0.828
0.748
0.671
0.681
0.489
0.701
0.617
0.606
0.573
0.568
0.502
0.164
Exp 1
Exp 2
Exp 3
Exp 4
Exp 5
Exp 6
Exp 7
Accuracy
Avg. Accuracy
1.000
0.800
0.600
0.400
0.200
0.000
0.955
0.915
0.921
0.855
0.803
0.760
0.633
0.850
0.806
0.831
0.765
0.760
0.688
0.567
0.710
0.652
0.649
0.637
0.669
0.532
0.255
Exp 1
Exp 2
Exp 3
Exp 4
Exp 5
Exp 6
Exp 7

Chapter 3

Multilingual Sentiment Analysis with Glassdoor Employee Reviews

3.1 Introduction

Sentiment analysis is one of the most popular Natural Language Processing (NLP) applications designed to extract emotions from the text. For English texts, the most common sentiment analysis technique is the lexicon-based (or rule-based) approaches such as Harvard IV-4 dictionary (Stone & Hunt 1968) and Loughran-McDonald dictionary (Loughran & McDonald 2011), where positive and negative words are predefined and stored in separate lists. Machine learning (ML) and deep learning (DL) are also effective tools for sentiment analysis, which rely on word embeddings to extract features from the text. Word embeddings are learned representations that capture relationships between words, and words that have similar meanings also share a close word embedding. Some examples of popular pre-trained word embedding are Google's Word2Vec (Mikolov et al. 2013), Stanford's GloVe (Pennington et al. 2014), and Facebook's Fasttext (Bojanowski et al. 2017). Pre-training enables word embeddings to learn from large datasets, therefore, they capture the semantic meaning of a word better than traditional word embeddings such as bag-of-words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF). More recently, the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018) model uses a Transformer architecture which takes into account the sequence of all words and their positions in a sentence from both left-to-right and right-to-left contexts simultaneously. Compared to previous methods, BERT significantly improved the accuracy of sentiment analysis.

The proliferation of online and social platforms has led to an exponential growth in the volume and diversity of textual data. While sentiment analysis has traditionally been applied to English language data, it is important that this tool be extended to other languages to facilitate global communication and promote cultural diversity. Multilingual sentiment analysis offers several key benefits for international business and finance. First, it enables businesses to better understand the sentiment of their customers and stakeholders across linguistic and cultural barriers. This knowledge can inform strategic decision-making, such as product development, marketing, and customer service. Second, multilingual sentiment analysis can aid in detecting emerging trends and issues in global markets, enabling businesses to stay ahead of the trend and capitalise on opportunities. Third, it can help mitigate risks associated with negative sentiment and public relations crises, allowing businesses to respond in a timely and effective manner. Finally, multilingual sentiment analysis can facilitate cross-border collaboration and communication, leading to increased efficiency and productivity.

Despite these potential benefits, the development and implementation of multilingual sentiment analysis are not without challenges. One major obstacle is the lack of labelled data for languages other than English, which is essential for training accurate sentiment analysis models. Additionally, linguistic and cultural nuances pose significant challenges for accurately capturing sentiment across languages. However, these challenges are being actively addressed through ongoing research and development efforts.

Although translating multilingual text into English with the help of automatic neural machine translation (ANMT) is a common practice, it is not an ideal solution for sentiment analysis. This is because translation can often result in loss of meaning and cultural context, leading to inaccuracies in sentiment analysis. Additionally, sentiment analysis models trained on translated data may not perform as well as those trained on original multilingual data, as the nuances of each language can impact the sentiment of the text in different ways. Therefore, it is important to conduct sentiment analysis directly on the original multilingual text to ensure the accuracy and relevance of the results. Advances in natural language processing and machine learning techniques are enabling the development of multilingual sentiment analysis models that can accurately analyse sentiment in multiple languages simultaneously, further highlighting the importance of direct analysis of original multilingual data.

In this paper, we perform multilingual sentiment analysis using Glassdoor employee reviews written in German, French, Portuguese and Spanish in the original language and their translation to English, we explore the impact of ANMT on sentiment classification. We observe an increase in sentiment misclassification when using translated texts. Through further analysis of the characteristics of the texts, we realise the reason for this is not primarily translation quality. Instead, factors including the language, complexity, length and number of grammatical errors in the translated text are all responsible for misclassifications in the post-translation sentiment analysis. A recent study on the impact of machine translation on sentiment classification using pre-trained language models is Poncelas et al. (2020), which focuses on indirect machine translation and an automatic bidirectional Gated Recurrent Unit (GRU) classifier. In contrast to their work, we use an automatic system for translation and pre-trained language models for sentiment classification. To the best of our knowledge, our work is also the first to empirically assess the characteristics of the translated texts to assess the impact of ANMT on sentiment misclassification.

Compared to sentiment analysis using translated text, we show that a more efficient way to handle multilingual sentiment analysis is to use a pre-trained multilingual model. We demonstrate that applying DistilmBERT, mBERT and XLM-R directly to multilingual texts can produce highly accurate sentiment predictions. These models also have the advantage of zero-shot transfer, a method that transfers the knowledge learnt from a resource-abundant language to solve tasks in other low-resourced languages. The intuition behind this is that high-resource languages have the advantage of learning text representations from a more abundant amount of data which in turn leads to better performance when fine-tuning the models in downstream tasks. Furthermore, the practical significance of zero-shot cross-lingual transfer is to assist the cases where labelled multilingual texts are inadequate. In practice, labelling English texts can be easier than multilingual texts, and there are many readily available datasets of labelled English texts. With zero-shot transfer, we can utilise the labelled English texts and reduce the cost of labelling the multilingual texts. Lastly, we show evidence that zero-shot cross-lingual transfer does not rely on the vocabulary memorisation of the pre-trained language models but on the syntactic language similarity between languages.

This study makes a significant contribution to the literature on multilingual sentiment analysis. It provides a comprehensive empirical analysis of the impact of ANMT on sentiment analysis, highlighting the importance of analysing original multilingual text rather than translated text. Moreover, we identify key factors that impact sentiment misclassification rates, including prediction probabilities and text attributes such as language, sentiment, and readability. The study also demonstrates the benefits of zero-shot transfer which is particularly useful when labelled multilingual data is unavailable. Finally, we provide a practical guide for the effective use of zero-shot transfer, demonstrating that it is more effective for foreign languages that are more syntactically similar to English.

The structure of this paper is as follows. After this introduction, we review the previous work on pre-trained language models for English and multilingual aspects in Section 3.2. Next, we introduce the data in Section 3.3 and experimental setups in Section 3.4. We explore the effect of translation on multilingual sentiment classification, the factors contributing to this effect, and the practical implication of zero-shot cross-lingual transfer. The results are discussed in Section 3.5. Finally, we summarise the paper and make suggestions for future work in Section 3.6.

3.2 Related Work

3.2.1 Contextual Language Models

In recent years, pre-trained language models have received increasing attention and development in the NLP field. A language model is a probability distribution over sequences of words. By training on a large amount of texts, the model learns the occurrence and probability distribution of words and can make predictions of words in a given context. At the earlier stage, many studies focused on word embeddings such as Word2Vec and GloVe to capture the similarities between words. However, these static word embeddings are context-independent, for instance, the word embedding for a word is always the same even though in different contexts the meaning of a word may change. By contrast, the pre-trained language models take into account the contextual meaning of words. Dai & Le (2015) and Ramachandran et al. (2017) are the first to introduce the idea of pre-training a set of contextual representations and fine-tuning them to perform a board range of supervised downstream tasks. They improve sequence learning with the long short-term memory recurrent networks (LSTM RNNs) and pre-train sequence auto-encoder and encoder-decoder pairs.

Shortly after, Peters et al. (2017, 2018) propose a deep bidirectional language model, Embeddings from Language Models (ELMo). Compared to the previous unidirectional LSTM, ELMo uses a bidirectional LSTM which allows the model to view both right-to-left and left-to-right content and is, therefore, more efficient at capturing word-to-word relationships. ELMo concatenates the outputs of the forward and backward LSTMs at each position in the input text. This concatenation is done to capture information from both directions. While this approach does provide valuable context from both the left and right sides of a word, it has a limitation that it does not allow the model to consider interactions or dependencies between words in both directions simultaneously. The concatenation essentially fuses the two directions into a single vector, potentially losing some nuances of interaction. As an essential advancement and the foundation for numerous sophisticated NLP models such as BERT, Robustly Optimised BERT Pretraining Approach (RoBERTa) (Liu et al. 2019) and Cross-Lingual Language Model-RoBERTa (XLM-R) (Conneau et al. 2020), the development of the Transformer model and its encoder-decoder architecture allows for data and models to be trained in parallel.

In the BERT model, several Transformer encoders are stacked on top of each other over the training tasks of Masked Language Model (MLM) and Next Sentence Prediction (NSP). In MLM, 15% of the words in the input sequence are randomly replaced by a [MASK] token for BERT to predict given the context while the NSP performs a binary classification to indicate whether or not two sentences follow each other given a [SEP] token as a separator between sentences. The performance of BERT marked a significant milestone, leading to substantial improvements in 11 fundamental NLP tasks. The advent of BERT ushered in a new era, and since then a large number of pre-trained language models have emerged.

BERT has several variants, including the DistilBERT (Sanh et al. 2019), Enhanced Language Representation with Informative Entities (ERNIE) (Zhang et al. 2019), RoBERTa, A Lite BERT (ALBERT) (Lan et al. 2019), etc. DistilBERT introduces a triple loss combining language modelling, distillation and cosine-distance losses. DistilBERT is 40% smaller, 60% faster than BERT while maintaining 97% of its accuracy. ERNIE introduces a knowledge masking strategy, including entity-level and phrase-level masks incorporating knowledge graphs, to replace the random masks in BERT. RoBERTa makes several changes to the BERT model, including training the model longer with larger batches and more data, eliminating the NSP task, training on longer sequences, and dynamically changing the Mask position during pre-training. ALBERT proposes two parameter-reduction strategies to reduce memory consumption and speed up training. In addition, ALBERT also improves the NSP task of BERT through a self-supervised loss that focuses on modelling inter-sentence coherence.

3.2.2 Multilingual Language Models

One similarity between all the previously discussed language models is that they are all pre-trained on English corpus. Effectively their same tasks can be pre-trained on a multilingual corpus in order to process text in other languages in the downstream tasks. Multilingual BERT (mBERT) is an instance of BERT trained on the concatenation of the top 104 languages with the largest Wikipedias. The distilled version of mBERT maintains half the size of mBERT. Instead of doing its own pre-training, DistilmBERT inherits some of mBERT's parameters as initialisation and then performs knowledge distillation.

Although mBERT is pre-trained in more than 100 languages, the model itself is not optimised for multilingualism, most of the vocabulary is not shared across languages, so the cross-linguistic knowledge that can be learned is very limited. In response to this, the cross-lingual language model (XLM) modifies BERT in the following ways (Lample & Conneau 2019): Firstly, the XLM model uses byte-pair encoding (BPE) for subword tokenisation whereas BERT uses WordPiece. Both techniques used to handle the problem of tokenising text into smaller units. BPE is known for its dynamic vocabulary creation, making it adaptive to specific training data and better at handling out-of-vocabulary (OOV) words. WordPiece, on the other hand, relies on a fixed vocabulary which may have limitations in handling OOV words compared to BPE. BPE allows XLM to create a shared vocabulary that includes common subword units across languages. This shared vocabulary is essential for enabling cross-lingual transfer because it ensures that similar subword units in different languages are represented consistently.

Secondly, each training sample in XLM contains two sentences with the same meaning but in different languages, rather than one sample from the same language as in BERT. BERT aims at predicting masked tokens, whereas, in XLM we can use the context information of one language to predict the masked tokens of another language for each set of sentences. Since different random words in a sentence pair will be masked, the model can use the translation information to predict the token. XLM also considers language IDs and information about the order of tokens in different languages, i.e. positional coding, to help the model to learn the relationship between tokens in different languages. Inspired by RoBERTa, Conneau et al. (2020) scale up XLM by increasing the amount of data by several orders of magnitude. It is trained on 2.5TB CommonCrawl-100 data Wenzek et al. (2020) of 100 languages. Conneau et al. (2020) demonstrate their XLM-RoBERTa (XLM-R) significantly outperforms mBERT and XLM on a variety of cross-lingual benchmarks.

3.2.3 Zero-shot Cross-lingual Transfer

The pre-trained multilingual language models have shown success in transferring the knowledge of one language to another in a zero-shot manner, and the models can still perform well on languages not seen during training when they are fine-tuned for downstream tasks for one or more languages (Pires et al. 2019, Artetxe et al. 2020, Rezaee et al. 2021). Fine-tuning with domain-specific data helps to update the pre-trained parameters and to achieve more accurate representations, but it relies on labelled data to supervise the task. With review-based text, labelling can be easier if there were star ratings associated with the text, the numerical value can be a judge of sentiment. However, in practice, the majority of text such as annual reports, financial statements, tweets, regulations and news comes unlabelled. Manual annotation on multilingual text requires higher standards of human resources hence extensively expensive compared to the English language. Alternatively, zero-shot cross-lingual transfer can utilise the labelled data in resource-rich languages to make predictions in other languages that lack labelled data.

Zero-shot cross-lingual transfer relies on the help of the “pivot” or “bridge” language to transfer knowledge between source and target languages. Since the models are pre-trained on a large amount of text and multiple languages at the same time, the pivot language may share some structural similarities with the target language. Pires et al. (2019) find that even though mBERT is not pre-trained with explicit cross-lingual supervision, it is still able to perform cross-lingual generalisation and the transfer works best between typologically similar languages. Karthikeyan et al. (2020) point out that vocabulary memorisation of the mBERT plays little role in the zero-shot transfer. Lauscher et al. (2020) empirically correlates the zero-shot transfer performance of mBERT and XLM-R with linguistic proximity between source and target languages. Across several tasks Part-of-speech (POS) tagging, dependency parsing, named-entity recognition, natural language inference and question answering, they show high correlations between zero-shot transfer results and syntactic and phonological language similarities. To expand Lauscher et al. (2020)'s work, we evaluate the linguistic proximity between source and target languages over the task of sentiment classification.

3.3 Data

We use anonymous employee reviews from Glassdoor.com, one of the largest online review platforms. A registered and verified reviewer has to provide textual comments with a minimum of five words on the pros and cons for the company in their preferred language, options including English, French, German, Dutch, Portuguese, Spanish and Italian. There is also an optional text field to provide advice to management. In this paper, we only consider comments from *pros* and *cons* columns in multilingual sentiment analysis because the column names already reflect the ground truth sentiment of the text, thus making the labelling more accurate. We develop a web crawler using Python to collect non-English reviews of S&P 500 companies from 2008 to 2021 and label the text from *pros* column as positive and from *cons* column as negative, this way both sentiment classes are evenly distributed in our sample.

Glassdoor.com displays reviews only in the default language of the country associated with the visitor's domain. For instance, on glassdoor.com (US), glassdoor.co.uk (United Kingdom) and glassdoor.ca (Canada), etc, reviews are displayed in English whereas on glassdoor.de (Germany), reviews are displayed in German. In addition to reviews that were originally written in that language, Glassdoor also provides fast translation for non-local languages through a host on Google Translate API [1]. Nevertheless, visitors can check reviews in other languages by using the filter feature.

We collect 52,857 reviews in total and store them in two datasets to distinguish their format. Dataset 1 (D1): 31,024 reviews were originally written in foreign languages. Dataset 2 (D2): 21,833 reviews were originally written in English but translated by Glassdoor to a foreign language depending on the user domain. We use D1 as the primary source to study the impact of translation on sentiment classification, and D2 for robustness check. Since the reviews in D1 were not translated by Glassdoor, we translate them with the help of several translate AIPs that adopt ANMT: Google, Bing (Microsoft), Argos, Sogou and DeepL. We primarily use the Google translated reviews for the analysis to match Glassdoor's choice of translator, in the next section we will discuss how we use other translators to evaluate the translation quality. Figure 3.1 displays the word count distribution of positive and negative reviews for each language before and after translation.

[INSERT Figure 3.1 ABOUT HERE]

To ensure there are sufficient representations for each language, we limit the languages to which have at least 200 reviews in both datasets and end up with four foreign languages: German, French, Portuguese and Spanish. A summary of D1 and D2 is reported in Table 3.1.

[INSERT Table 3.1 ABOUT HERE]

[1] More information can be found in Local Language section of this page: `https://help.glassdoor.com/s/article/Ratings-on-Glassdoor?language=en_US`

3.4 Experimental Setup

One of the most straightforward solutions to dealing with multilingual texts is to translate them into languages we are familiar with, such as English. This is not only because it is easier to understand, but also because most sentiment analysis models are built in English, from the simplest dictionary methods to pre-trained language representations. However, the challenge we still face today is that despite how sophisticated NLP models are, it is inevitable that cultural meanings and linguistic details may be lost in translation. ANMT may be more accurate for word-to-word or phrase-to-phrase translation, but its accuracy might be affected when the text becomes complex and contains professional terminology. Therefore, it is in our interest to investigate the difference in the performance of a sentiment classifier on reviews written in the original language and reviews that have been translated, and to explain the cause of such difference if there was any.

The first step is to identify the cases where a model's sentiment prediction changed before and after translation. Using the data from D1, we fine-tune and evaluate three multilingual models: DistilmBERT, mBERT and XLM-R[2] for French, German, Portuguese, and Spanish reviews. For their translated texts, we use DistilBERT, BERT and XLM-R. The details of the models are summarised in Table 3.2. During fine-tuning, we use AdamW optimiser and select the learning rate of 2e-5, batch size of 32 and training epoch of 5. Since the positive and negative classes are evenly distributed in our data, we report the binary accuracy as the main evaluation metric and oversee the confusion matrix for additional explanation.

[INSERT Table 3.2 ABOUT HERE]

In the second step, we extract features from the translated texts that could potentially affect the sentiment classifier, and use them to train supervised machine

[2] This model can be downloaded from `https://huggingface.co/xlm-roberta-base`

learning models to classify those cases we identified in step one. The importance score of the features from the ML classifiers can help to explain the factors that affect the sentiment prediction before and after translation, and why the pre-trained language models failed on the translated texts. Our first focus is on translation quality, for which there are various methods and metrics. For instance, Multidimensional Quality Metrics (MQM) (Lommel et al. 2014) is a comprehensive framework designed to evaluate the quality of machine-generated translations. MQM employs a specific set of criteria to assess various aspects, including fluency, adequacy, and terminology accuracy, therefore it offers a comprehensive evaluation of translation quality. Other common approaches for assessing translation quality include human evaluation, where experts or bilingual reviewers compare the translation against the source text and rate its quality, and the use of automated metrics such as Bilingual Evaluation Understudy (BLEU) (Papineni et al. 2002), which calculate quality scores based on the similarity between the translation and a reference translation.

In this paper, we employ BLEU to evaluate the translation quality for its simplicity and computational efficiency. BLEU calculates a score based on the n-grams matches of a candidate sentence to a few references. The score ranges from 0 to 1, where 1 implies a perfect match and the candidate is identical to the references. Since Glassdoor uses Google's translate API as the default option, the English reviews used for sentiment classification were also translated using Google's service. Our objective is to evaluate the quality of this translation in comparison to other translation APIs. Therefore, for the same text, we consider Google's translation as the candidate and compare it with translations generated by Bing, Argos, Sogou, and DeepL.

Other features we include are the sentiment of the original review, denoted as *Senti*; the length of the review, denoted as *Len*; *Emo* and *Abb* are binary variables to recognise whether a review contains any emoticons and abbreviations, respectively; *Lang_deu, Lang_fra, Lang_por, Lang_spa* are language dummies for reviews originally written in German, French, Portuguese and Spanish, respectively; the number of noun phrases, *NP*; the number of verb phrases, *VP*; and the number of grammatical errors measured by Python's LanguageTool package, *GE*; the averaged sentiment prediction probabilities of pre-, post-translation and their difference, denoted as *Pre_prob*, *Post_prob* and *Diff_prob*, respectively.

Furthermore, we also include a few measures to assess the complexity and readability of the translated text using the python library Textstat[3]. We compute the Flesch Reading-Ease score, the Flesch-Kincaid Grade Level, the Fog Scale, and the Dale-Chall score, denoted as *FE*, *FG*, *FOG* and *DC*, respectively. The Flesch Reading-Ease score ranges from 0 to 100, a text with a higher score is easier to read. However, other measures are interpreted as the minimum grade level needed to comprehend the text, and a lower score indicates the text is less complex. Therefore, the results of Flesch Reading-Ease and other measures correlate approximately inversely: a text with a comparatively high value on the Flesch Reading-Ease score should have a lower value on the grade-level score.

Finally, we study the zero-shot cross-lingual transfer ability of DistilmBERT, mBERT and XLM-R on Glassdoor employee reviews by training and validating the models using English reviews (from D2), and evaluating them on German, French, Portuguese and Spanish reviews (from D1). Following Ahuja et al. (2022)'s work, we explain zero-shot performance by analysing the relatedness between the pivot and target languages from two aspects: the effect of *vocabulary overlap* and the *linguistic similarity.* The *vocabulary overlap* is defined as the percentage of unique tokens that are common to the vocabularies of both the pivot and target languages. Supposing V_p and V_p are the tokenised vocabularies of the pivot and target languages:

$$Overlap = \frac{V_p \cap V_t}{V_p \cup V_t} \tag{3.1}$$

The *linguistic similarity* is measured by typological vectors from lang2vec (Littell

[3] `https://pypi.org/project/textstat/`

et al. 2017), a release of the URIEL project[4] to enable multilingual NLP on less-resourced languages. We consider the following features from lang2vec: syntax, phonology and inventory. Syntax captures information about the grammatical relationships between words in a language. It helps understand how sentences are structured. Phonology refers to the study of the sound patterns and pronunciation rules of a language. It includes information about the sounds used in a language and their meanings or distinctions. Lastly, inventory relates to the presence of natural classes of sounds (consonants and vowels) in a language. It involves categorising sounds into groups based on shared phonological features.

3.5 Result Discussion

3.5.1 Translation Results

We evaluate the model's overall performance as well as the performance of each foreign language. Table 3.3 reports the accuracy change before and after translation, and the Newey-West adjusted t-statistics of their difference. In Panel A, DistilmBERT, mBERT and XLM-R demonstrate that on average using the multilingual models to directly classify the sentiment of foreign-language reviews can significantly reduce the relative classification error rate by 14% to 33%. This finding is robust, as shown in Panel B, reverse translation also increases the task's overall relative misclassification rate by 10% (XLM-R) to 46% (DistilmBERT).

[INSERT Table 3.3 ABOUT HERE]

In Panel A, German and Spanish translations have no significant effect on sentiment prediction, however, we find that the distilled BERT models are relatively 17% more accurate when Portuguese reviews have been translated into English. This result is counter-intuitive but their confusion matrices in Figures 3.2a and 3.2b suggest

[4] `https://www.cs.cmu.edu/~dmortens/projects/7_project/`

that DistilmBERT is less effective at correctly predicting the positive sentiment in Portuguese because it predicts more False Negatives (FNs) and less True Positives (TPs), which results in a lower recall, the ratio of true positives to total (actual) positives. After translating them into English, the positive sentiment becomes more easily detected by the model. For instance, a positive Portuguese comment "Salários, benefícios e jornada de trabalho." is predicted to be 51% negative by DistilmBERT, whereas DistilBERT predicts its translation "Salaries, benefits and working hours." to be 68% positive.

[INSERT Figure 3.2 ABOUT HERE]

In contrast to Portuguese, the BERT and XLM-R models suggest that translating French reviews to English increases the relative sentiment misclassification rate by around 30%. As shown in Figures 3.3c - 3.3f, the French-English translation caused a lower recall in BERT and a lower specificity (ratio of true negatives to total actual negatives) in XLM-R. In other words, after the French-English translation, BERT struggles more FNs whereas XLM-R mainly fails in False Positives (FPs). For instance, DistilmBERT, mBERT and XLM-R corrected predicted the negative French comment "Evolution de carrière au bon vouloir du service dans lequel on travaille (disparité)." with the probabilities of 73%, 64% and 100%, respectively. However, after translating it into English "Career evolution at the good will of the service in which we work (disparity).", all models falsely identified it to be above 98% positive.

[INSERT Figure 3.3 ABOUT HERE]

In Panel B, the results show that translating English reviews into German, French and Spanish also causes a decrease in the accuracy of the sentiment classifier. The difference in accuracy of Spanish-English translations was significant across all three types of models, so in the case of Spanish for example, the confusion matrices in Figures 3.4a - 3.4f explain that the performance dropped using the translated reviews was because of the reduced recall rates and the models have trouble to correctly predict positive sentiment with the translated Spanish. The overall performance in Table 3.3 confirmed our assumption that performing sentiment classification using reviews in their original language would be more accurate than in the translated one, and it affects all four languages we examined by different models. The confusion matrices of other entries from Table 3.3 are displayed in the Appendix.

[INSERT Figure 3.4 ABOUT HERE]

3.5.2 Feature Analysis Results

In the test set from the previous section, the sentiment of 996 out of 9,309 reviews was misclassified by one of the models after translation. Based on the features we extracted, as discussed previously, we train a random forest (RF) and a decision tree (DT) classifier to predict these instances. We adopt the under-sampling methods to handle the imbalanced classes. Table 3.4 shows evaluation metrics of RF and DT, the misclassified sentiment can be relatively accurately distinguished by these two classifiers with the F-1 scores of 0.93 and 0.88, respectively. The DT works by recursively splitting the decision nodes that minimise the impurity of the split, therefore, during the training, we can compute how much each feature contributes to decreasing the weighted impurity and this result is the feature importance. RF is a collection of single DTs trained on a randomly selected subset of features, the feature importance of RF can be calculated by averaging the decrease in impurity over trees. Figure 3.5 displays the feature importance of RF and DT which are useful to explain the sentiment misclassification.

[INSERT Table 3.4 ABOUT HERE]

[INSERT Figure 3.5 ABOUT HERE]

Higher scores are more important in explaining post-translation sentiment misclassification. First, the most important features are the prediction probabilities of pre- and post-translation. They reflect how certain a model is when it predicts the sentiment for a review. By examining the mean of the features for the misclassified cases in Table 3.5 as well as their original predictions, we find that incorrect predictions have lower prediction probabilities than cases where the predictions are consistent and correct before and after translation. Also in general when a model makes the same prediction for original and translated reviews, the probabilities of such prediction tend to be lower for the translated texts, implying that the model behaved less confidently when processing the translated text. Second, the sentiment of the original text is also important, as discussed in the previous section, some models struggle with FNs and FPs because sentiment can be interpreted differently across languages. A positive sentence in English may be neutral or negative in other languages, this is more challenging to explain from the translation perspective but it is reflected in the prediction probabilities. Third, from the language perspective, French and Portuguese have higher scores than German and Spanish. The results from Table 3.3 Panel A also match this finding that the performance of French and Portuguese has been significantly impacted by translation.

[INSERT Table 3.5 ABOUT HERE]

In addition, the readability and the number of noun phrases in the translated text also is an important factor. Through a manual review of the misclassed cases, we discover that sentiment misclassification is more likely to occur on less complex and shorter texts because they provide less meaningful contextual information. Many short reviews are just phrases and incomplete sentences thus making it more difficult to judge the sentiment. Furthermore, We assume that the presence of abbreviations is a challenge for translators because they require knowledge of the cultural context. For instance, in French “RH” is short for “Ressources Humaines” which means “Human Resources” or “HR” in English. In the absence of context, the translators failed to accurately translate “RH”. In the other French example, “CE pas si mauvais que ça.” could be translated as “CE is not that bad.” if the abbreviation was kept in its original form. A slight change in the casing, “Ce”, on the other hand, corresponds to the English pronoun “it”. We find of the translators all ignored the upper casing letter in the spelling and translated the word incorrectly. Google translates this sentence to “This is not as bad as that.”, Bing considers “CE” as a spelling error and translates it as “It’s not that bad.”. In the context of an employee review, “CE” in fact stands for “Conseil d’Entreprise” meaning employer’s committee. However, despite such translation mistakes, grammatical errors are considered more important for the cause of post-translation sentiment misclassification. Poncelas et al. (2020) explain that once translation quality reaches a certain threshold, it is not correlated to the performance of the sentiment classifier.

3.5.3 Zero-shot Transfer Results

We use English as the pivot language to train DistilmBERT, mBERT and XLM-R, and evaluate the sentiment prediction on German, French, Portuguese and Spanish reviews. To eliminate the variance coming from the model itself, we run each model five times and report their averaged results in Table 3.6. All of the multilingual models can produce relatively good results, with an overall accuracy of 77.9%, 85.7% and 95% for DistilmBERT, mBERT and XLM-R, respectively. DistilmBERT has the weakest zero-shot prediction ability in general and particularly in German reviews, but it produces similar results for other languages. As the model size increases, the zero-shot cross-lingual transfer ability becomes significantly stronger. The XLM-R model performs particularly well, with comparable accuracy to the previous overall result in Table 3.3 Panel A when it is fine-tuned on multilingual texts. Looking at the results for each language, we find that mBERT and XLM-R are the most accurate in predicting sentiment for German and the least accurate for French. This is possible because syntactically and phonologically, German and English are the closest, while French and English are the farthest.

[INSERT Table 3.6 ABOUT HERE]

Recall the vocabulary size of the models from Table 3.1 DistilmBERT and mBERT use WordPiece tokenizer (Wu et al. 2016*b*) with the vocabulary of 110K tokens. XLM-R adopts the SentencePiece tokenizer (Sennrich et al. 2016) with the vocabulary of 250K tokens. We compute the cosine distance between English (the source language) and German, French, Portuguese and Spanish (the target languages) for each typological vector from lang2vec as a similar score, then correlate them and the overlap rate with the performance of DistilBERT, mBERT and XLM-R using the Pearson's correlation coefficient. The results in Table 3.7 indicate that zero-shot transfer is highly correlated to syntactic language similarity but not to other features. Furthermore, zero-shot performance is not significantly correlated to vocabulary overlaps. The same finding has also been concluded by Karthikeyan et al. (2020) that vocabulary memorisation of the pre-trained language representations plays little role in zero-shot transfer.

[INSERT Table 3.7 ABOUT HERE]

3.6 Conclusion and Future Work

To break down language barriers for worldwide communication, the transfer and exchange of information between domains require the use of NLP techniques to process texts across different languages. This paper deploys three advanced pre-trained NLP models DistilmBERT, mBERT and XLM-R for the multilingual sentiment analysis using Glassdoor employee reviews in German, French, Portuguese and Spanish as well as their translation to English. All models have an overall accuracy of over 94%, however, there is a statistically significant decline in the overall accuracy of 0.8% to 1.5% when using the translated texts. This finding is robust to the same task with the foreign-to-English translations. In terms of language, the accuracy of the DistilmBERT model improved by 0.7% after translating Portuguese into English, as the model had difficulties in identifying positive sentiment and therefore has a low recall. After translation, positive sentiment became more easily detected by the model. In the French-English translation, BERT has a lower recall and failed in identifying FNs, while XLM-R has a lower specificity and fails in identifying FPs.

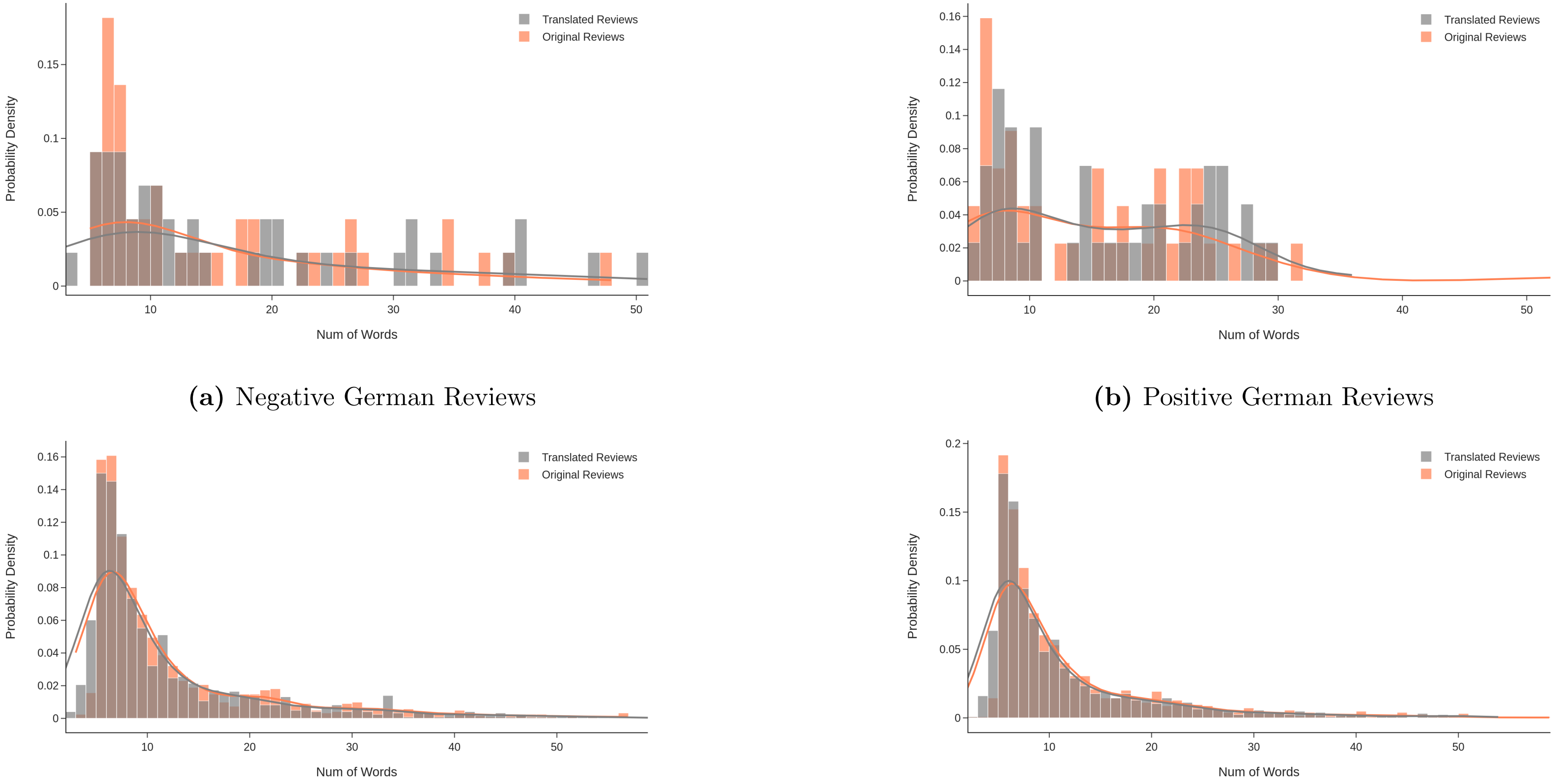
We extract the characteristics from the reviews and build two ML classifiers to further investigate the factors that contributed to sentiment misclassification. We report that the prediction probabilities before and after the translation, and their difference are the most important elements in identifying sentiment misclassification. When the predicted class are the same, the prediction probability for the translated text tends to be slightly lower than for the original text. In addition, other elements such as the original sentiment of the text, the language, and the complexity and readability are all factors affecting the accuracy of a sentiment classifier. The results of this analysis also suggest that the grammatical errors in the translated text are more important than the quality of the translation.

Finally, we demonstrate the zero-shot ability of DistilmBERT, mBERT and XLM-R, these cross-lingual models are useful when the foreign language is lean-resourced or lacks of labelled data. Zero-shot learning is a well-established technique, with the knowledge transfer between languages we can extend more practical applications for processing multilingual texts. Our empirical results confirm that even when these models are fine-tuned using English reviews only, they still produce relatively accurate predictions for the sentiment of German, French, Portuguese and Spanish reviews. We compute the similarity between English and other languages and the models' vocabulary overlaps. The results indicate that zero-shot transfer of mBERT and XLM-R are highly correlated to syntactic language similarity but not vocabulary overlaps.

Our research highlights an opportunity for future studies in the business, management, and finance domains to expand their interest in mining multilingual texts. Although ANMT is a convenient tool, its computational cost and time requirements significantly increase when dealing with long texts or large quantities of text. Furthermore, we found that while translation quality remains high, sentiment may shift after translation, as indicated by a slight decrease in the prediction probability of translated text. Directly processing multilingual texts with multilingual language models, therefore, offers a more efficient approach to sentiment classification problems. To further improve the accuracy and efficiency of multilingual sentiment analysis, future work could explore few-shot transfer by including additional fine-tuning on a few target-language instances. We believe that these advanced NLP models hold great potential for handling complex real-world applications and can be widely generalised across various disciplines and fields.

3.7 Figures and Tables for Chapter 3

Figure 3.1: Review Length Distribution Before and After Translation.



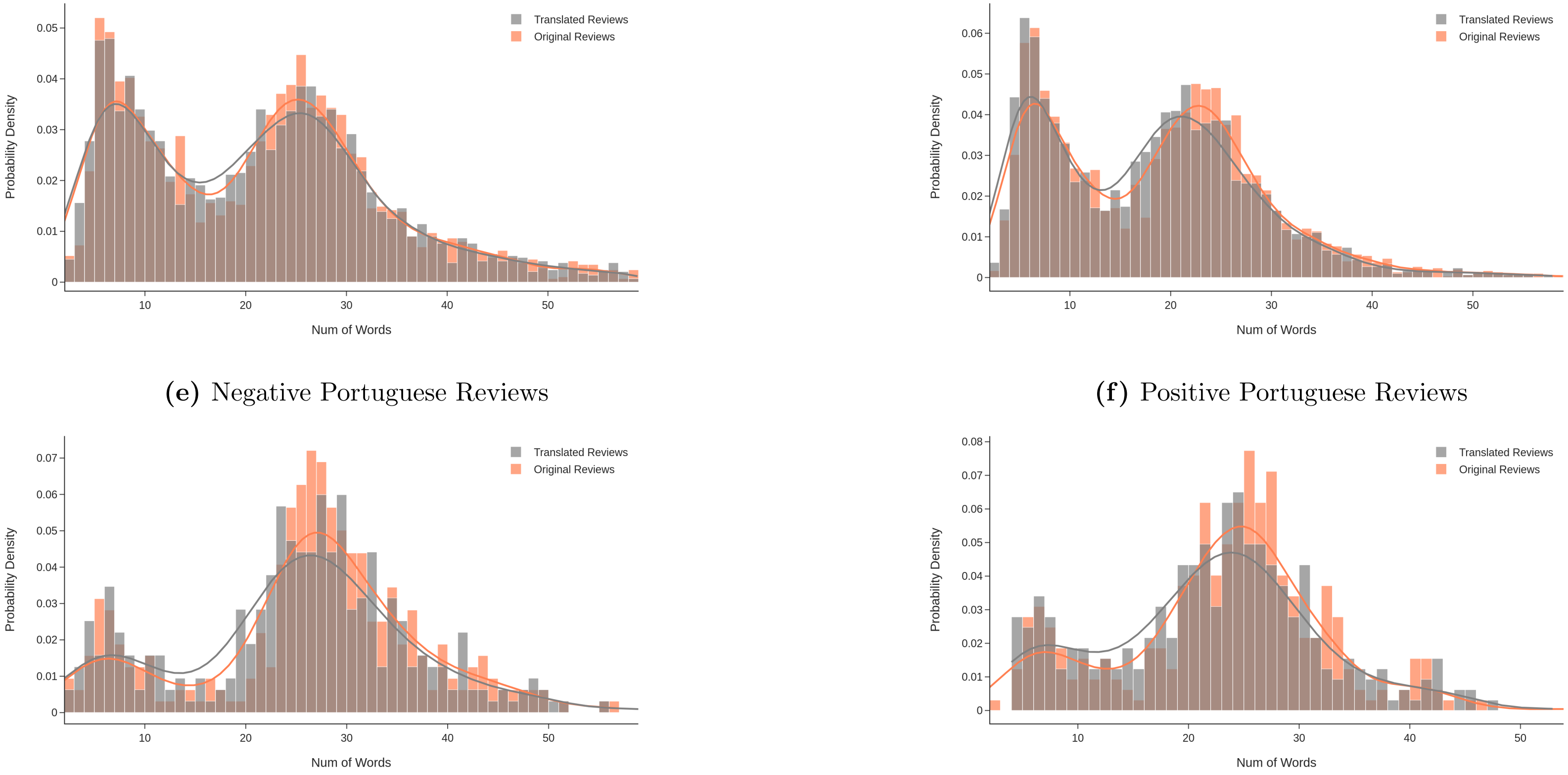
(a) Negative German Reviews

(b) Positive German Reviews

(c) Negative French Reviews

(d) Positive French Reviews

Figure 3.1: Review Length Distribution Before and After Translation. *(cont.)*



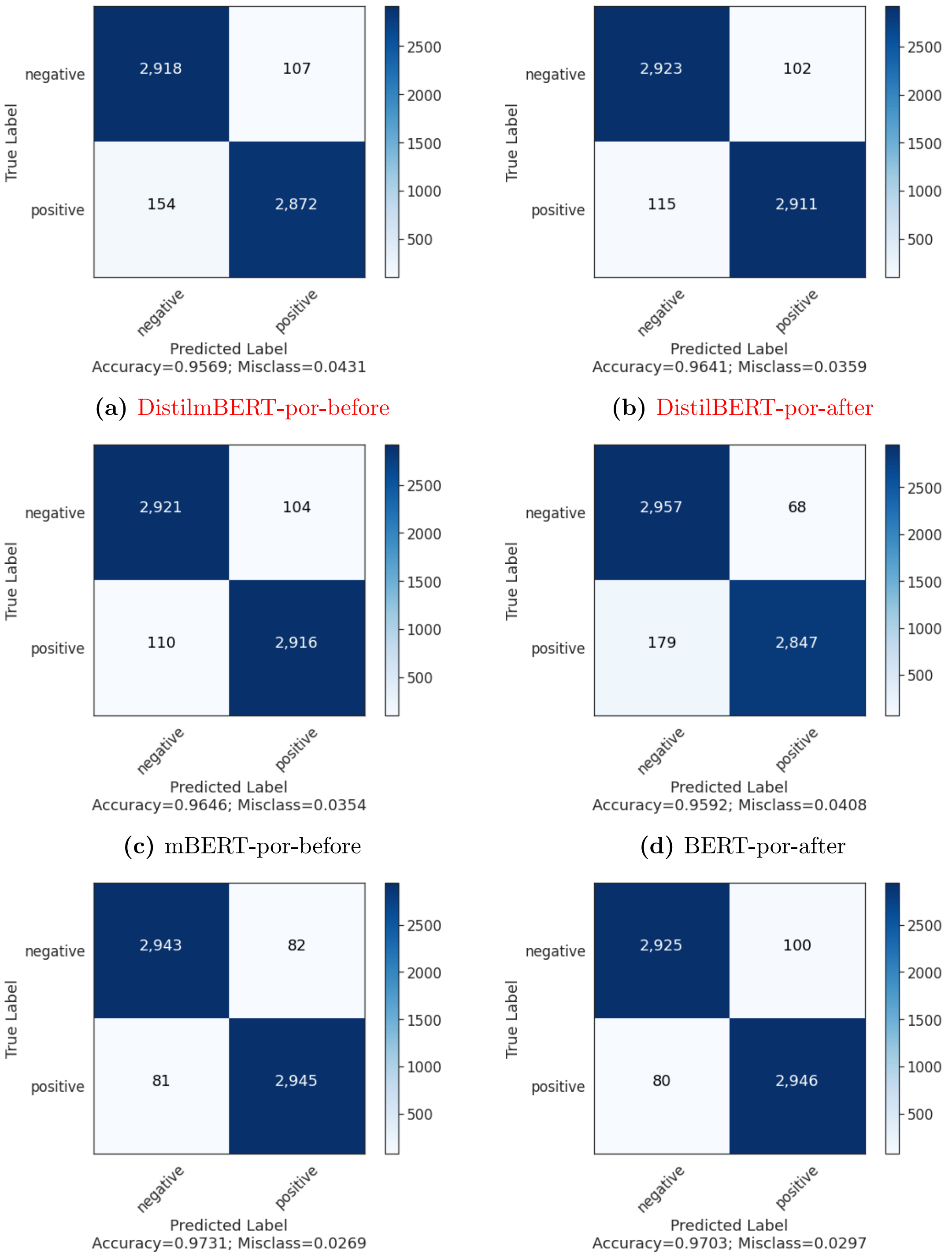
(e) Negative Portuguese Reviews

(f) Positive Portuguese Reviews

(g) Negative Spanish Reviews

(h) Positive Spanish Reviews

Figure 3.2: Portuguese-English Translation Analysis



(a) DistilmBERT-por-before

(b) DistilBERT-por-after

(c) mBERT-por-before

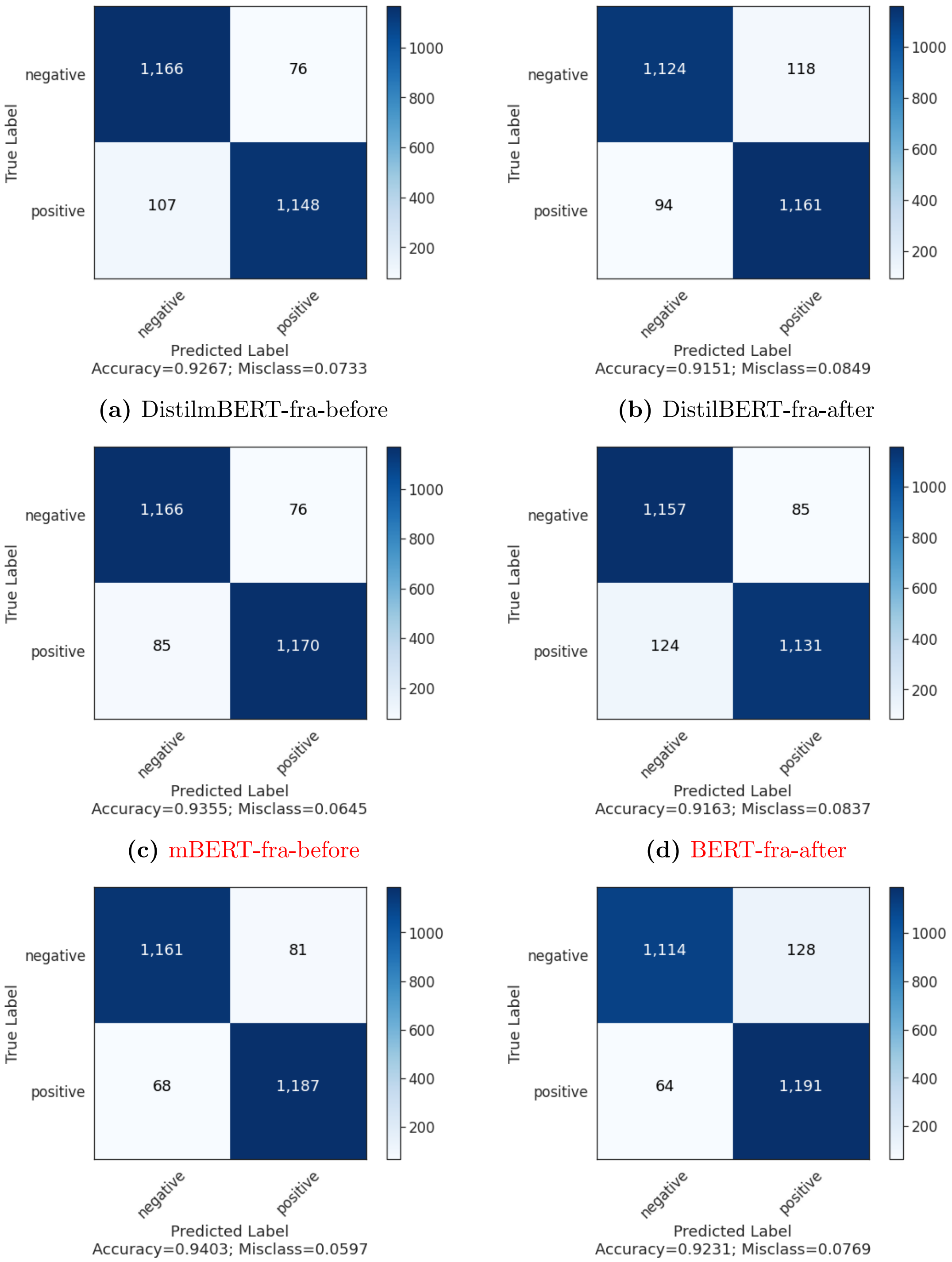
(d) BERT-por-after

(e) XLM-R-por-before

(f) XLM-R-por-after

Notes: The caption in red indicates the change in accuracy after translation is statistically significant.

Figure 3.3: French-English Translation Analysis



(a) DistilmBERT-fra-before

(b) DistilBERT-fra-after

(c) mBERT-fra-before

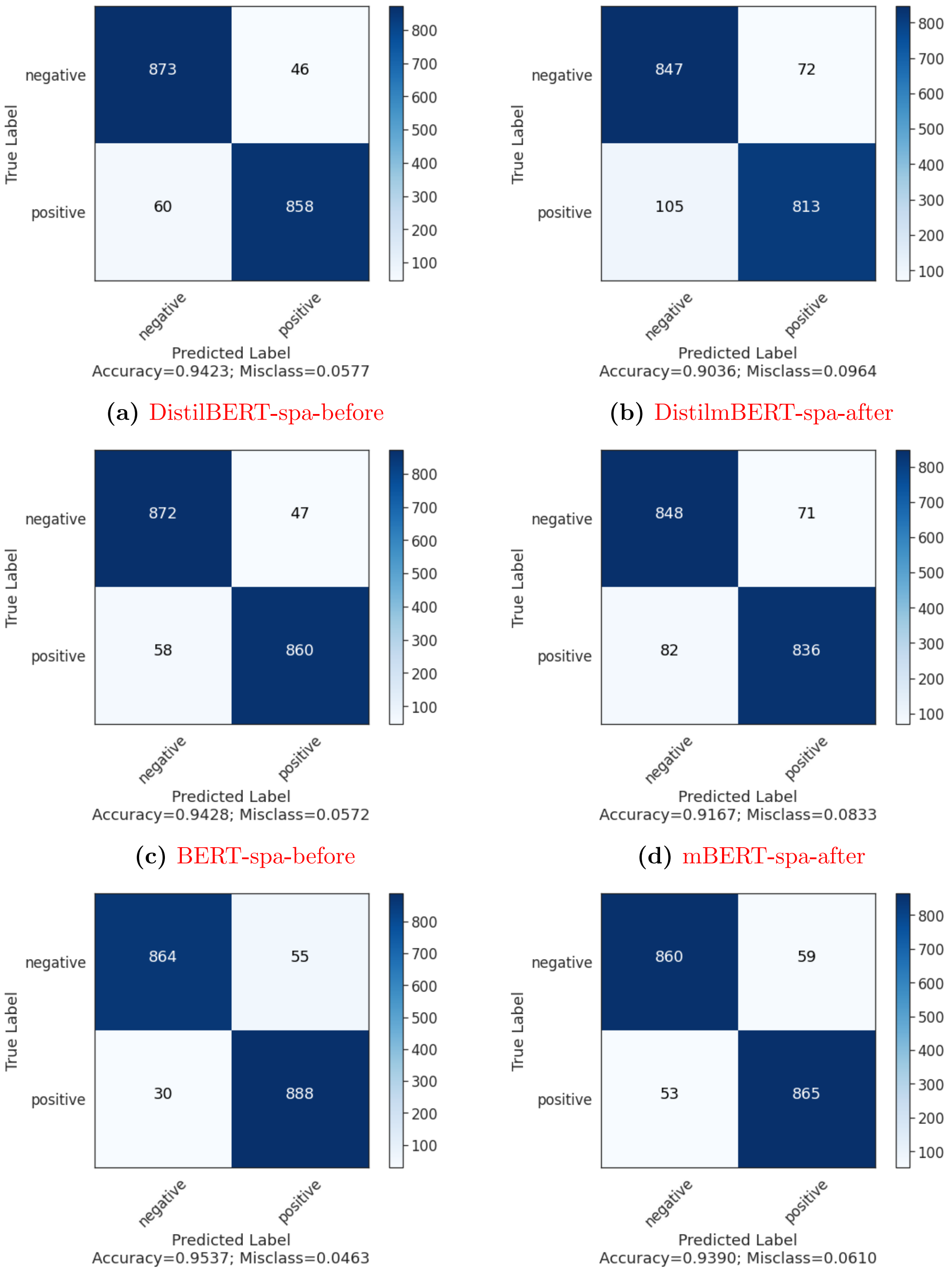
(d) BERT-fra-after

(e) XLM-R-fra-before

(f) XLM-R-fra-after

Notes: The caption in red indicates the change in accuracy after translation is statistically significant.

Figure 3.4: English-Spanish Translation Analysis



(a) DistilBERT-spa-before

(b) DistilmBERT-spa-after

(c) BERT-spa-before

(d) mBERT-spa-after

(e) XLM-R-spa-before

(f) XLM-R-spa-after

Notes: The caption in red indicates the change in accuracy after translation is statistically significant.

Figure 3.5: Feature Importance

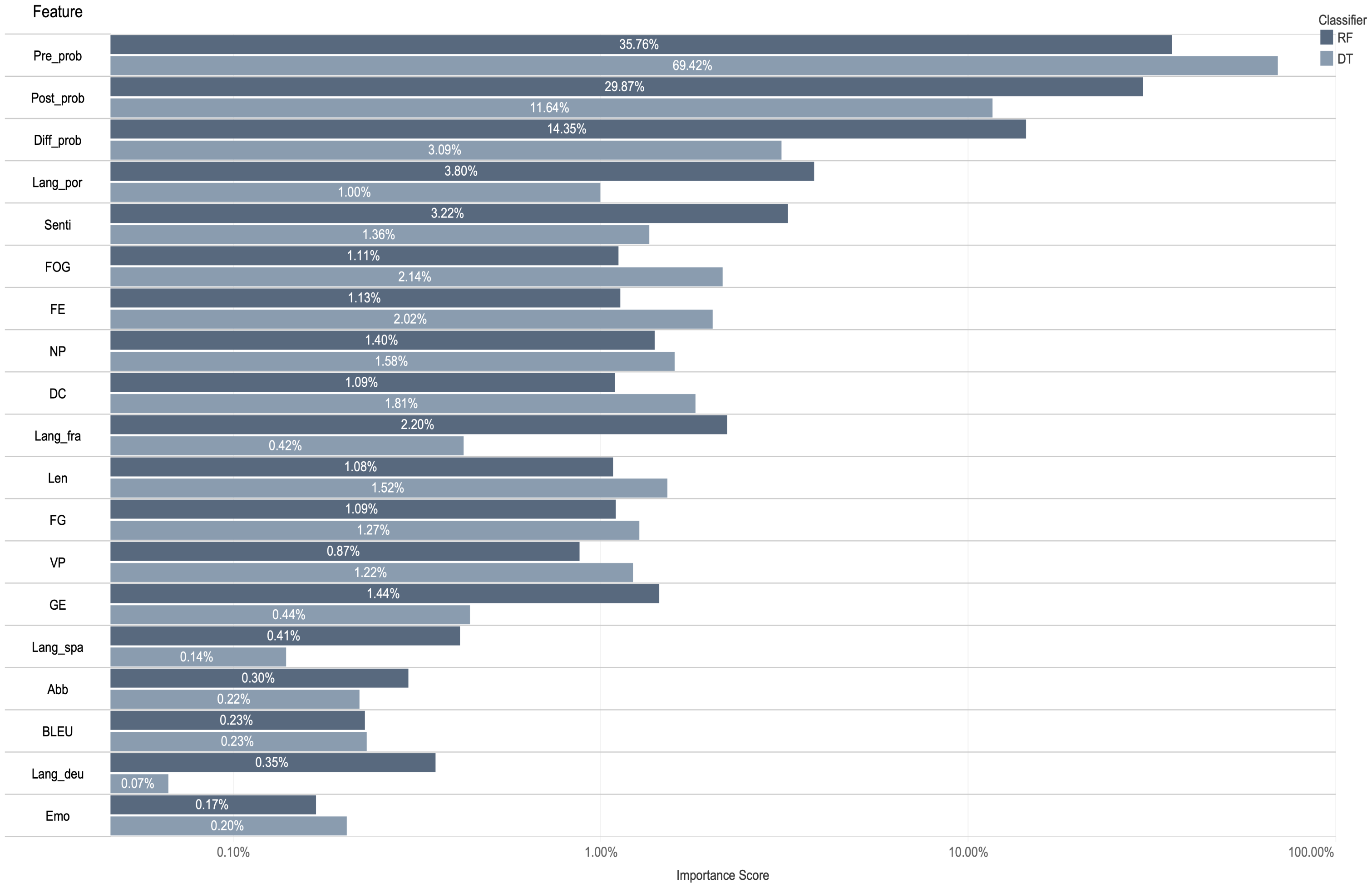


Table 3.1: Description of Datasets

	Dataset 1 (D1)	**Dataset 2 (D2)**
Original Language:	Foreign	English
Translation	Not Available	Provided by Glassdoor
Size	31,024	21,833
Language	Portuguese: 20,039 (65%) French: 8,170 (26%) Spanish: 2,525 (8%) German: 290 (1%)	Portuguese: 2,196 (10%) French: 8,058 (37%) Spanish: 6,042 (28%) German: 5,537 (25%)
Demo	Original: Pros: Projets intéressants, des technologies innovantes. Cons: Évolution limitée par le management.	Original: Pros: Lots of material available to develop new skills. Cons: Shadow decisions in top management. Translated to: Pros: Beaucoup de matériel disponible pour développer de nouvelles compétences. Cons: Décisions fantômes au sein de la haute direction.

Notes: Glassdoor reviews were not provided with foreign to English translations and we later translated the reviews in D1 via Google, Amazon, Argos, Sogou and the DeepL Translate APIs.

Table 3.2: Model Description

Model	**#L**	**#H**	**#A**	**#Params**	**Lg**	**#V**
distilbert-base-cased	6	768	12	66M	eng	30k
bert-base-cased	12	768	12	110M	eng	30k
distilbert-base-multilingual-cased	6	768	12	134M	multi	110k
bert-base-multilingual-cased	12	768	12	177M	multi	110k
xlm-roberta-base	12	768	12	270M	multi	250k

Notes: #L = the number of layers; #H = hidden size; #A = number of attention heads; #Params = number of parameters; Lg = Language of the pre-trained corpus; #V: vocabulary size.

Table 3.3: Changes in Sentiment Prediction Accuracy in Percentage after Translation

Panel A: Foreign-to-English Translation (D1)

	DistilmBERT	DistilBERT		mBERT	BERT		XLM-R	XLM-R	
	Before	After	Diff	Before	After	Diff	Before	After	Diff
deu	86.02	92.47	-6.45	92.47	94.62	-2.15	94.62	90.32	4.30
t-stats			(-1.42)			(-0.59)			(1.11)
fra	92.67	91.51	1.16	93.55	91.62	1.92***	94.03	92.31	1.72**
t-stats			(1.52)			(2.59)			(2.41)
por	95.69	96.41	-0.73**	96.46	95.92	0.55	97.31	97.03	0.28
t-stats			(-2.05)			(1.57)			(0.93)
spa	93.40	94.15	-0.7	94.60	94.90	-0.30	96.25	95.35	0.90
t-stats			(-0.57)			(-0.25)			(0.82)
Overall	94.62	93.87	0.75**	95.51	94.00	1.51***	96.33	95.57	0.75***
t-stats			(2.20)			(4.64)			(2.60)

Panel B: English-to-Foreign Translation (D2)

	DistilBERT	DistilmBERT		BERT	mBERT		XLM-R	XLM-R	
	Before	After	Diff	Before	After	Diff	Before	After	Diff
deu	93.32	89.60	3.72***	93.26	90.72	2.54***	94.50	94.21	0.30
t-stats			(3.88)			(2.73)			(0.37)
fra	93.49	91.44	2.05***	93.49	92.81	0.68	94.50	94.34	0.16
t-stats			(2.74)			(0.95)			(0.25)
por	94.28	92.99	1.29	94.99	95.85	-0.86	96.28	96.71	-0.43
t-stats			(0.99)			(-0.77)			(-0.44)
spa	94.23	90.36	3.86***	94.28	91.67	2.61***	95.37	93.90	1.47**
t-stats			(4.40)			(3.10)			(1.98)
Overall	93.73	90.84	2.89***	93.81	92.29	1.52***	94.92	94.43	0.49
t-stats			(6.28)			(3.46)			(1.27)

Notes: This table summarises the change in accuracy in (%) per foreign language as well as the model's overall performance using the original and the translated reviews. In Panel A, text is translated from foreign languages (before) to English (after). In Panel B, text is translated from English (before) to foreign languages (after). Newey-West adjusted t-statistics are given in the parentheses, ** and *** indicate the significance at the 5% and 1% levels, respectively. Null hypothesis: Diff = 0.

Table 3.4: Evaluation Metrics of RF and DT

	Precision	**Recall**	**F1-score**	**Accuracy**
RF	0.933	0.933	0.933	0.933
DT	0.881	0.881	0.881	0.881

Table 3.5: Mean Value of Features

	Misclassified = No	**Misclassified = Yes**
BLEU	0.970	0.936
Senti	0.502	0.483
Len	19.523	15.950
Emo	0.046	0.052
Abb	0.085	0.101
Lang_deu	0.008	0.029
Lang_fra	0.257	0.365
Lang_por	0.665	0.517
Lang_spa	0.070	0.088
NP	6.228	4.936
VP	1.957	1.742
GE	0.270	0.340
Post_prob	0.906	0.846
Pre_prob	0.992	0.906
Diff_prob	-0.085	-0.061
FE	42.618	51.546
FG	10.701	9.119
FOG	13.320	11.577
DC	10.613	10.184

Notes: Misclassified = No: Using the translated text, the predictions of DistilmBERT, mBERT and XLM-R were consistent and correct; Misclassified = Yes: Using the translated text, at least one model misclassified the sentiment.

Table 3.6: Accuracy of Zero-shot Prediction

	DistilmBERT	**mBERT**	**XLM-R**
deu	0.768	0.862	0.957
fra	0.783	0.845	0.929
por	0.782	0.860	0.964
spa	0.783	0.861	0.951
Overall	0.779	0.857	0.950

Table 3.7: Pearson's Correlation Between the Pivot and Target Languages in Zero-shot Learning

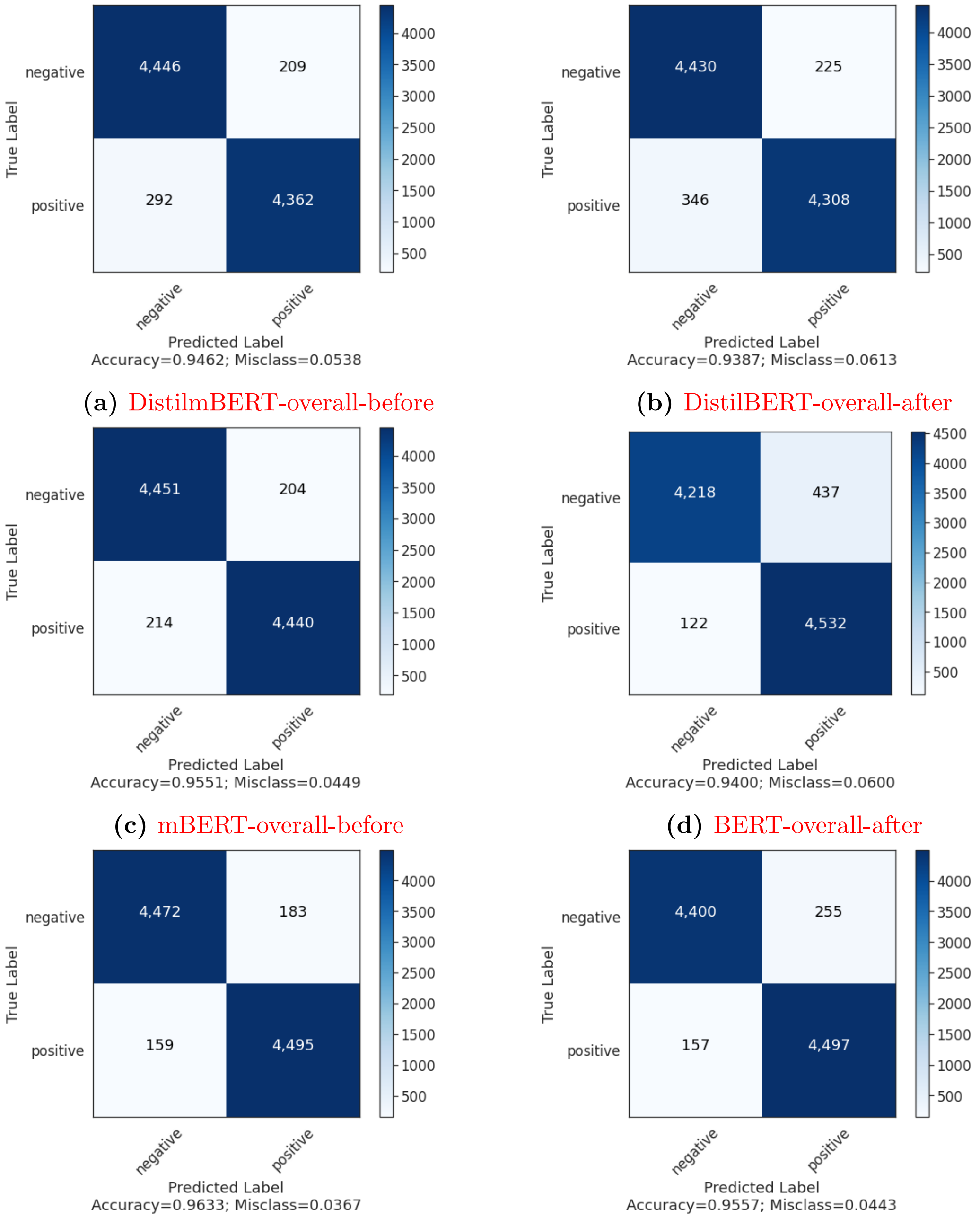
	Syntax	**Phonology**	**Inventory**	**Overlap Rate**
DistilmBERT	0.352	-0.635	-0.509	0.393
p-value	(0.648)	(0.365)	(0.492)	(0.607)
mBERT	0.967**	0.276	0.326	0.574
p-value	(0.033)	(0.724)	(0.674)	(0.426)
XLM-R	0.904**	0.808	0.206	0.309
p-value	(0.096)	(0.192)	(0.794)	(0.691)

Notes: P-values are given in the parentheses, ** indicates the significance at the 5% level.

B Appendices for Chapter 3

B.1 Confusion Matrices of Translation Analysis

Overall Foreign-to-English Translation (D1).



(a) DistilmBERT-overall-before

(b) DistilBERT-overall-after

(c) mBERT-overall-before

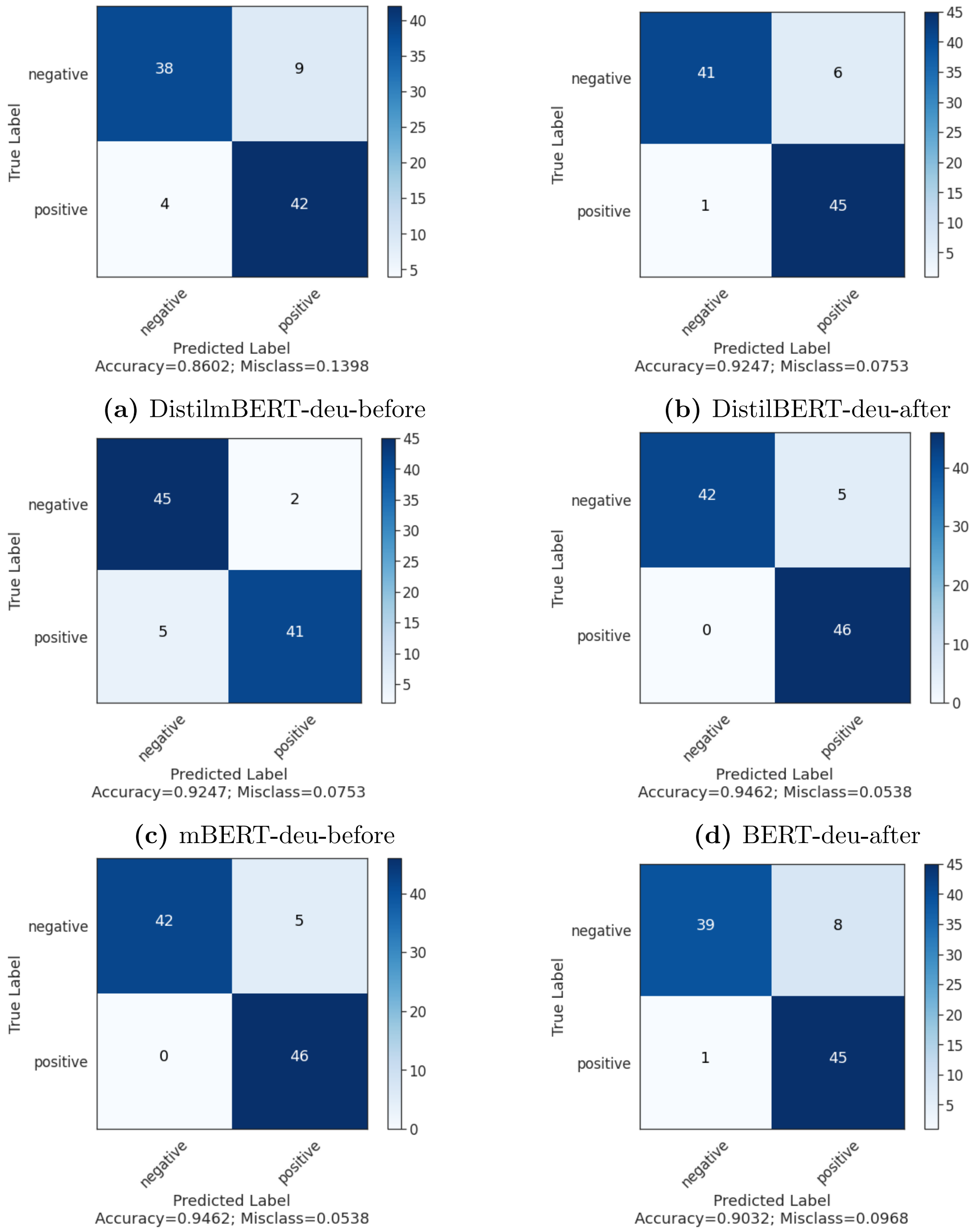
(d) BERT-overall-after

(e) XLM-R-overall-before

(f) XLM-R-overall-after

Notes: The caption in red indicates the change in accuracy after translation is statistically significant.

German-English Translation



(a) DistilmBERT-deu-before

(b) DistilBERT-deu-after

(c) mBERT-deu-before

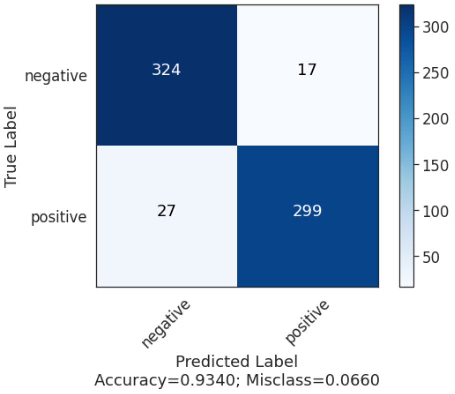
(d) BERT-deu-after

(e) XLM-R-deu-before

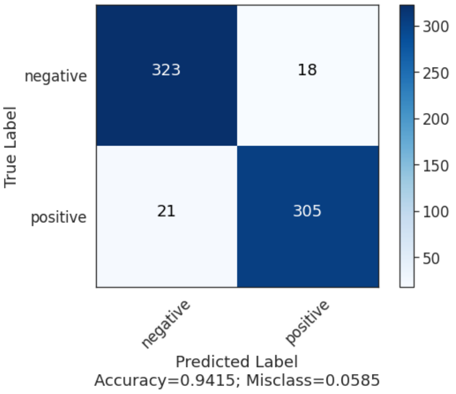
(f) XLM-R-deu-after

Notes: The caption in red indicates the change in accuracy after translation is statistically significant.

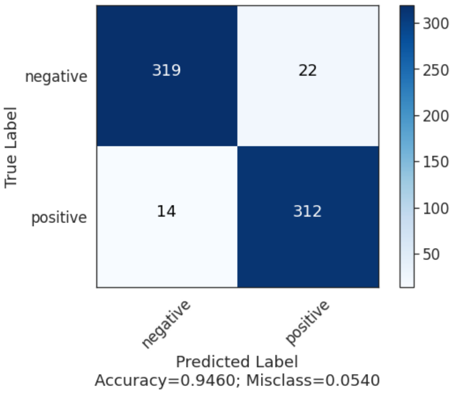
Spanish-English Translation



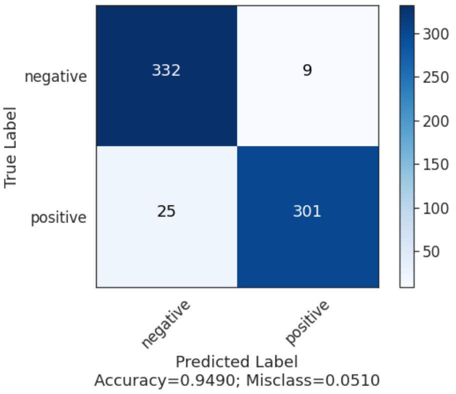
(a) DistilmBERT-spa-before



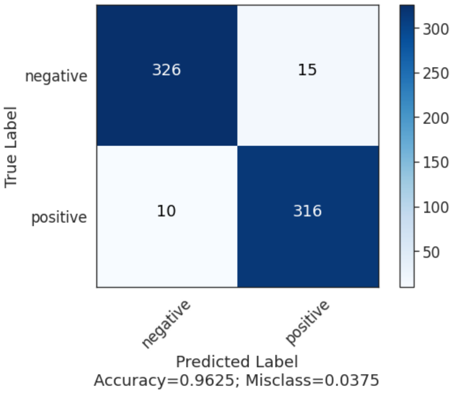
(b) DistilBERT-spa-after



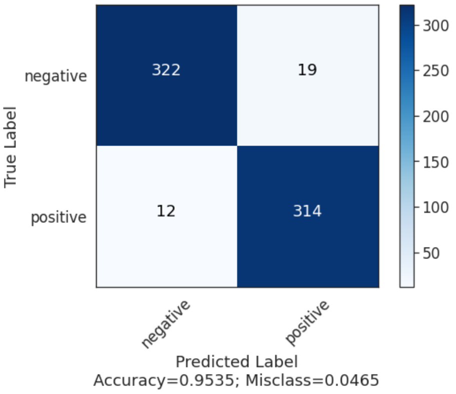
(c) mBERT-spa-before



(d) BERT-spa-after



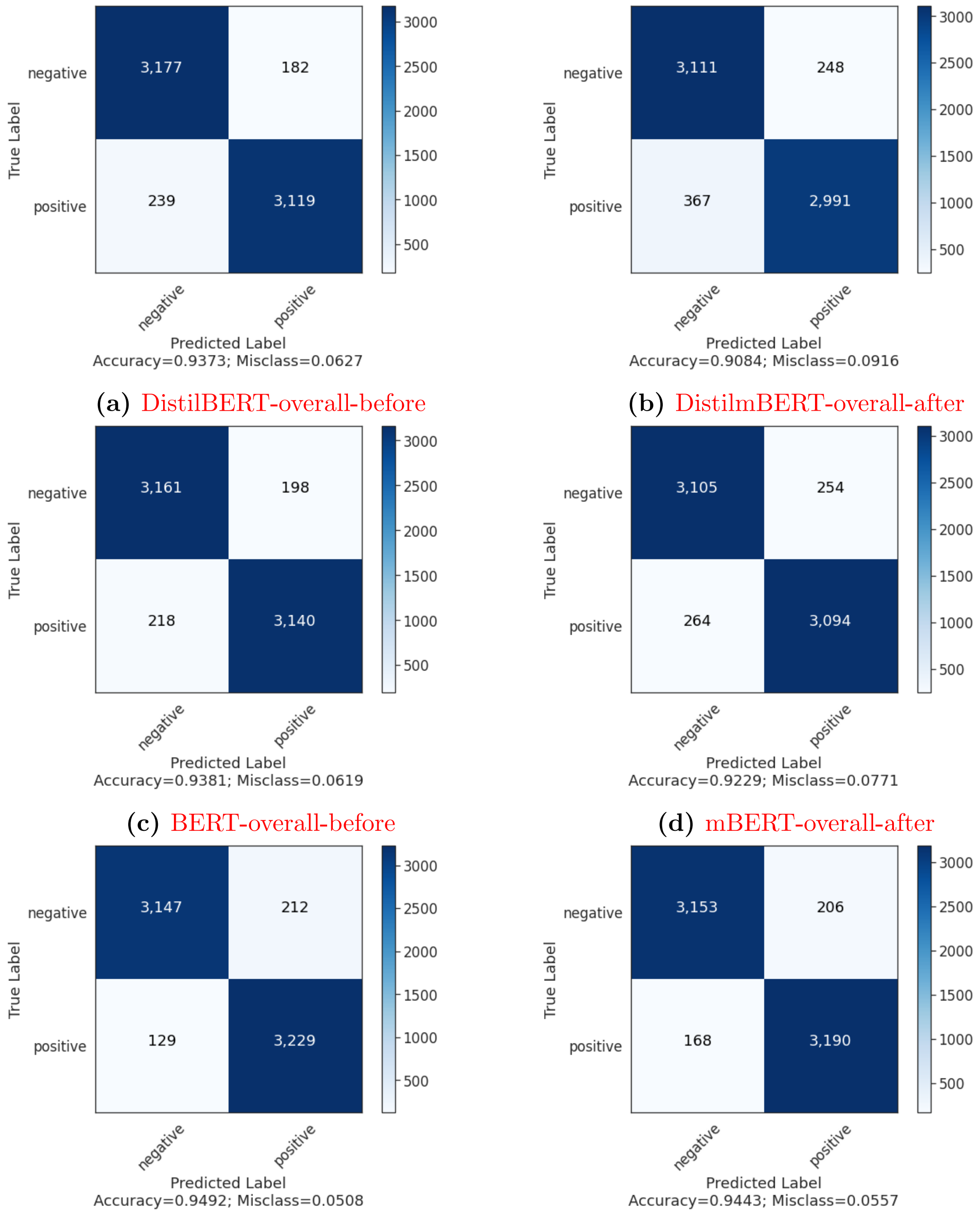
(e) XLM-R-spa-before



(f) XLM-R-spa-after

Notes: The caption in red indicates the change in accuracy after translation is statistically significant.

Overall English-to-Foreign Translation (D2).



(a) DistilBERT-overall-before

(b) DistilmBERT-overall-after

(c) BERT-overall-before

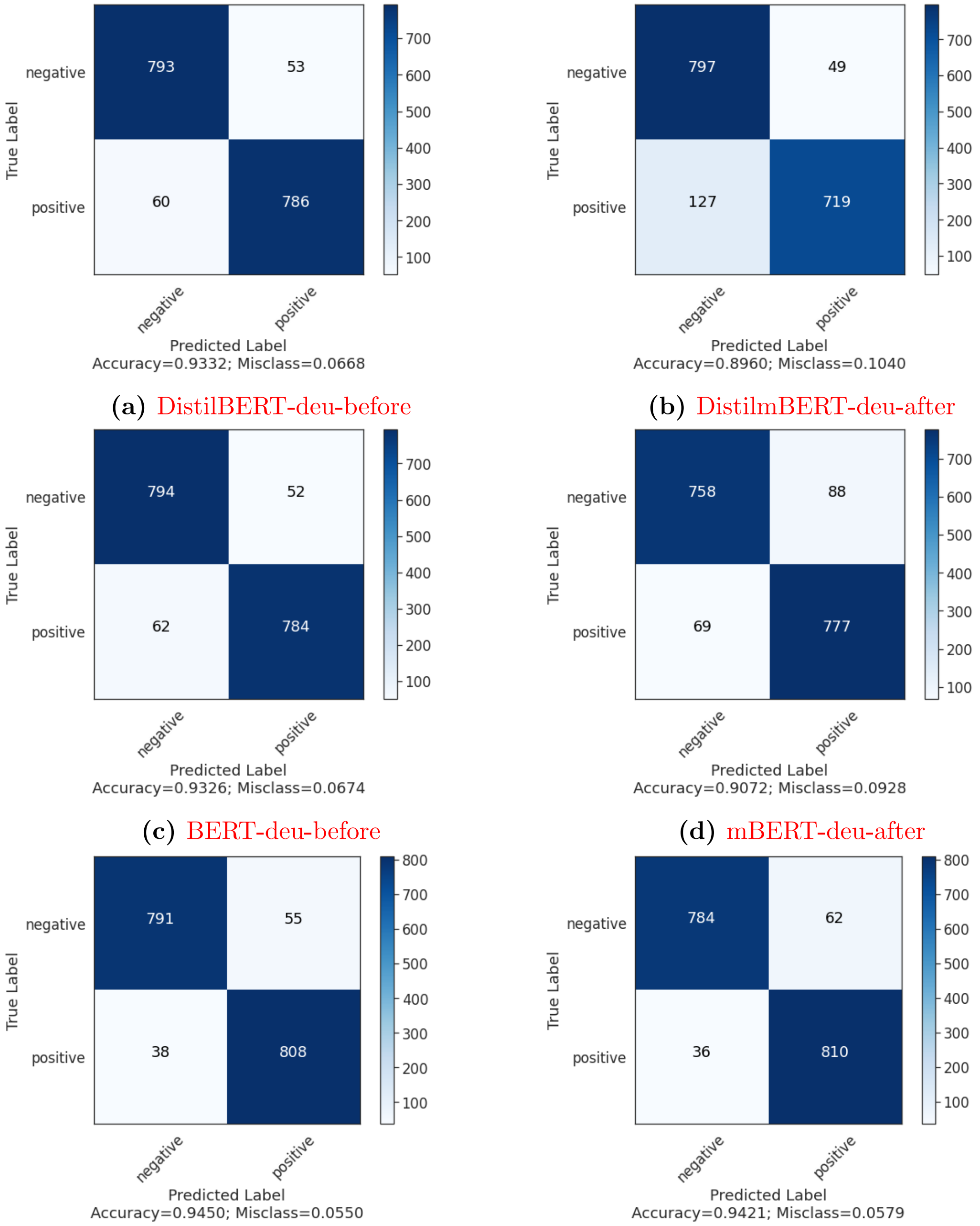
(d) mBERT-overall-after

(e) XLM-R-overall-before

(f) XLM-R-overall-after

Notes: The caption in red indicates the change in accuracy after translation is statistically significant.

English-German Translation



(a) DistilBERT-deu-before

(b) DistilmBERT-deu-after

(c) BERT-deu-before

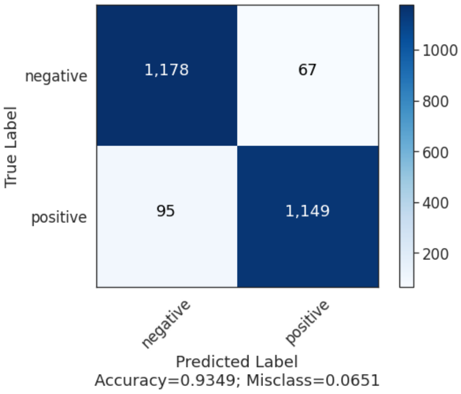
(d) mBERT-deu-after

(e) XLM-R-deu-before

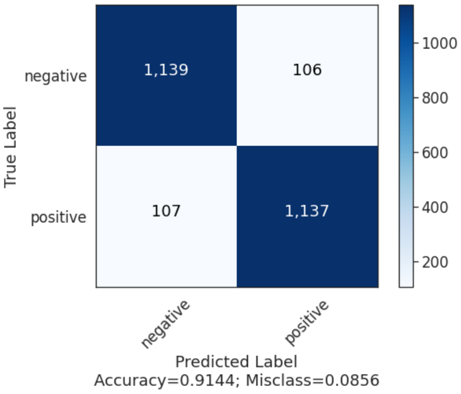
(f) XLM-R-deu-after

Notes: The caption in red indicates the change in accuracy after translation is statistically significant.

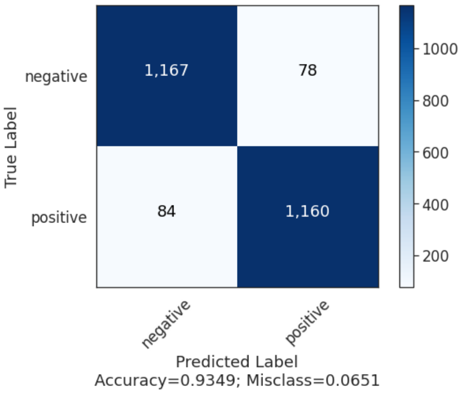
English-French Translation



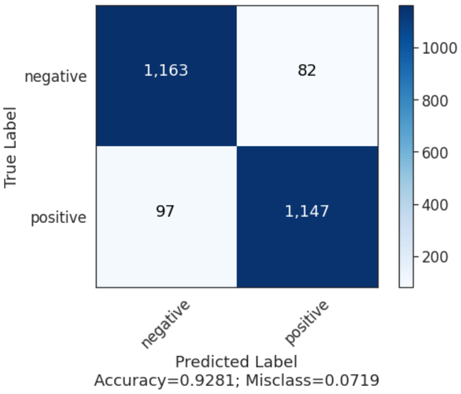
(a) DistilBERT-fra-before



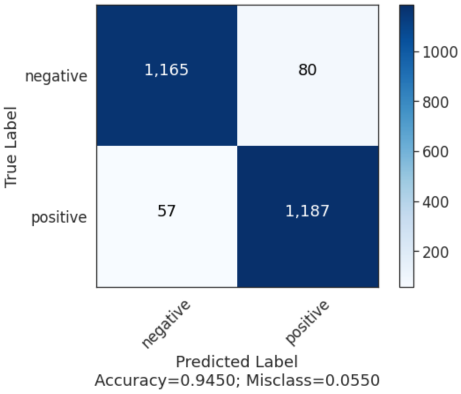
(b) DistilmBERT-fra-after



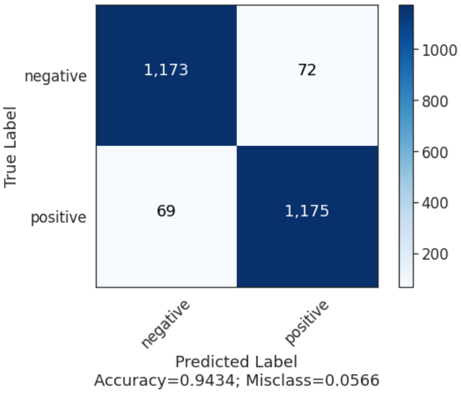
(c) BERT-fra-before



(d) mBERT-fra-after



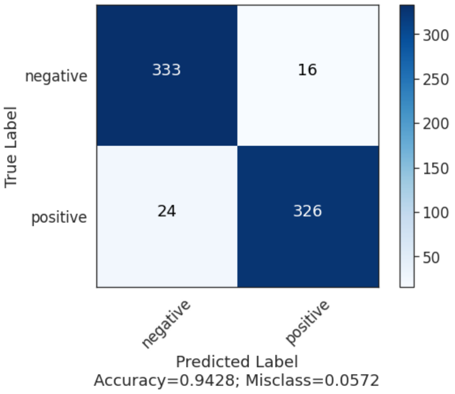
(e) XLM-R-fra-before



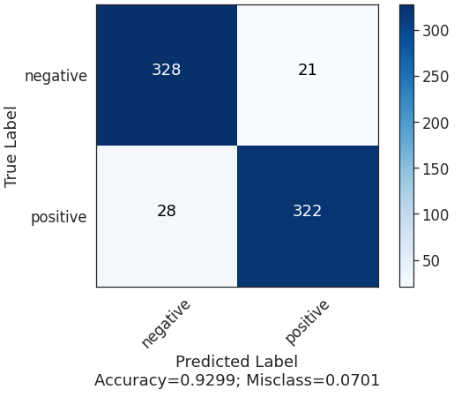
(f) XLM-R-fra-after

Notes: The caption in red indicates the change in accuracy after translation is statistically significant.

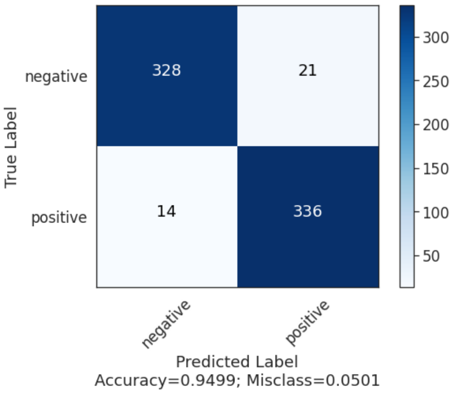
English-Portuguese Translation



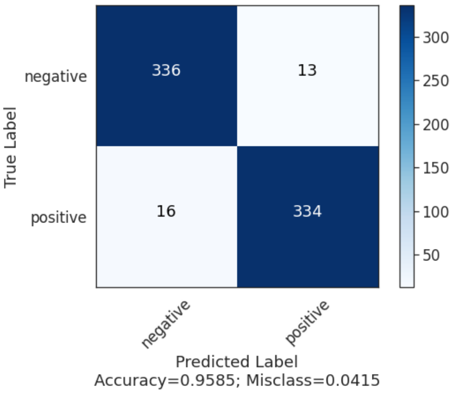
(a) DistilBERT-por-before



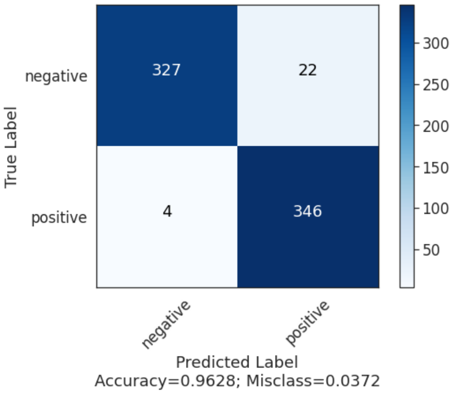
(b) DistilmBERT-por-after



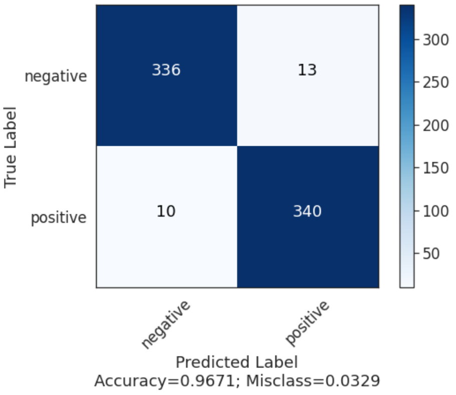
(c) BERT-por-before



(d) mBERT-por-after



(e) XLM-R-por-before



(f) XLM-R-por-after

Notes: The caption in red indicates the change in accuracy after translation is statistically significant.

Chapter 4

BERT Employee Sentiment and Stock Returns

4.1 Introduction

With the development of the Internet and the convenience of online platforms, it has become increasingly easy for everyone to exchange information through social media. This has led to a proliferation of employee reviews on social media sites and dedicated platforms such as Glassdoor. Employees can easily write and post reviews of their companies, providing valuable insights into their workplace experiences and overall satisfaction levels. In turn, companies and researchers can leverage these reviews to gain a better understanding of employee sentiment and corporate culture. In this paper, we aim to investigate the relationship between employee satisfaction and stock returns using 1,352,736 Glassdoor employee reviews for 617 large US companies listed in the S&P500 from 2008 to 2021.

Glassdoor offers a platform for email-verified employees to anonymously leave a review for their company. Each review consists of an overall rating and five sub-ratings: Work-Life Balance, Culture and Values, Career Opportunities, Compensation and Benefits, and Senior Leadership. These ratings are on a scale of 1 to 5 stars. Accompanying the ratings, reviewers can freely express their opinions in text boxes for pros, cons, and advice to management (minimum characters for pros and cons, advice is optional). Reviewers are given the option to select statements about their approval of the CEO, the company's business outlook, and whether they would recommend the company to a friend. They can also provide information about their employment length and job status (current or former, full-time or part-time).

Numerous studies demonstrate the importance of prioritising the numerical star rating as a measure of employee satisfaction and sentiment. These studies indicate that companies with satisfied employees are more likely to be productive, motivated, and committed to the company, leading to improved financial performance and higher stock returns (Edmans 2011, Melián-González et al. 2015, Huang et al. 2015, Symitsi et al. 2018, Stamolampros et al. 2019, Corritore et al. 2020, Green et al. 2019). In their research, Wolter et al. (2019) investigate the impact of systematic changes in employee satisfaction on customer outcomes. They reveal that employee satisfaction trajectories significantly influence customer satisfaction and repatronage intentions, particularly for companies where there are substantial employee-customer interactions. Using employee predictions of companies' six-month business outlook from Glassdoor, Huang et al. (2020) find that the average employee outlook can predict future operating performance. This effect is particularly pronounced for firms that receive less attention from analysts and investors. They also find the predictability is greater when the disclosures are aggregated from a larger and more diverse employee base.

Nonetheless, topic modelling has also been a popular textual analysis tool for Glassdoor employee reviews. Using these reviews, Symitsi et al. (2021) show that integrating unsupervised textual techniques into standard data analysis and models reveals hidden factors influencing key operational and financial indicators, such as job satisfaction, employee turnover, and financial performance. Other studies apply different textual analysis methods to extract information from Glassdoor employee opinions. For instance, dictionary-based text analysis programs such as DICTION, Linguistic Inquiry and Word Count (LIWC), and WordNet have been employed

(Stamolampros et al. 2019, Corritore et al. 2020). Data-mining software such as IBM Watson has also been used (Dabirian et al. 2017, 2019). Furthermore, Tambe et al. (2020) adopt cluster analysis of text from Glassdoor employee reviews and find that information technology (IT) workers tend to prefer companies that use emerging technologies because they highly value technology and learning opportunities on the job. In another study, Campbell & Shang (2022) use an inverse regression approach to assign importance weights to words and demonstrate that employee comments frequently mention attributes associated with corporate misconduct.

Previous analyses on employee sentiment, however, have mainly focused on Glassdoor ratings, which may ignore the valuable information contained in the text comments of the reviews. To bridge this gap, our work is the first to assess employee sentiment directly from the text comments, and the first to apply the NLP methods- Encoder Representations from Transformers (BERT) (Devlin et al. 2018) and the Loughran-McDonald (LM) dictionary (Loughran & McDonald 2011) for sentiment analysis on Glassdoor reviews. BERT is a state-of-the-art natural language processing model that has shown superior accuracy compared to traditional dictionary-based methods (Sousa et al. 2019, González-Carvajal & Garrido-Merchán 2020, Mishev et al. 2020, Zhao et al. 2021, Stevenson et al. 2021, Zhu et al. 2022). Meanwhile, the LM dictionary remains one of the most popular and widely used methods in finance for sentiment analysis for its simplicity and ease of use. The LM dictionary has been used in numerous studies to measure the sentiment of financial news and reports, stock market discussions, and social media posts (Rogers et al. 2011, Engelberg et al. 2012, Huang et al. 2014, Tsai & Wang 2017, Feuerriegel & Gordon 2019, Mishev et al. 2020, Frankel et al. 2022, Huang et al. 2023).

We also use the numerical star ratings as a measure of employee sentiment in addition to BERT and LM, such that we can investigate whether different sentiment measures, at varying levels of accuracy and complexity, can contribute differently to the empirical results. We first explore the determinants of employee sentiment using logistic regressions. We show that a similar set of factors (e.g. Overall, Work-

Life Balance, Culture & Values and Senior Leadership Ratings etc.) tend to drive employee sentiment in BERT, LM and the overall rating. However, the effect sizes of BERT are larger than LM, suggesting that, as a more advanced model, BERT is better suited to capture the employee sentiment in Glassdoor reviews compared to LM and compared to the star ratings.

Second, we report that the analysis of employee sentiment variation and stock returns yields different indications when using BERT, LM and the ratings in the value-weighted portfolios, the high sentiment portfolio sorted by BERT archives a positive and significant alpha of 0.12%, however, according to LM and the ratings the low and mid sentiment portfolios have better performance. We also report the value-added effect of BERT in comparison to LM and rating by sorting the portfolios by the difference of these sentiment measures.

In addition, we examine employee sentiment and stock returns across different industries. We show that BERT, LM and the ratings produce comparable results for most equal-weighted portfolios, and companies with high employee sentiment in the Consumer Staples, Health Care, and IT industries outperform companies with low employee sentiment in the same industry by 0.56% to 1.04%. However, this evidence only holds for BERT in the value-weighted portfolios.

Third, we apply the Fama-MacBeth two-step regression approach to estimate the impact of employee sentiment, firm characteristics and topics on monthly stock excess returns over various horizons. The results suggest that the positive employee sentiment measured by BERT significantly increases the 1- and 3-month ahead stock returns. However, the sentiment estimated by LM or the ratings does not show any significant evidence to predict stock returns in either the short or long term. We show evidence that the textual data can be more informative and interpretable than star ratings for sentiment analysis as the star ratings often oversimplify the sentiment expressed in a review.

Moreover, we build upon the work of Hales et al. (2018), who use selling, general and administrative (SG&A) expenses as a measure of wages and benefits paid to employees, and find a positive association between SG&A expenses and employment outlook. They also report that employees tend to express a more negative outlook towards restructuring charges, such as plant closures or layoffs. In our study, we extend their analysis by examining the industry-specific effects of employee-related costs on both employee satisfaction and stock returns. We reveal that industries that have lower costs in production and manufacture, and higher costs in operation and management, exhibit a clear trend where staff are more important, and employee morale is also more important for performance.

The structure of this paper is as follows. After this introduction, we will provide a detailed description of the data used in our analysis in Section 4.2, which includes summary statistics of employee review data and firm-level characteristics. In Section 4.3, we will present our textual analysis, which focuses on sentiment analysis and topic modelling. Section 4.4 will discuss the results of our analysis, covering the determinants of employee sentiment, portfolio sorting by employee sentiment, and stock returns predictability through Fama-MacBeth regressions. We also study the relationship between employee-related costs and stock returns across industries in Section 4.4. Finally, we will conclude our paper in Section 4.5, where we will summarise our findings and discuss possible future research directions.

4.2 Data

4.2.1 Employee Review Data

The data used in this study includes a collection of Glassdoor employee reviews for constituents of the S&P 500 index from January 2008 to March 2021. Glassdoor is a website that provides an online platform for employees to anonymously share their experiences, opinions, and insights about their current or former employers. These reviews are shared in the form of ratings and written comments covering various aspects of the job, such as the overall company culture, Work-life Balance, Senior

Leadership, Compensation & Benefits, and Career Opportunities. The ratings are on a scale of 1 to 5, where 5 is the highest rating and 1 is the lowest. In addition to the numerical ratings, employees have the option to provide both positive and negative comments about the company, highlighting both the pros and cons of working there and any advice to management. This comment section offers an open-ended opportunity for the employee to elaborate on their experience and express their views on specific aspects of the company. To extract meaningful insights from the comments, we apply sentiment analysis and topic modelling, two techniques that have been demonstrated to be effective in revealing underlying emotions and themes in text data. We will discuss the details of these methods and the results they yield in the next section.

Reviewers are also asked to provide their perspectives on certain elements related to the company's reputation and future prospects, and we construct dummy variables to quantify them. One of these elements is their willingness to recommend the company to a friend, which is represented by 1 if yes, and 0 otherwise. Another important aspect is the CEO approval rating, given the option of "APPROVE", "NO OPINION", or "DISAPPROVE", we assign the values of 1, 0, and -1 correspondingly. Lastly, the business outlook rating provides an indication of the employee's perception of the company's future prospects and potential for growth. It is measured using a scale of "POSITIVE", "NEUTRAL", or "NEGATIVE", represented by values of 1, 0, and -1, respectively. These additional ratings provide a more comprehensive understanding of the reviewer's experience and their view of the company.

The dataset we compile contains not only employee reviews of companies but also characteristics of the reviewers, including the length of the reviewer's employment, which is represented in ranges of 0, 1, 5, 8, 10 or more years. We also categorise the reviewer's employment type, with options for full-time or part-time work, coded as 1 and 0, respectively. Finally, we include the job type which identifies whether the reviewer is a current or former employee, and is coded as 1 and 0, respectively. To ensure that the reviews for each company are representative, a criterion is set such that each company has to have at least 30 employee reviews per month. Our final employee dataset contains 1,352,736 employee reviews for 617 unique companies.

4.2.2 Summary Statistics

To retrieve the relevant firm-level characteristics, we manually match company names to PERMNO identifiers in the Center for Research in Securities Prices (CRSP) and Compustat databases. A full description of the variables is provided in Table C16 in the appendix, and their summary statistics are reported in Table 4.1, with Panel A for employee reviews and Panel B for monthly firm characteristics.

[INSERT Table 4.1 ABOUT HERE]

On average, employees are satisfied with their company, as indicated by an overall rating of 3.48 in Panel A. However, the average sub-ratings are all lower than the overall rating, including Work-Life Balance (2.87), Culture & Values (2.72), Senior Leadership (2.55), Career Opportunities (2.81), and Compensation & Benefits (2.95). These lower sub-ratings indicate that employees may have some concerns regarding these areas. However, the ratings should be evaluated in the context of company or industry standards, as well as employee expectations and requirements. The average employee's sentiment towards their company's CEO and business outlook is slightly positive, and employees are generally willing to recommend the company to their friends. Moreover, the average employment duration is 3.22 years, 57% of the reviewers are current employees, and 85% of them work full-time.

Given that the focus of our examination comprises the S&P 500 constituents, it is worth noting that these firms are highly liquid large-cap entities. Therefore, as indicated in Panel B of Table 4.1, the average market capitalisation of our reviewed firm is around $35 million and their level of illiquidity as measured by Amihud (2002) is exceedingly low. We scale the trading volume by the shares outstanding to calculate the turnover and obtain an average turnover of 2.41 for the firms that we have reviewed. The average Book-to-Market (BM), Return on Assets (ROA), and Return on Equity (ROE) ratios are 0.52, 0.14, and 0.18, respectively. The average age of the companies is 28.69 years suggesting the firms are well-established with a track record of longevity. Finally, the mean excess return is 0.84, showing that the companies had generated a positive return over and above the market benchmark during our period of analysis.

Furthermore, we extend the work of Hales et al. (2018) to explore several employee-related costs which may influence employee satisfaction and stock returns including Cost of Goods Sold (COGS), Selling, General and Administrative (SG&A) Expenses, Research and Development (R&D) Expenses, Stock Compensation (STKCO) Expenses, and Restructuring Costs (RC). COGS, for example, refers to the direct costs associated with producing goods, including materials, labour, and manufacturing overheads. Employee satisfaction may be affected if the company is not paying its employees a fair wage or is cutting corners to reduce COGS. Similarly, SG&A expenses cover the indirect costs associated with running the company's day-to-day operations such as salaries, bonuses, and benefits for non-production employees. If SG&A expenses are reduced to increase profits, employees may feel undervalued and less satisfied with their jobs.

On the other hand, R&D expenses can indirectly impact employee satisfaction. Investing in R&D can lead to the development of new products, which can create new job opportunities and enhance employee engagement. Moreover, companies that invest in R&D demonstrate a commitment to innovation and progress, which can boost employee morale and job satisfaction. STKCO reflects the expenses associated with the issuance of stock options or other equity-based compensation to employees. Stock compensation plans can motivate employees to work harder and align their interests with those of the company. However, employees may become demotivated if they perceive that the company is not fairly compensating them through stock options. Finally, RC represents the expenses incurred in restructuring activities such as downsizing, closure of facilities, and implementing new business strategies.

High restructuring costs may result in job losses and lower job security, leading to reduced employee satisfaction. The statistical summary of these employee-related costs is presented in Table 4.1 Panel B.

4.3 Textual Analysis

In this study, we aim to gain deeper insights into employee satisfaction and its impact on company performance by analysing the textual comments from Glassdoor employee reviews. However, extracting meaningful information from the vast amounts of text data can be a challenge, therefore, we adopt two popular text analysis techniques - sentiment analysis and topic modelling - to analyse the comments of pros, cons and advice to management.

4.3.1 Sentiment Analysis

Sentiment analysis is the process of identifying and extracting emotions from text, which is useful for understanding employees' opinions and attitudes towards their company. There are many methods available for sentiment analysis, each varying in their level of complexity. Simpler methods, such as the dictionary-based methods involve using pre-built lists of words associated with positive, negative, or neutral sentiments, it is easy to use and computationally efficient. Whereas the more advanced deep learning model, such as the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018), excels in its ability to capture the context and meaning of words in a sentence. In Chapter 2, we compare the effectiveness of 31 sentiment analysis methods using text comments from 20,000 Glassdoor employee reviews. We show that BERT exhibits an accuracy one-third higher than that of the dictionary-based methods when measured against employee star ratings. However, despite this superior performance, dictionary-based methods continue to be the preferred choice in the finance industry. In this study, we apply both BERT[1] and the Loughran-McDonald (LM) dictionary (Loughran & McDonald 2011) to extract the sentiment from employee comments.

BERT is a pre-trained language model that can be fine-tuned for sentiment analysis, we first tokenise the text using the BERT tokeniser, which breaks down the text into individual words or subwords that BERT can understand. We then convert each tokenised review into a numerical format that can be fed into BERT, called an input representation. The overall rating of each review is used as the label for training, and the data is split into training, validation, and test sets. During fine-tuning, BERT learns how to associate the input representations with the corresponding labels. Once the model is trained, we can use it to predict the sentiment of new employee reviews by tokenising each review, converting it to an input representation, and passing it through the fine-tuned BERT model. The output of the model is a probability distribution over a positive and a negative class, and the sentiment class with the highest probability is considered the predicted sentiment for the review.

In addition to BERT, we also incorporate the LM dictionary, which contains over 2,300 words and phrases that are commonly used in financial news and reports. Each word and phrase in the dictionary is assigned a score indicating whether it is positive or negative in sentiment. For example, the word "profit" is assigned a positive score, while the word "lawsuit" is assigned a negative score. Since it does not require additional training, we first pre-process each review using standard NLP techniques such as word tokenisation, stopword removal and stemming. Then, we count the number of positive and negative words in the review that are present in the LM dictionary. We use these counts to compute a sentiment score for each review, where a higher score indicates a more positive sentiment. We take the overall rating provided by the employees in their reviews as the ground truth and categorise the reviews as positive if the overall rating is 3 stars or above, and as negative otherwise.

[1] The pre-trained model is downloaded from `https://huggingface.co/bert-base-cased`

From the sentiment predictions, we find BERT model achieves an accuracy of 90.2%, compared to 62.6% for the LM dictionary. In Table 4.2, we provide a set of reviews and their sentiment predictions produced by LM and BERT.

[INSERT Table 4.2 ABOUT HERE]

The LM dictionary has the advantage of its specificity to financial language, which makes it more accurate for financial sentiment analysis than more general dictionaries. However, it has some limitations which can distort the accuracy of the analysis. One limitation is that it relies on a fixed set of positive and negative words to determine the sentiment of a sentence, and as a result, it may fail to capture negation. Negation occurs when a word or phrase indicates the opposite of what is being said. For example, in the sentence “Leadership do not care about training or assisting you to help you do better.” the LM dictionary would classify it as positive because it only detects the presence of the words “leadership” and “better”, even though the overall sentiment expressed is negative. BERT, on the other hand, would recognise the context these words are in and correctly classify this as a negative sentiment. It is also difficult for the LM dictionary to detect sarcasm and irony in text because the sentiment expressed is often the opposite of what the words actually say. BERT can identify the sentiment based on the overall context of the text, allowing it to detect sarcasm and irony more accurately.

There are several reasons why we choose the LM and BERT models over other dictionaries or FinBERT. First, we recognise Glassdoor employee reviews cover a wide range of contexts, including both financial and general aspects. Employees not only share their own experiences about the work environment, company culture, and management but also discuss salary, compensation, bonuses and other financial incentives offered by the company. Compared to other dictionaries, the LM dictionary offers a domain-specific focus, which better captures the financial aspects of these reviews. Second, from our previous experiments in Chapter 2, we observe that BERT outperforms FinBERT in the majority of cases. While FinBERT could be considered as an alternative choice, BERT is pre-trained on a larger and more diverse dataset, making it more flexible and adaptable to various contexts where financial sentiment is embedded within a broader set of employee experiences and perspectives.

In addition to sentiment extraction from the text comments of employee reviews, we also measure the review's sentiment from its overall star rating. Star ratings provide a straightforward and easily understandable way to assess sentiment. They offer a quick summary of overall satisfaction or dissatisfaction without the need for extensive analysis or interpretation. If a review has 3 stars or above, we consider it to be positive otherwise negative. Though the majority of prior studies focus on the star ratings, they may lack the depth and context provided by text comments. It is in our interest to compare empirical outcomes produced by these different sentiment measures.

Nest, we follow Brown & Cliff (2004, 2005) and Chen et al. (2023)'s work and construct a monthly Employee Sentiment Index (ES) to track the employee sentiment changes over time. The ES for the company c in month i is computed by subtracting the number of negative reviews from the number of positive reviews and dividing the result by the total number of reviews.

$$ES_{c,i} = \frac{Num\,of\;Pos\;Reviews_{c,i} - Num\,of\;Neg\;Reviews_{c,i}}{Total\;Reviews_{c,i}} \tag{4.1}$$

We form 3 independent ES indices with the sentiment predictions of BERT, LM and the overall star ratings for each company. They are denoted as ES_BERT, ES_LM, and LM_Rating, respectively. The value of the index ranges from -1 to 1, where a higher value indicates a more positive employee sentiment during that month.

4.3.2 Topic Modelling

Topic modelling is an unsupervised learning technique that is commonly used to uncover to identify underlying themes or topics that are present in a large volume of texts without prior knowledge of their content. We follow the work of Schmiedel et al. (2019) and implement Latent Dirichlet Allocation (LDA) to extract meaningful topics from the employee reviews. The purpose of topic modelling is to assist sentiment analysis such that we can investigate the specific aspects of employee opinions that influence their overall sentiment.

The process of topic modelling involves identifying patterns of co-occurrence among words in a corpus and grouping them together into topics. After the reviews are pre-processed and vectorised using similar techniques as the sentiment analysis, we apply the most commonly used approach for topic modelling is Latent Dirichlet Allocation (LDA), which is a probabilistic generative model that assumes that each document in a corpus is a mixture of topics, and each topic is a distribution over words. The goal of LDA is to estimate the distribution of topics in each document and the distribution of words in each topic. We utilise the topic coherence score which evaluates the degree to which the top words in each topic are semantically related to each other. Through experimentation, we find that a model with 13 topics provided the best balance between semantic coherence and model complexity and then we manually label each topic based on the top words associated with it.

Table 4.3 reports the extracted topics and 20 keywords in each topic. In the regression analysis, we interact the topics with BERT and LM sentiment predictions individually. This enables us to examine the sentiment expressed within each of the 13 identified topics and evaluate the areas where employee sentiment was particularly strong or weak. As indicated in Table 4.4, our analysis using both BERT and LM sentiment analysis techniques shows that the overall employee reviews on Glassdoor tend to be more positive than negative across most topics. However, the difference between the proportion of positive and negative reviews predicted by BERT is generally higher compared to LM. For instance, in the "Work-Life Balance" topic, BERT predicts that 84.5% of reviews are positive and 15.5% are negative, while LM predicts that 58.8% are positive and 41.2% are negative. These findings suggest that BERT may be more sensitive to detecting positive sentiment in employee reviews, while LM may be more pessimistic in its predictions. We summarise the accuracy of LM and BERT at both the topic and industry levels in Table C17 in the appendix.

[INSERT Table 4.3 ABOUT HERE]

[INSERT Table 4.4 ABOUT HERE]

Moreover, we also compute the topic distribution across 11 Global Industry Classification Standard (GICS) industries. The summary in Table 4.5 reports that Compensation & Benefits, Career Opportunity, and Job Security are the three most frequently mentioned topics. However, there are also notable differences in topic distribution among industries. For instance, in the Consumer Discretionary, Consumer Staples, Communication Services, and Financials industries, "Customer Service" is a highly mentioned topic. In contrast, the IT industry had a greater focus on Work-Life Balance.

[INSERT Table 4.5 ABOUT HERE]

4.4 Result Discussion

4.4.1 Determinant of employee sentiment

We first examine the determinants of employee sentiment through the logistic regressions, as shown in Table 4.6 we include both BERT (Columns 1-3), LM (Columns 4-6) and the overall star ratings (Columns 7-9) as the sentiment measurements. We construct ES_BERT, ES_LM and ES_Rating to observe the sentiment changes over time but since they are continuous values, we transform them into categorical variables before the logistic regression analysis. If the ES is equal to or greater than 0, we consider the employee sentiment to be positive for the month, otherwise negative. However, it is worth noting that converting a continuous variable into a binary one may entail a loss of information. By framing logistic regression as a classification task, we aim to provide a clearer understanding of the determinants of sentiment direction, rather than focusing on the exact sentiment magnitude. Therefore, the choice of a binary classification approach aligns with our research objectives and allows for a more straightforward interpretation of the results. Specifically, the logistic regression is as follows:

$$P(Y = 1|X) = 1/(1 + exp(-z)) \tag{4.2}$$

Where $P(Y = 1|X)$ is the probability of positive employee sentiment (1) given the independent variables X. z is the linear combination of the independent variables, which is calculated as:

$$z = \beta_0 + \beta_1 X_{i,j} + \beta_2 Y_{i,j} + \beta_3 Z_{i,j} \tag{4.3}$$

$X_{i,j}$ are the ratings of the reviews, $Y_{i,j}$ are reviewer characteristics (e.g. employment length, current vs. former employee, full-time vs. part-time employee), and $Z_{i,j}$ are firm-level variables related to employee costs.

The correlation matrix of the variables is reported in Table C18 in the appendix. Since ES_Rating is transformed from the overall ratings and they are highly correlated, we omit the Overall Rating in the last set of regressions. The results indicate that while similar factors tend to drive employee sentiment in BERT, LM and the ratings, their effect sizes may differ. ES_BERT and ES_Rating share closer coefficients in terms of sub-ratings and reviewer characteristics than ES_LM. We convert the coefficients from Table 4.6 to odds ratios by e^{β} for simpler interpretation. For example, one unit increase in the Overall and Culture & Values ratings is associated with 57% [2] and 31% increases, respectively in the odds of positive ES_BERT. Whereas one unit increase in the same ratings is associated with 18% and 14% increases, respectively in the odds of positive ES_LM.

[INSERT Table 4.6 ABOUT HERE]

In addition, when firm-level characteristics are not controlled for, every unit increase in Senior Leadership Rating, the odds of positive employee sentiment decrease by 12% for ES_BERT and ES_Rating, respectively, and 7% for ES_LM. This shows that employees' perceptions of senior leadership play a slightly negative role in shaping employee sentiment according to BERT and the ratings. The Career Opportunities rating increases the probability of positive ES_LM, while it is not significantly related to ES_BERT and ES_Rating. On the other hand, the Recommend to Friend rating increases the odds of positive ES_BERT and ES_Rating, but it is not significantly related to ES_LM. Similar to the findings of Hales et al. (2018), we discover higher COGS and SG&A are associated with a higher likelihood of positive ES_BERT, ES_LM and ES_Rating, STKCO shows a significant positive relationship with ES_BERT and ES_Rating but not with ES_LM. Higher COGS, SG&A and STKCO could indicate greater financial stability and profitability for the company, which may in turn lead to more positive employee sentiment. Besides, companies that invest in their employees and offer competitive salaries and benefits may have higher COGS and SG&A, hence having more positive employee sentiment.

4.4.2 Portfolio sorted by employee sentiment

To explore the relationship between employee sentiment and stock returns, we sort the S&P 500 constituents into tercile portfolios based on their Glassdoor employee sentiment index, as measured by ES_BERT, ES_LM and ES_Rating. This process

[2] Calculated as $e^{\beta} - 1 = e^{0.45} - 1 = 0.57$. The following figures are derived based on the same formula.

is carried out for each month from January 2008 to March 2021. We first rank the firms based on their average employee sentiment index scores and divide them into three groups of equal-sized portfolios based on their percentile rankings. The top third firms with the highest ES are sorted into the High sentiment portfolio, the bottom third firms with the lowest ES are sorted into the Low sentiment portfolio, and the remaining firms are sorted into the neutral sentiment portfolio (i.e. the Mid group). We also construct a long-short portfolio using employee sentiment where the investor takes a long position in the high ES portfolio and a short position in the low ES portfolio (i.e. the High-Low group). The stocks in the portfolios are rebalanced on a monthly basis, and both equal-weighted and value-weighted portfolios are constructed.

Table 4.7 Panel A presents the average excess returns for different portfolios constructed based on the employee sentiment index as measured by ES_BERT, ES_LM and ES_Rating. The standard errors are calculated using Newey & West (1987) to account for heteroscedasticity and autocorrelation. In value-weighted portfolios, all portfolios have positive excess returns, however, the largest average excess return goes to different portfolios among these three groups. For ES_BERT, ES_LM and ES_Rating, the highest return is generated by the High portfolio (1.56%), the Low portfolio (1.59%), and the Mid portfolio (1.45%), respectively. For the long-short portfolios, only ES_BERT generates a positive and statistically significant spread of 0.25% (t-statistic of 2.53). In addition, the overall market is performing well, as shown by the positive and statistically significant market excess returns of 1.36%, 1.37% and 1.35% for ES_BERT, ES_LM and ES_Rating, respectively. For equal-weighted portfolios, both ES_BERT and ES_LM have the highest returns in the Low sentiment group which leads to a negative High-Low portfolio spread, but the differences are not statistically significant.

[INSERT Table 4.7 ABOUT HERE]

To confirm that the performance of portfolios sorted by employee sentiment is not influenced by risk factors, we also run the Fama-French-Carhart (FFC) four-factor model:

$$R_{i,t} = \alpha + \beta_{\mathrm{MKT}}\mathrm{MKT}_t + \beta_{\mathrm{HML}}\mathrm{HML}_t + \beta_{\mathrm{SMB}}\mathrm{SMB}_t + \beta_{\mathrm{MOM}}\mathrm{MOM}_t + \varepsilon_{i,t} \tag{4.4}$$

Where $R_{i,t}$ represents the return on the individual portfolio in month t, MKT_t is the market return, SMB stands for the Small Minus Big factor, HML represents the High Minus Low factor, and MOM is the momentum factor.

The results are presented in Table 4.7 Panel B. For both value- and equal-weighted portfolios, ES_BERT produces positive and statistically significant alphas of 0.07% to 0.12%, respectively, in High portfolios. However, neither of the alphas for the long-short portfolios is statistically significant, suggesting that companies with high or low employee sentiment as measured by BERT are not significantly different in terms of their impact on returns. On the other hand, according to ES_LM, stocks with lower employee sentiment achieve significant alphas of 0.12% to 0.21%, and the value-weighted long-short portfolios that buy stocks in the High portfolio and sell stocks in the Low portfolio could lead to a loss of 0.28% in alpha. Nonetheless, the portfolio sorted by ES_Rating shares a similar conclusion as ES_BERT in the value-weighted setting and find High sentiment group outperform the market.

4.4.3 Value-added of BERT in portfolio sorting

In the last section, we discuss the portfolios sorted by different but independent sentiment measures. However, it would be useful to compare the performance of the portfolios sorted by the differentiation of two sentiment measures. This way, we can examine whether the more advanced sentiment measures or the assessment of text comments offer more empirical value. Specifically, we construct three sets of portfolios sorted by the monthly differences between BERT-LM, BERT-Rating, and LM-Rating, respectively. For example, in Table 4.8, the stocks are categorised into tercile portfolios based on their Δ(ES_BERT-ES_LM) every month and rebalanced at the end of each month. The high sentiment portfolio consists of the top third of firms with the highest Δ(ES_BERT-ES_LM), while the low sentiment portfolio includes the bottom third with the lowest Δ(ES_BERT-ES_LM). The difference between the two is sorted into the long-short portfolio, denoted as High-Low. The remaining firms are allocated to the neutral sentiment portfolio, also referred to as the Mid group. Table 4.9 and Table 4.10 are set up in the same manner.

[INSERT Table 4.8 ABOUT HERE]

[INSERT Table 4.9 ABOUT HERE]

[INSERT Table 4.10 ABOUT HERE]

The results from Table 4.8 demonstrate that BERT shows more advantages than LM with high employee sentiment portfolios. This is indicated by the significant alphas of 0.17% and 0.23% in the value-weighted and equal-weighted settings, respectively. These findings illustrate the value-added nature of BERT over LM. Investors who allocate their investments to portfolios classified as high sentiment using BERT are likely to achieve higher returns compared to those who invest in portfolios classified as high sentiment using LM.

Additionally, when comparing sentiment measured by text versus rating, we observe the outperformance of BERT over star ratings in the Mid and Low sentiment groups, as shown in Table 4.9. The value-weighted Mid Δ(ES_BERT-ES_Rating) portfolio demonstrates a 1.92% return and a 0.35% alpha, while the equal-weighted Low Δ(ES_BERT-ES_Rating) achieves a 1.83% return and a 0.34% alpha. Similarly, the Low Δ(ES_LM-ES_Rating) also reports a significant alpha of 0.32% in Table 4.10. These results suggest that BERT and LM provide additional nuanced information about mid-to-low employee sentiment that is not fully captured by the overall rating.

Furthermore, we investigate whether the relationship between employee sentiment and stock returns varies across different industries. We sort the stocks for each of the 11 GICS industries into tercile portfolios based on ES_BERT, ES_LM and ES_Rating and report the average excess returns and the FFC 4-factor alphas for each industry in Table 4.11. When considering ES_BERT, it is observed that the long-short portfolios from the Materials and Real Estate industries underperform the benchmark and produce significant and negative alphas. Whilst portfolios from Consumer Staples, Health Care, and IT yield significant and positive alphas indicating that purchasing stocks with high employee sentiment and selling stocks with low sentiment in these industries can achieve positive alphas of 0.56% to 1.04%. In general, the sentiment indices produce comparable results for most equal-weighted portfolios. For the value-weighted portfolios, ES_BERT reports significant negative alphas of -1.13% and -1.05% for the Materials and Real Estate industries, respectively, and positive alphas of 1.04%, 0.66% and 0.56% for the Consumer Staples, Health Care and IT industries, respectively. However, some of these effects are not significantly captured by ES_LM and ES_Rating.

[INSERT Table 4.11 ABOUT HERE]

4.4.4 Predicting stock returns: Fama-MacBeth regressions

In the previous section, we examine the effectiveness of using employee sentiment as a factor for portfolio sorting. In this section, we explore the impact of employee sentiment, firm characteristics and topics on stock returns. In particular, we use the Fama-MacBeth two-step regression approach to estimate the relationship between these variables and monthly stock excess returns over various horizons (1, 3, 6, 9, and 12-month ahead). The Fama-MacBeth regression is a two-step procedure that estimates the coefficients of a regression model separately for each time period and then takes the average of the coefficients across all time periods. We first run individual cross-sectional regressions for each month to estimate the risk premiums associated with different factors. The cross-sectional regression equation is specified as follows:

$$R_{i,t} = \beta_0 + \beta_1 ES_{i,t} + \beta_2 X_{i,t} + \beta_3 T_{i,t} + \varepsilon_{i,t} \tag{4.5}$$

Where $R_{i,t}$ is the excess return for stock i in month t, $ES_{i,t}$ is the monthly employee sentiment index, $X_{i,t}$ are firm-level characteristics for stock i in month t, and $T_{i,t}$ is the topic-sentiment interaction for stock i in month t.

This outputs a set of firm-specific coefficients for each independent variable at each horizon. In the second step, we take the cross-sectional average of these coefficients to estimate the overall relationship between our independent variables and monthly excess returns. We run three sets of regressions using the prediction of BERT, LM and overall star ratings independently. This allows us to compare the empirical differences between the three sentiment analysis methods. We report the regression results in Table 4.12.

[INSERT Table 4.12 ABOUT HERE]

In Panel A, it is observable that ES_BERT is positive and significant at 5% to 10% level for the univariate (Column 1) and multivariate (Column 2-4) regressions indicating that employee satisfaction predicts 1- and 3-month ahead stock returns. We take into account the interaction between the sentiment index and topic dummy to control for any confounding effects that may arise from varying levels of sentiment across different topics. The topic-specific sentiment tends to impact mid to long-term stock returns. For the 3-month ahead returns, positive sentiment in Customer Service, Work Environment, and Working Hours topics can predict significant increases in stock returns of 1.99%, 2.65%, and 2.42%, respectively. In addition, the positive effect of Work Environment continues to be present in 6- and 12-month ahead returns, with a stronger impact in the long term (3.79% and 4.79%, respectively). When considering the 12-month ahead returns, sentiment related to Organisational Strategy, Leadership, and Job Security play a more significant role in stock returns. As shown in Column 7, positive sentiment in each of these aspects leads to significant positive returns of 3.81%, 3.40% and 3.72%, respectively. In

Panel B, the sentiment estimated by LM does not show any significant evidence to predict stock returns in the short or long term, and thus the topic-specific sentiment was statistically insignificant. One exception is in Column 14, where positive sentiment in Communication contributes to positive 12-month ahead stock returns of 8.24% at the 5% significance level. The insignificance of stock prediction of the overall star ratings is observed in Panel C as well. Using the ES_Rating as a measure of overtime employee sentiment fails to foresee stock price changes in either the short or the long term.

4.4.5 Employee-related costs and stock returns across industries

Following prior research by Hales et al. (2018), who utilise SG&A expenses as an indicator of employee compensation and discover a positive correlation between SG&A expenses and employment prospects. Additionally, Hales et al. (2018) observe that employees generally hold a pessimistic view when it comes to restructuring charges, such as plant shutdowns or workforce reductions. Therefore, in the previous section, we control for employee-related costs such as COGS, SG&A, R&D, STKCO, and RC, on employee satisfaction and stock returns for all firms in all industries. However, the results suggest that the impact of these costs on employee satisfaction and stock returns is relatively weak when analysed across all industries because these costs can vary greatly depending on the industry in question and different industries may have different cost structures and employee expectations. To solve this problem, we conduct a more detailed examination focusing on the industry-specific effects of employee-related costs on employee satisfaction and stock returns.

Specifically, we focus on employee sentiment extracted from text comments instead of numerical star ratings as they offer richer information, contextual understanding, and help uncover root causes of sentiment. Moreover, we develop this analysis using predictions of BERT over LM because it is more accurate and it consistently provides more insightful results in our portfolio sorting and stock return prediction analyses.

We first review the differences in the allocation of employee-related costs across the 11 GICS industries. As reported in Table 4.13, the utilities industry have the highest proportion of COGS (96.60%), this is possibly due to the extensive infrastructure such as power plants, pipelines, and distribution networks required for building and maintaining the generation of electricity, gas and water. The operation of real estate companies is also capital-intensive, expenses such as property purchase costs, property management, maintenance, and property taxes can result in a substantial amount of COGS (83.98%). The energy sector encompasses oil, natural gas, and renewable energy sources, all of which involve resource extraction and processing. Extracting resources, refining them, and converting them into usable energy often require substantial expenditures, leading to high COGS (83.11%). Meanwhile, these industries face ongoing challenges and must be agile in adapting to changes, which can lead to high RC as they strive to remain competitive and environmentally sustainable.

[INSERT Table 4.13 ABOUT HERE]

The financials, IT, and communication services industries have relatively higher STKCO allocations (3.39%, 3.31% and 2.94%, respectively). These industries operate in highly competitive environments, they offer stock-based incentives to motivate employees and to attract top talent. In addition, the healthcare, IT, and communication services industries tend to have higher SG&A and R&D costs due to their unique operational demands. In healthcare, regulatory requirements and the need for advanced medical research drive substantial R&D investments, while complex administrative tasks and compliance necessitate significant SG&A expenditures. Similarly, in the IT sector, continuous innovation and rapid technological advancements demand substantial R&D investments, while the competitive landscape requires significant marketing and administrative efforts. Communication services industries face the constant need to upgrade and expand networks and technologies, resulting in higher R&D costs, while intense competition and the need to maintain customer service and support infrastructure drive up SG&A expenses.

To further examine the effects of industry-specific employee-related costs and employee satisfaction on stock returns, we categorise industries into low and high groups based on their values of COGS, RC, STKCO, SG&A, and R&D in Table 4.13. This approach allows us to compare the performance of companies with lower versus higher levels of these expenses. A summary of the groups of industries can be found in Table C19 in the appendix. Next, we sort the stocks from each group into tercile portfolios by ES_BERT, as we did in the previous section. The average returns and Fama-French 4-factor estimated alphas for both value- and equal-weighted long-short portfolios are reported in Table 4.14.

[INSERT Table 4.14 ABOUT HERE]

The results from Table 4.14 suggest that employee sentiment and their related costs may have a combined effect on stock returns. For the value-weighted portfolios, in industries with lower RC, high employee sentiment companies tend to outperform low employee sentiment companies by 0.25%. Low restructuring costs can indicate a stable work environment without frequent upheavals and uncertainty. In a low-restructuring-cost scenario where employees are valued and supported, companies can benefit from satisfied individual and team performance.

By contrast, in industries with higher STKCO and SG&A, high employee sentiment companies significantly outperform low employee sentiment companies by 0.39% and 0.54%, respectively. The equal-weighted portfolios also draw the same conclusion on SG&A. Investing in SG&A expenses allows companies to drive revenue growth and market positioning, while positive employee sentiment enhances productivity and overall company performance. The simultaneous presence of high SG&A expenses and employee sentiment can create a virtuous cycle of positive performance and growth for the company.

Furthermore, for the equal-weighted portfolios, high employee sentiment companies in lower-COGS industries show more advantages than low employee sentiment companies. Among these industries, companies often rely on efficient operations and cost-effective processes to produce goods or services. Combined with high employee sentiment companies can achieve increased operational efficiency and cost savings. On the contrary, higher-COGS industries often require a strong focus on cost management due to the significant expenses involved in producing goods or services. Low employee sentiment companies may prioritise cost-cutting measures to mitigate the impact of high COGS. Their emphasis on cost reduction can also lead to improved competitiveness and better financial performance.

We also investigate the predictability of stock returns within each industry using the Fama-Macbeth regression. This analysis controlled for employee sentiment as measured by BERT, the book-to-market ratio, company size, and employee-related costs. The results from Table 4.15 show that COGS expenses are significantly negatively correlated with stock returns in the Utility industry at a 5% significance level. With a one-unit increase in COGS within the Utility industry, the company's excess return could decrease by 3.52%. However, higher SG&A expenses can lead to an increase in stock returns by 0.40% and 0.55% in the IT and Utility industries, respectively. In addition, higher R&D expenses in the Consumer Staples industry can contribute to higher returns, but the effect is marginally small (0.01%).

[INSERT Table 4.15 ABOUT HERE]

4.5 Conclusion

This paper is the first to apply BERT for sentiment analysis on employee reviews and empirically study the relationship between employee sentiment and stock returns. We also use the review's overall star rating and the predictions of LM on the text comments to measure employee sentiment. Through the logistic regressions, we identify a common set of factors, such as Overall and Culture & Values ratings, Senior Leadership Rating, COGS, SG&A, and STKCO, that tend to drive employee sentiment across all three measures. While the concern about information loss is valid when we transform the continuous ES into categorical values, it is essential to consider that the choice of analysis method is often guided by the specific research context and objectives. In this study, our goal is to investigate the determinants of sentiment direction, specifically whether it is positive or negative. Logistic regression with binary classification served this purpose effectively by providing a clear and interpretable way to understand these determinants. Future research may explore alternative methods such as the ordinal logistic regression, which is well-suited for handling ordinal or continuous outcomes. These methods could be particularly relevant when the research objectives emphasise the nuanced differences in sentiment strength.

When sorting stocks into tercile portfolios based on sentiment measures (ES_BERT, ES_LM, and ES_Rating), we observe greater variability in portfolio returns among the sentiment measures. ES_BERT and ES_Rating indicate that portfolios with medium to high sentiment tend to outperform the market benchmark, while ES_LM suggests that the low sentiment portfolio outperforms the market. Our analysis also provides evidence of the value-added nature of BERT in investment decision-making. We show that BERT provides more advantages compared to LM when analysing high employee sentiment portfolios. However, the assessment of text, whether using BERT or LM, appears to be more empirically indicative than relying solely on the overall rating.

We further analyse the relationship between employee sentiment and stock returns across different GICS industries, finding BERT to be the most interpretable measure. According to BERT, industries such as Consumer Staples, Health Care, and IT experience significant outperformance of the benchmark in portfolios with high employee sentiment, while the Materials and Real Estate industries show underperformance. Moreover, using the Fama-MacBeth two-step regressions, our analysis reveals that positive sentiment expressed in specific topics, such as Customer Service, Work Environment, and Working Hours, can effectively predict significant increases in stock returns in both the short and long term when using BERT. In contrast, LM and overall ratings do not provide substantial evidence for predicting stock returns. In addition, we investigate the industry-specific effects of employee-related costs on both employee satisfaction and stock returns. Our findings highlight the significant impact of employee satisfaction on stock returns, particularly in industries with lower expenses directly related to production or manufacturing (COGS and RC), but higher expenses related to general operation and management of the business (SG&A and R&D).

Overall, our research demonstrates the feasibility and benefits of using text comments to assess employee sentiment in the empirical finance study. Although previous studies have frequently relied on overall star ratings as a measure of employee sentiment, these ratings lack the comprehensive understanding and specificity provided by text comments. This additional context can be crucial for evaluating the overall performance of a company. Therefore, we advocate for future research to explore the use of state-of-the-art language models in analysing textual data in finance, bridging the gap between NLP and financial applications and enhancing decision-making processes.

4.6 Tables for Chapter 4

Table 4.1: Summary Statistics of Employee Reviews and Firm-level Characteristics.

Panel A: Employee reviews

	Count	Mean	Std.	Q1	Median	Q3
Overall Rating	1,352,736	3.48	1.22	1.00	4.00	5.00
Work-Life Balance Rating	1,352,736	2.87	1.68	0.00	3.00	5.00
Culture and Values Rating	1,352,736	2.72	1.85	0.00	3.00	5.00
Senior Leadership Rating	1,352,736	2.55	1.65	0.00	3.00	5.00
Career Opportunities Rating	1,352,736	2.81	1.65	0.00	3.00	5.00
Compensation & Benefits Rating	1,352,736	2.95	1.63	0.00	3.00	5.00
Recommend To Friend	1,041,983	0.30	0.95	0.00	1.00	1.00
CEO Approval	953,286	0.35	0.74	-1.00	1.00	1.00
Business Outlook	934,510	0.31	0.78	-1.00	1.00	1.00
Employment Length	1,352,736	3.22	5.12	0.00	1.00	20.00
Current Employee	1,352,736	0.57	0.50	0.00	1.00	1.00
Full-time Employee	1,061,125	0.85	0.36	0.00	1.00	1.00

Panel B: Firm-level Characteristics

	Count	Mean	Std.	Q1	Median	Q3
Size	42,601	35,091,046.24	75,663,683.74	17,174.40	14,270,164.50	2,232,278,808.54
BM	36,934	0.52	0.63	0.00	0.37	43.15
ROA	38,246	0.14	0.10	-1.01	0.13	1.85
ROE	36,974	0.18	0.74	-39.33	0.14	37.04
Age	42,665	28.69	11.30	1.00	28.00	53.00
Excess Return	42,773	0.84	4.52	-17.02	1.26	12.68
Illiquidity	38,743	0.00	0.00	0.00	0.00	0.00
Turnover	42,746	2.41	6.14	0.00	1.72	1,117.43
COGS	41,623	1,099.85	2,414.46	-199.00	396.00	34,973.33
SG&A	36,675	311.19	647.74	-307.67	118.33	13,298.00
R&D	27,686	88.40	231.38	-1.33	15.87	5,062.67
STKCO	41,168	22.30	74.54	-50.80	5.67	1,133.33
RC	37,047	-15.87	101.97	-3,552.67	-2.67	2,166.67

Table 4.2: LM and BERT Sentiment Analysis Results

Review	Overall Rating	LM	BERT
Medical benefits and motivating employees to do more than working for pennies. No advancement opportunities and no internal growth - no loyalty to employees.	1	positive	negative
Pay is good, people around you are good and help the best they can. The company is always trying to do better with tech and new things. You are pushed to the limits, the pay is not worth the stress that you have to deal with on a day to day basis. Leadership do not care about training or assisting you to help you do better.	2	positive	negative
Good basic training, above average rewards trips, valuable experience Sales reps are expendable, sales managers aren't always concerned about longevity of employees or their interests, primary goal for managers is their quarterly bonuses, everything else is secondary. Good place to start. Stop training sales reps to sell around product short falls and limitations and spend resources improving products.	3	negative	positive
Excellent training programs and employee wellness initiatives. Tend to cut contracts too often. Great place to develop top-notch professional skills.	4	negative	positive
Opportunity to learn about business operations and how to use many of the products in the store. Sometimes communication becomes a problem at a higher level and creates confusion between sales associates and merchandising associates. Overall amazing experience if you are willing to learn and grow.	5	negative	positive

Notes: This table reports a summary of sentiment predictions made by LM and BERT for a set of employee reviews where LM failed to correctly predict the sentiment, with the overall rating serving as the ground truth (3-star and above is considered a positive review). Each review contains comments in the *Pros*, *Cons* and *Advice to Management* columns. Certain words in the review are highlighted in red or blue, indicating whether the LM considers the words positive or negative, respectively.

Table 4.3: Topic Distribution and Keywords

Topics ID	Label	Proportion	Top 20 Keywords
1	Customer Service	13.0%	customer, sale, store, service, goal, number, product, account, commission, bank, pressure, money, branch, base, metric, deal, incentive, quota, representative, order
2	Career Opportunity	10.4%	opportunity, career, growth, training, advancement, development, experience, program, advance, path, progression, move, movement, mobility, start, challenge, stability, possibility, ton, potential
3	Organisational Strategy	4.9%	business, industry, market, focus, talent, brand, term, strategy, organisation, division, marketing, change, resource, innovation, investment, model, structure, unit, result, acquisition
4	Compensation & Benefits	15.2%	benefit, pay, salary, health, insurance, vacation, plan, option, family, package, bonus, time, travel, employer, stock, match, schedule, pension, patient, incentive
5	Work Environment	4.8%	people, office, location, area, site, job, building, space, bit, perk, gym, campus, place, plant, city, parking, facility, ladder, move, world
6	Communication	6.9%	management, people, change, communication, meeting, coworker, structure, benefit, direction, idea, advice, talk, style, operation, thing, department, improvement, expectation, turnover, line
7	Team Dynamics	4.8%	team, environment, staff, support, atmosphere, member, fun, workload, expectation, workplace, colleague, event, experience, community, task, activity, stress, bit, teamwork, pace
8	Leadership	4.0%	leadership, level, leader, decision, culture, organisation, diversity, change, director, role, group, idea, entry, talent, woman, lack, individual, direction, executive, vision
9	Management	7.6%	manager, department, policy, hr, issue, problem, feedback, review, person, management, rule, friend, favouritism, performance, procedure, case, door, practice, system, situation
10	Working Hours	6.2%	day, hour, week, time, shift, supervisor, call, food, schedule, month, break, holiday, center, weekend, night, coffee, lunch, partner, overtime, minute
11	Job Security	9.2%	year, company, bonus, performance, cost, layoff, increase, job, review, contract, promotion, raise, profit, stock, end, budget, performer, contractor, morale, price
12	Work-Life Balance	7.7%	life, balance, culture, benefit, compensation, environment, work-life, flexibility, salary, good, perk, home, mobility, analyst, package, pace, environment, stability, carrier, workplace
13	Technology	5.2%	product, process, technology, system, engineer, quality, software, tech, tool, engineering, oracle, group, development, solution, resource, innovation, developer, datum, application, bureaucracy

Notes: This table reports the 13 topics we extracted from the employee reviews using Latent Dirichlet Allocation (LDA) and the topic distribution among the reviews. The last column provides the top 20 keywords in the topic, based on which we manually label the topic.

Table 4.4: Sentiment Predicted by BERT and LM Across Topics.

	BERT		LM	
Topic Labels	Neg. Reviews (%)	Pos. Reviews (%)	Neg. Reviews (%)	Pos. Reviews (%)
Customer Service	24.9	75.1	46.6	53.4
Career Opportunity	15.9	84.1	25.4	74.6
Organisational Strategy	22.6	77.4	39.0	61.0
Compensation & Benefits	18.1	81.9	41.4	58.6
Work Environment	16.8	83.2	40.7	59.3
Communication	34.8	65.2	56.0	44.0
Team Dynamics	14.9	85.1	38.4	61.6
Leadership	26.8	73.2	35.2	64.8
Management	31.4	68.6	47.8	52.2
Working Hours	23.4	76.6	50.5	49.5
Job Security	26.4	73.6	42.0	58.0
Work-Life Balance	15.5	84.5	41.2	58.8
Technology	21.7	78.3	45.1	54.9

Notes: This table reports the proportions of positive and negative reviews predicted by BERT and LM for different topics.

Table 4.5: Topic Distribution per Industry (in %).

Topic Labels/GICS	10	15	20	25	30	35	40	45	50	55	60
Compensation and Benefits	15.27	14.84	19.12	11.74	15.82	17.09	12.81	18.34	11.96	14.63	9.79
Career Opportunity	15.36	11.69	11.90	7.42	8.69	11.14	13.85	10.10	9.38	12.47	12.72
Job Security	14.70	12.32	11.73	5.17	6.93	10.12	9.02	11.11	10.03	13.29	16.29
Customer Service	3.00	7.74	4.78	28.37	15.75	7.46	12.20	4.43	16.57	2.59	5.98
Management	8.34	8.38	7.86	7.57	8.15	8.87	7.44	6.54	7.07	9.34	9.71
Communication	8.05	8.30	7.88	7.20	8.32	7.14	6.34	5.03	8.49	8.18	8.92
Work-Life Balance	7.58	5.99	6.68	3.84	5.34	7.53	10.05	12.31	6.97	7.19	5.61
Organizational Strategy	6.22	7.54	5.83	2.23	5.87	5.21	6.51	5.55	4.03	7.23	5.90
Work Environment	5.77	6.40	5.18	4.08	5.17	4.99	5.18	4.50	5.64	6.77	5.97
Working Hours	3.23	4.26	4.35	12.72	8.39	4.65	3.53	2.46	6.15	3.96	4.10
Leadership	4.12	5.38	4.71	2.45	4.05	5.33	4.48	4.08	4.28	7.44	5.80
Team Dynamics	3.29	3.19	3.64	5.39	5.89	5.56	4.91	4.72	4.09	3.05	6.67
Technology	5.09	3.97	6.35	1.82	1.63	4.90	3.68	10.84	5.33	3.87	2.55

Notes: The Global Industry Classification Standard (GICS) classification is as follows. 10: Energy, 15: Materials, 20: Industrials, 25: Consumer Discretionary, 30: Consumer Staples, 35: Health Care, 40: Financials, 45: Information Technology, 50: Communication Services, 55: Utilities, 60: Real Estate.

Table 4.6: Determinants of Employee Sentiment

	ES_BERT			ES_LM			ES_Rating		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Overall Rating	0.45***		0.41***	0.16***		0.18***	–		–
	(0.05)		(0.06)	(0.04)		(0.04)	–		–
Work-Life Balance Rating	-0.05		-0.01	-0.02		-0.00	-0.05		-0.07*
	(0.03)		(0.04)	(0.03)		(0.03)	(0.03)		(0.04)
Culture & Values Rating	0.27***		0.23***	0.13***		0.12***	0.28***		0.29***
	(0.03)		(0.04)	(0.02)		(0.03)	(0.03)		(0.04)
Senior Leadership Rating	-0.13***		-0.06	-0.07**		-0.06	-0.13***		-0.08
	(0.04)		(0.05)	(0.03)		(0.04)	(0.04)		(0.06)
Career Opportunities Rating	0.06		0.00	0.09***		0.07*	-0.04		-0.07
	(0.04)		(0.05)	(0.03)		(0.04)	(0.04)		(0.05)
Compensation & Benefits Rating	-0.02		-0.02	-0.03		-0.03	-0.07*		-0.04
	(0.04)		(0.04)	(0.03)		(0.04)	(0.03)		(0.05)
Recommend To Friend	0.18***		0.24***	0.08		0.03	0.31***		0.35***
	(0.07)		(0.08)	(0.05)		(0.06)	(0.06)		(0.08)
CEO Approval	0.01		-0.02	-0.01		-0.01	-0.15***		-0.16**
	(0.06)		(0.07)	(0.05)		(0.06)	(0.06)		(0.07)
Business Outlook	0.07		0.03	0.02		0.08	-0.05		-0.11
	(0.08)		(0.09)	(0.06)		(0.07)	(0.08)		(0.09)
Employment Length	-0.08		-0.13	-0.05		-0.09	0.13*		0.03
	(0.08)		(0.10)	(0.05)		(0.07)	(0.07)		(0.09)
Current Employee	-0.00		0.00	0.01		0.00	-0.01		-0.01
	(0.01)		(0.01)	(0.01)		(0.01)	(0.01)		(0.01)
Full-time Employee	-0.53***		-0.38***	-0.33***		-0.26***	-0.48***		-0.46***
	(0.09)		(0.11)	(0.07)		(0.10)	(0.09)		(0.12)
COGS		0.08**	0.08***		0.08***	0.08***		0.07**	0.06**
		(0.03)	(0.03)		(0.02)	(0.02)		(0.03)	(0.03)
RC		-0.00	-0.00		0.00	0.00		-0.00	-0.00
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
STKCO		0.01*	0.01*		0.00	-0.00		0.01**	0.00*
		(0.01)	(0.01)		(0.00)	(0.00)		(0.01)	(0.00)
SG&A		0.11***	0.10***		0.06***	0.06***		0.08***	0.05**
		(0.02)	(0.02)		(0.02)	(0.02)		(0.02)	(0.02)
R&D		-0.00	-0.00		0.00**	0.00**		0.00*	0.00*
		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)
Constant	1.41***	1.59***	0.43	1.54***	1.29***	0.71***	0.27*	1.65***	-0.29
	(0.18)	(0.19)	(0.29)	(0.13)	(0.14)	(0.22)	(0.16)	(0.17)	(0.27)
Observations	42,773	42,773	42,773	42,773	42,773	42,773	42,773	42,773	42,773
Pseudo-R2(%)	0.0449	0.0212	0.0632	0.00885	0.00885	0.0183	0.105	0.0229	0.124

Notes: This table reports the employee sentiment index (ES) determinants. Using the text comments, the ES is constructed by aggregating the sentiment predictions generated by BERT and the Loughran-McDonald (LM) dictionary for the reviews of each company on a monthly basis. We also use the review's overall star rating as a judge of the sentiment, we consider a review to be positive if it has 3 or more stars otherwise negative. The ES is calculated by subtracting the number of negative reviews from the number of positive reviews and dividing the result by the total number of reviews. Since ES_Rating is transformed from the overall ratings and they are highly correlated, we omit the Overall Rating in the last set of regressions. Firm-level cluster-adjusted standard errors are provided in parentheses. The asterisks *, **, and ***, respectively, denote the significance at the 10%, 5%, and 1% levels.

Table 4.7: Portfolio Returns Sorted by Employee Sentiment.

Panel A. Portfolio average excess returns

	Value-weighted			Equal-weighted		
	ES_BERT	ES_LM	ES_Rating	ES_BERT	ES_LM	ES_Rating
Low	1.31***	1.59***	1.30***	1.27***	1.32***	1.12
	(3.61)	(4.16)	(3.37)	(2.46)	(2.51)	(2.14)
Mid	1.34***	1.37***	1.45***	1.12	1.20***	1.15***
	(3.72)	(3.80)	(4.05)	(2.43)	(2.64)	(2.53)
High	1.56***	1.27***	1.44***	1.24***	1.11	1.35***
	(4.18)	(3.34)	(3.73)	(2.56)	(2.30)	(2.74)
High-Low	0.25***	-0.32	0.14	-0.02	-0.21	0.23
	(2.53)	(-1.55)	(0.70)	(-0.16)	(-1.47)	(1.21)
β^{MKT}-RF	1.36***	1.37***	1.35***	1.17	1.17	1.16
	(3.73)	(3.74)	(3.66)	(2.40)	(2.40)	(2.39)

Panel B. Portfolio alphas

	Value-weighted											
	ES_BERT				ES_LM				ES_Rating			
	Low	Mid	High	High-Low	Low	Mid	High	High-Low	Low	Mid	High	High-Low
Alpha	-0.01	0.02	0.12**	0.13	0.21**	0.00	-0.07	-0.28*	-0.05	0.16*	0.02	0.07
	(-0.09)	(0.31)	(2.08)	(1.22)	(2.21)	(0.05)	(-0.77)	(-1.64)	(-0.39)	(1.88)	(0.22)	(0.33)
β^{MKT}	0.97***	0.99***	1.04***	0.08*	1.01***	1.02***	0.97***	-0.03	0.96***	0.98***	1.06***	0.10**
	(43.47)	(51.43)	(37.53)	(1.63)	(41.21)	(42.74)	(58.72)	(-1.01)	(37.65)	(39.84)	(50.15)	(2.39)
SMB	0.13***	-0.06**	-0.07*	-0.20***	0.05*	-0.03	-0.02	-0.08	0.15***	-0.05	-0.10**	-0.25***
	(4.52)	(-2.21)	(-1.84)	(-3.2)	(1.82)	(-0.96)	(-0.72)	(-1.44)	(2.9)	(-1.57)	(-2.29)	(-2.81)
HML	0.04*	0.07*	-0.11***	-0.15***	0.01	0.06	-0.08**	-0.09	-0.07*	0.09***	-0.03	0.04
	(1.66)	(1.75)	(-3.31)	(-3.14)	(0.27)	(1.10)	(-2.26)	(-1.54)	(-1.90)	(3.41)	(-1.03)	(0.70)
MOM	0.04**	-0.02	-0.02	-0.07***	-0.03	-0.00	0.03	0.06	-0.01	-0.04***	0.05**	0.06
	(2.37)	(-0.75)	(-1.50)	(-2.69)	(-0.98)	(-0.10)	(1.02)	(1.03)	(-0.37)	(-2.46)	(2.24)	(1.45)
	Equal-weighted											
	ES_BERT				ES_LM				ES_Rating			
	Low	Mid	High	High-Low	Low	Mid	High	High-Low	Low	Mid	High	High-Low
Alpha	0.09	-0.02	0.07*	-0.02	0.12*	0.05	-0.03	-0.15	-0.08	0.03	0.19*	0.27
	(1.05)	(-0.38)	(1.76)	(-0.2)	(1.76)	(0.73)	(-0.55)	(-1.35)	(-0.75)	(0.37)	(1.74)	(1.32)
β^{MKT}	1.01***	0.99***	1.0***	-0.01	1.01***	1.01***	0.97***	-0.04*	1.00***	0.99***	1.00***	0.00
	(86.82)	(84.82)	(89.42)	(-0.62)	(64.49)	(73.19)	(72.09)	(-1.69)	(60.86)	(63.01)	(54.20)	(0.14)
SMB	0.1**	-0.08***	-0.02	-0.12	0.13***	-0.11***	-0.02	-0.14**	0.14**	-0.09***	-0.05	-0.20
	(2.24)	(-3.38)	(-0.61)	(-1.52)	(2.71)	(-2.93)	(-0.48)	(-1.98)	(2.09)	(-2.72)	(-0.95)	(-1.62)
HML	0.05***	-0.01	-0.04	-0.09***	0.01	0.03	-0.05*	-0.06	-0.04	0.03	0.01	0.04
	(3.67)	(-0.48)	(-1.48)	(-2.80)	(0.31)	(0.70)	(-1.91)	(-1.33)	(-1.50)	(1.33)	(0.18)	(0.89)
MOM	0.05*	0.01	-0.06***	-0.10***	-0.00	0.03	-0.03	-0.03	-0.03	0.05	-0.02	0.01
	(1.87)	(0.55)	(-3.09)	(-2.78)	(-0.06)	(1.01)	(-1.52)	(-0.64)	(-0.94)	(1.35)	(-0.38)	(0.18)

Notes: This table presents the returns of equal- and value-weighted portfolios, sorted based on the monthly Employee Sentiment Index (ES) generated by BERT, LM and the overall star ratings, between January 2008 and March 2021. Stocks are divided into tercile portfolios according to their ES each month and are rebalanced at the end of the month. The top third of firms with the highest ES are placed in the High Sentiment portfolio, the bottom third with the lowest ES are placed in the Low Sentiment portfolio, and the difference is sorted into the long-short portfolio, represented by High-Low. The remaining firms are placed in the Neutral Sentiment portfolio, also known as the Mid group. These portfolios are risk-adjusted. Panel A summarises the average excess returns of these portfolios, as well as the market excess returns (MKT-RF). Panel B summarises the Fama-French-Carhart (FFC) four-factor estimations. The values are reported in percentage terms. The T-statistics with Newey-West adjustments are provided in parentheses. The asterisks *, **, and ***, respectively, indicate the significance at the 10%, 5%, and 1% levels.

Table 4.8: Portfolio sorted by Δ(ES_BERT-ES_LM)

Panel A. Portfolio average excess returns

	Value-weighted	Equal-weighted
Low (ES_BERT-ES_LM)	1.54***	1.43***
	(3.47)	(2.44)
Mid (ES_BERT-ES_LM)	1.36***	1.32
	(3.13)	(2.17)
High (ES_BERT-ES_LM)	1.66***	1.60***
	(3.54)	(2.85)
High-Low (ES_BERT-ES_LM)	0.12	0.17
	(0.76)	(0.77)

Panel B. Portfolio alphas

	Value-weighted				Equal-weighted			
	Low (ES_BERT-ES_LM)	Mid (ES_BERT-ES_LM)	High (ES_BERT-ES_LM)	High-Low (ES_BERT-ES_LM)	Low (ES_BERT-ES_LM)	Mid (ES_BERT-ES_LM)	High (ES_BERT-ES_LM)	High-Low (ES_BERT-ES_LM)
Alpha	0.08	-0.09	0.17*	0.09	0.00	-0.08	0.23**	0.23
	(0.73)	(-0.78)	(1.72)	(0.56)	(0.01)	(-1.20)	(1.96)	(1.12)
β^{MKT}	0.99***	0.99***	1.01***	0.02	1.02***	1.00***	0.97***	-0.05
	(36.77)	(38.17)	(30.62)	(0.33)	(36.77)	(46.54)	(38.93)	(-0.98)
SMB	0.01	0.02	-0.03	-0.04	-0.01	0.03	-0.02	-0.01
	(0.18)	(0.50)	(-0.63)	(-0.45)	(-0.32)	(0.92)	(-0.46)	(-0.13)
HML	0.04	0.06	-0.10*	-0.14	-0.02	0.03	-0.01	0.02
	(0.92)	(1.35)	(-1.75)	(-1.51)	(-0.45)	(0.73)	(-0.12)	(0.19)
MOM	0.00	0.05*	-0.05*	-0.06	0.03	0.01	-0.04	-0.07
	(0.20)	(1.73)	(-1.71)	(-1.30)	(0.42)	(0.77)	(-0.69)	(-0.55)

Notes: This table presents the returns of equal- and value-weighted portfolios, sorted based on the monthly Employee Sentiment Index (ES) differences generated by BERT and LM, between January 2008 and March 2021. Stocks are divided into tercile portfolios according to their ΔES each month and are rebalanced at the end of the month. The top third of firms with the highest ΔES are placed in the High Sentiment portfolio, the bottom third with the lowest ΔES are placed in the Low Sentiment portfolio, and the difference is sorted into the long-short portfolio, represented by High-Low. The remaining firms are placed in the Neutral Sentiment portfolio, also known as the Mid group. These portfolios are risk-adjusted. Panel A summarises the average excess returns of these portfolios. Panel B summarises the Fama-French-Carhart (FFC) four-factor estimations. The values are reported in percentage terms. The T-statistics with Newey-West adjustments are provided in parentheses. The asterisks *, **, and ***, respectively, indicate the significance at the 10%, 5%, and 1% levels.

Table 4.9: Portfolio sorted by Δ(ES_BERT-ES_Rating)

Panel A. Portfolio average excess returns

	Value-weighted	Equal-weighted
Low (ES_BERT-ES_Rating)	1.45***	1.83***
	(3.41)	(3.10)
Mid (ES_BERT-ES_Rating)	1.92***	1.27
	(4.09)	(2.02)
High (ES_BERT-ES_Rating)	1.51***	1.47
	(3.34)	(2.30)
High-Low (ES_BERT-ES_Rating)	0.06	-0.35
	(0.27)	(-1.31)

Panel B. Portfolio alphas

	Value-weighted				Equal-weighted			
	Low (ES_BERT-ES_Rating)	Mid (ES_BERT-ES_Rating)	High (ES_BERT-ES_Rating)	High-Low (ES_BERT-ES_Rating)	Low (ES_BERT-ES_Rating)	Mid (ES_BERT-ES_Rating)	High (ES_BERT-ES_Rating)	High-Low (ES_BERT-ES_Rating)
Alpha	-0.01	0.35***	-0.19	-0.18	0.34**	-0.19	-0.00	-0.35
	(-0.07)	(3.12)	(-1.20)	(-0.70)	(1.95)	(-1.31)	(-0.02)	(-1.43)
β^{MKT}	0.92***	1.00***	1.07***	0.15*	1.01***	1.00***	0.98***	-0.03
	(25.94)	(22.46)	(19.17)	(1.79)	(32.70)	(32.43)	(42.93)	(-0.68)
SMB	-0.01	-0.06	0.06	0.06	-0.05	-0.05	0.10**	0.15**
	(-0.13)	(-1.37)	(1.05)	(0.71)	(-1.47)	(-1.31)	(2.14)	(2.14)
HML	0.09*	0.01	-0.10	-0.19*	0.08*	0.00	-0.08	-0.16*
	(1.64)	(0.29)	(-1.50)	(-1.68)	(1.64)	(0.09)	(-1.57)	(-1.80)
MOM	0.01	-0.03	0.03	0.02	0.04	0.03	-0.07*	-0.12**
	(0.25)	(-1.11)	(0.93)	(0.51)	(1.02)	(0.42)	(-1.84)	(-2.40)

Notes: This table presents the returns of equal- and value-weighted portfolios, sorted based on the monthly Employee Sentiment Index (ES) differences generated by BERT and Rating, between January 2008 and March 2021. Stocks are divided into tercile portfolios according to their ΔES each month and are rebalanced at the end of the month. The top third of firms with the highest ΔES are placed in the High Sentiment portfolio, the bottom third with the lowest ΔES are placed in the Low Sentiment portfolio, and the difference is sorted into the long-short portfolio, represented by High-Low. The remaining firms are placed in the Neutral Sentiment portfolio, also known as the Mid group. These portfolios are risk-adjusted. Panel A summarises the average excess returns of these portfolios. Panel B summarises the Fama-French-Carhart (FFC) four-factor estimations. The values are reported in percentage terms. The T-statistics with Newey-West adjustments are provided in parentheses. The asterisks *, **, and ***, respectively, indicate the significance at the 10%, 5%, and 1% levels.

Table 4.10: Portfolio sorted by Δ(ES_LM-ES_Rating)

Panel A. Portfolio average excess returns

	Value-weighted	Equal-weighted
Low (ES_LM-ES_Rating)	1.77***	1.88***
	(3.70)	(3.10)
Mid (ES_LM-ES_Rating)	1.84***	1.41
	(3.82)	(2.29)
High (ES_LM-ES_Rating)	1.42***	1.41***
	(3.13)	(2.50)
High-Low (ES_LM-ES_Rating)	-0.35	-0.47
	(-1.25)	(-1.86)

Panel B. Portfolio alphas

	Value-weighted				Equal-weighted			
	Low (ES_LM-ES_Rating)	Mid (ES_LM-ES_Rating)	High (ES_LM-ES_Rating)	High-Low (ES_LM-ES_Rating)	Low (ES_LM-ES_Rating)	Mid (ES_LM-ES_Rating)	High (ES_LM-ES_Rating)	High-Low (ES_LM-ES_Rating)
Alpha	0.10	0.20	-0.16	-0.26	0.32*	-0.12	-0.05	-0.36
	(0.76)	(1.52)	(-0.99)	(-0.97)	(1.91)	(-0.84)	(-0.49)	(-1.58)
β^{MKT}	1.02***	1.01***	0.97***	-0.05	1.02***	1.02***	0.95***	-0.07***
	(32.53)	(29.07)	(34.53)	(-1.04)	(53.07)	(34.12)	(44.50)	(-2.69)
SMB	0.07	-0.10*	0.02	-0.05	0.08**	-0.13***	0.05	-0.03
	(1.36)	(-1.81)	(0.33)	(-0.51)	(2.15)	(-2.52)	(0.96)	(-0.45)
HML	0.11***	0.00	-0.11	-0.22**	0.06	0.04	-0.10	-0.15*
	(2.47)	(0.08)	(-1.58)	(-2.18)	(1.59)	(0.67)	(-1.56)	(-1.90)
MOM	-0.00	-0.02	0.03	0.03	0.04	-0.02	-0.02	-0.06
	(-0.01)	(-0.63)	(0.63)	(0.39)	(0.79)	(-0.43)	(-0.71)	(-0.96)

Notes: This table presents the returns of equal- and value-weighted portfolios, sorted based on the monthly Employee Sentiment Index (ES) differences generated by LM and Rating, between January 2008 and March 2021. Stocks are divided into tercile portfolios according to their ΔES each month and are rebalanced at the end of the month. The top third of firms with the highest ΔES are placed in the High Sentiment portfolio, the bottom third with the lowest ΔES are placed in the Low Sentiment portfolio, and the difference is sorted into the long-short portfolio, represented by High-Low. The remaining firms are placed in the Neutral Sentiment portfolio, also known as the Mid group. These portfolios are risk-adjusted. Panel A summarises the average excess returns of these portfolios. Panel B summarises the Fama-French-Carhart (FFC) four-factor estimations. The values are reported in percentage terms. The T-statistics with Newey-West adjustments are provided in parentheses. The asterisks *, **, and ***, respectively, indicate the significance at the 10%, 5%, and 1% levels.

Table 4.11: Industry-specific: Long-short Portfolio Returns Sorted by Employee Sentiment

	Value-weighted portfolios High-Low						Equal-weighted portfolios High-Low					
	Average excess return			4-Factor alpha			Average excess return			4-Factor alpha		
GICS Industry	ES_BERT	ES_LM	ES_Rating	ES_BERT	ES_LM	ES_Rating	ES_BERT	ES_LM	ES_Rating	ES_BERT	ES_LM	ES_Rating
Energy	0.15	0.14	-0.15	0.08	0.22	-0.08	-0.19	-0.45	-0.11	-0.30	-0.35	0.03
	(0.33)	(0.23)	(-0.37)	(0.17)	(0.45)	(-0.20)	(-0.32)	(-1.02)	(-0.24)	(-0.60)	(-0.94)	(0.05)
Materials	-1.03	0.52	-0.38	-1.13**	0.57	-0.15	-1.04	0.87	0.27	-1.13***	1.05**	0.56
	(-1.50)	(1.13)	(-0.81)	(-1.96)	(1.00)	(-0.35)	(-1.75)	(1.57)	(0.56)	(-2.46)	(2.01)	(1.23)
Industrials	-0.13	0.02	0.02	-0.31	0.07	-0.06	0.05	-0.00	0.13	0.06	0.05	0.05
	(-0.56)	(0.07)	(0.09)	(-1.47)	(0.30)	(-0.22)	(0.17)	(-0.01)	(0.51)	(0.28)	(0.28)	(0.19)
Consumer Discretionary	0.46	0.38	0.93***	0.37	0.48	0.95**	0.56	0.34	0.76	0.36	0.30	0.68**
	(1.16)	(0.93)	(2.58)	(0.81)	(1.39)	(2.12)	(1.60)	(1.32)	(2.15)	(1.03)	(1.11)	(1.94)
Consumer Staples	0.89***	-0.06	0.34	1.04***	0.06	0.54	0.40	0.69	0.38	0.60*	0.72***	0.57*
	(2.89)	(-0.16)	(0.94)	(3.31)	(0.17)	(1.56)	(1.08)	(2.19)	(1.23)	(1.79)	(2.55)	(1.66)
Health Care	0.52	-0.17	0.09	0.66**	-0.39	-0.01	0.43	-0.26	0.51	0.57**	-0.43	0.55*
	(1.88)	(-0.61)	(0.31)	(2.31)	(-1.52)	(-0.03)	(1.56)	(-1.01)	(1.46)	(2.31)	(-1.54)	(1.78)
Financials	0.45	-0.85***	0.05	0.35	-0.62*	-0.14	-0.14	-0.49***	0.28	-0.08	-0.29	0.42
	(1.26)	(-2.73)	(0.17)	(1.05)	(-1.88)	(-0.43)	(-0.49)	(-2.65)	(0.86)	(-0.26)	(-1.37)	(1.19)
Information Technology	0.64	0.38	-0.08	0.56*	0.52	-0.25	0.85***	0.67	0.66	0.91***	0.88***	0.65*
	(2.04)	(1.13)	(-0.16)	(1.86)	(1.25)	(-0.59)	(4.06)	(2.39)	(2.25)	(3.89)	(3.16)	(1.88)
Communication Services	0.13	0.38	1.10	0.11	0.52	0.61	0.89	0.67	1.18	1.14*	0.88***	1.41**
	(0.20)	(1.13)	(1.27)	(0.19)	(1.25)	(0.84)	(1.30)	(2.39)	(1.50)	(1.65)	(3.16)	(2.08)
Utilities	0.12	0.10	0.03	0.16	0.21	0.20	0.22	-0.24	0.07	0.24	-0.15	0.19
	(0.50)	(0.53)	(0.08)	(0.63)	(1.17)	(0.71)	(0.83)	(-0.92)	(0.20)	(0.85)	(-0.60)	(0.65)
Real Estate	-0.84	-0.90	0.05	-1.05**	-0.68	-0.24	-0.78	-0.38	0.42	-0.84**	-0.21	0.33
	(-1.76)	(-1.38)	(0.09)	(-2.10)	(-1.18)	(-0.39)	(-2.05)	(-0.96)	(0.96)	(-1.99)	(-0.51)	(0.82)

Notes: This table presents the returns of equal- and value-weighted long-short portfolios across each of the 11 GICS industries. In each industry, the stocks are sorted based on the monthly Employee Sentiment Index (ES) generated by BERT, LM and overall star ratings. The stocks are divided into tercile portfolios based on ES each month, and this table focuses on the High-Low group, which represents the difference between the group with the highest ES and the group with the lowest ES. Stocks are rebalanced at the end of each month, and the returns are adjusted for the risk-free rate. This table provides the average monthly excess returns and the Fama-French-Carhart (FFC) four-factor alpha in percentage terms. T-statistics with Newey-West adjustments are provided in parentheses. The asterisks *, **, and ***, respectively, indicate the significance at the 10%, 5%, and 1% levels.

Table 4.12: Fama-Macbeth Regression for Return Predictability

Panel A							
	ER(t+1)			ER(t+3)	ER(t+6)	ER(t+9)	ER(t+12)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Employee Sentiment							
ES_BERT	0.25*	0.12**	0.19**	0.30**	0.23	0.17	0.15
	(1.70)	(2.03)	(2.44)	(2.01)	(1.27)	(0.59)	(0.51)
Firm Characteristics							
log(BM)		-0.05	-0.26	-0.04	-0.22	-0.34	-0.19
		(-0.56)	(-1.63)	(-0.25)	(-1.07)	(-1.30)	(-0.70)
log(Size)		0.17	-0.13	0.28	0.42	0.46	0.55*
		(1.15)	(-0.70)	(0.81)	(1.51)	(1.58)	(1.70)
log(Illiquidity)		0.13	-0.02	0.34	0.55*	0.75***	0.67***
		(1.30)	(-0.12)	(0.90)	(1.86)	(2.84)	(2.65)
ROA		-0.96	-1.12	-1.21	0.30	1.24	0.29
		(-1.56)	(-1.37)	(-0.99)	(0.20)	(0.78)	(0.19)
ROE		-0.16	0.59	-0.11	-0.54	-0.09	-1.12
		(-0.47)	(1.03)	(-0.14)	(-0.79)	(-0.11)	(-1.41)
Turnover		0.03	-0.03	0.10	0.08	0.06	0.04
		(0.87)	(-0.57)	(0.66)	(0.51)	(0.59)	(0.26)
Age		-0.00	-0.01	-0.01	-0.01	-0.00	-0.01
		(-1.23)	(-1.18)	(-0.99)	(-0.46)	(-0.05)	(-0.36)
COGS		0.03	0.02	-0.06	-0.01	0.10	0.07
		(0.94)	(0.27)	(-0.43)	(-0.06)	(0.76)	(0.47)
RC		0.00	-0.00	-0.00	0.00	0.00	0.00
		(0.06)	(-0.14)	(-0.53)	(0.67)	(0.51)	(0.28)
STKCO		0.00	0.01	-0.00	0.00	0.01*	0.00
		(0.89)	(1.13)	(-0.24)	(0.11)	(1.72)	(0.52)
SG&A		-0.05	-0.07	0.07	0.15	-0.03	0.03
		(-1.44)	(-0.97)	(0.50)	(0.70)	(-0.31)	(0.25)
R&D		0.01	0.01	0.02	0.01	0.01	-0.00
		(0.98)	(1.01)	(1.02)	(1.02)	(0.97)	(-0.92)
Topics x ES							
Customer Service			0.42	1.99*	2.60	0.70	2.96
			(0.71)	(1.68)	(1.51)	(0.30)	(1.46)
Career Opportunity			0.81	1.50	2.31	1.90	4.13
			(1.18)	(1.08)	(1.01)	(0.64)	(1.52)
Organisational Strategy			0.53	1.21	2.12	1.24	3.81**
			(0.89)	(1.39)	(1.37)	(0.62)	(2.08)
Compensation & Benefits			0.71	1.99	2.81	2.64	4.66
			(1.14)	(1.42)	(1.23)	(0.83)	(1.53)
Work Environment			0.65	2.65**	3.79*	2.72	4.79**
			(1.20)	(2.23)	(1.88)	(0.96)	(2.21)
Communication			0.02	1.00	1.65	1.18	3.27
			(0.03)	(0.82)	(0.98)	(0.52)	(1.54)
Team Dynamics			1.82	0.84	1.89	-0.63	0.11
			(1.05)	(0.62)	(1.12)	(-0.23)	(0.04)
Leadership			-0.40	0.66	1.50	2.10	3.40*
			(-0.76)	(0.49)	(0.86)	(0.97)	(1.71)
Management			0.23	2.33	5.91	6.16	6.90
			(0.34)	(1.20)	(1.29)	(1.08)	(1.49)
Working Hours			1.04	2.42*	2.12	0.90	2.07
			(1.51)	(1.77)	(1.27)	(0.46)	(1.05)
Job Security			0.13	1.65	1.98	1.22	3.72*
			(0.20)	(1.46)	(1.24)	(0.61)	(1.90)
Work-Life Balance			0.47	1.91	2.31	1.04	2.25
			(0.58)	(1.09)	(1.02)	(0.57)	(1.32)
Technology			0.47	1.02	1.68	1.37	3.54
			(0.72)	(0.83)	(0.94)	(0.54)	(1.41)
Observations	42,773	42,773	42,773	42,773	42,773	42,773	42,773
R-squared	0.02	0.05	0.19	0.19	0.18	0.19	0.19
Num of groups	111	110	110	110	110	110	110
Pseudo-R2(%)	0.015	0.046	0.190	0.185	0.185	0.189	0.190

Table 4.12: Fama-Macbeth Regression for Return Predictability - Continued

Panel B							
	ER(t+1)			ER(t+3)	ER(t+6)	ER(t+9)	ER(t+12)
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Employee Sentiment							
ES_LM	0.08	0.35	0.07	0.12	0.11	0.15	0.09
	(1.47)	(1.25)	(1.21)	(1.08)	(0.78)	(0.86)	(0.52)
Firm Characteristics							
log(BM)		-0.06	-0.18*	0.08	0.02	-0.12	-0.23
		(-0.72)	(-1.74)	(0.41)	(0.07)	(-0.48)	(-0.82)
log(Size)		0.07	0.00	0.57*	0.76**	0.79**	0.72*
		(0.66)	(0.03)	(1.93)	(2.00)	(2.18)	(1.73)
log(Illiquidity)		0.10	0.05	0.66	0.90**	0.91***	0.80**
		(1.16)	(0.54)	(1.62)	(2.12)	(3.05)	(2.64)
ROA		-0.78	-0.99	-0.70	0.57	0.95	-0.16
		(-1.28)	(-1.38)	(-0.50)	(0.33)	(0.61)	(-0.11)
ROE		-0.13	0.23	-0.64	-1.41	-0.85	-0.37
		(-0.41)	(0.71)	(-0.55)	(-1.08)	(-1.26)	(-0.51)
Turnover		0.01	-0.00	0.20	0.26	0.20***	0.18*
		(0.36)	(-0.07)	(1.27)	(1.45)	(2.84)	(1.98)
Age		-0.00	-0.01	-0.00	0.01	0.01	0.01
		(-0.89)	(-1.25)	(-0.31)	(0.61)	(0.97)	(0.35)
COGS		0.01	0.03	-0.01	0.01	0.10	0.12
		(0.30)	(0.53)	(-0.10)	(0.14)	(0.84)	(0.91)
RC		-0.00	-0.00	-0.00	-0.00	-0.00	-0.01
		(-0.64)	(-0.81)	(-0.94)	(-0.74)	(-0.87)	(-1.16)
STKCO		0.00	0.00	-0.01	-0.01	0.00	0.01
		(1.20)	(1.06)	(-0.72)	(-0.51)	(0.84)	(1.08)
SG&A		-0.01	-0.05	0.27	0.39	-0.02	-0.08
		(-0.24)	(-0.70)	(0.92)	(0.90)	(-0.18)	(-0.73)
R&D		0.01	0.01	0.02	0.01	0.01	-0.01
		(0.99)	(1.00)	(1.02)	(1.02)	(0.95)	(-1.00)
Topics x ES							
Customer Service			-1.52	0.24	2.05	1.09	2.35
			(-1.08)	(0.14)	(0.90)	(0.24)	(0.50)
Career Opportunity			-0.74	-0.05	2.05	1.86	5.03
			(-0.48)	(-0.03)	(0.69)	(0.38)	(0.81)
Organisational Strategy			-0.53	-0.80	2.65	3.46	7.49
			(-0.41)	(-0.53)	(1.06)	(0.79)	(1.32)
Compensation and Benefits			-1.05	-0.14	2.11	2.51	5.91
			(-0.71)	(-0.09)	(0.73)	(0.49)	(0.86)
Work Environment			-1.16	0.59	1.80	2.02	5.02
			(-0.71)	(0.30)	(0.67)	(0.39)	(0.88)
Communication			-1.37	0.21	2.88	4.69	8.24**
			(-0.96)	(0.15)	(1.46)	(1.47)	(2.02)
Team Dynamics			-0.81	-1.04	1.73	2.65	6.48
			(-1.11)	(-0.81)	(0.76)	(1.03)	(1.36)
Leadership			-2.46	-1.71	-0.07	0.68	2.78
			(-1.50)	(-0.84)	(-0.03)	(0.15)	(0.50)
Management			-2.05	-0.89	-0.19	-1.54	-1.37
			(-1.33)	(-0.51)	(-0.08)	(-0.31)	(-0.28)
Working Hours			-0.96	1.49	2.68	3.04	6.26
			(-0.62)	(0.96)	(1.23)	(0.84)	(1.41)
Job Security			-1.91	-1.43	0.53	1.14	4.45
			(-1.34)	(-0.76)	(0.21)	(0.24)	(0.82)
Work-Life Balance			-1.42	-0.24	2.49	1.97	3.95
			(-1.24)	(-0.21)	(1.32)	(0.59)	(0.90)
Technology			-1.10	0.22	1.58	0.65	3.63
			(-0.80)	(0.13)	(0.52)	(0.13)	(0.55)
Observations	42,773	42,773	42,773	42,773	42,773	42,773	42,773
R-squared	0.01	0.05	0.19	0.18	0.18	0.19	0.19
Num of groups	111	110	110	110	110	110	110
Pseudo-R2(%)	0.007	0.046	0.190	0.185	0.185	0.188	0.189

Table 4.12: Fama-Macbeth Regression for Return Predictability - Continued

Panel C							
	ER(t+1)			ER(t+3)	ER(t+6)	ER(t+9)	ER(t+12)
	(15)	(16)	(17)	(18)	(19)	(20)	(21)
Employee Sentiment							
ES_Rating	0.15	-0.10	-0.08	-0.01	0.03	0.01	0.10
	(1.23)	(-0.77)	(-0.79)	(-0.04)	(0.17)	(0.05)	(0.75)
Firm Characteristics							
log(BM)		-0.04	-0.03	-0.00	0.05	0.11	0.10
		(-1.64)	(-0.76)	(-0.04)	(0.21)	(0.48)	(0.51)
log(Size)		0.03	0.06	0.25	0.21	0.35	0.57
		(0.72)	(0.98)	(1.13)	(0.81)	(1.12)	(1.04)
log(Illiquidity)		0.01	0.05	0.21	0.12	0.24	0.40
		(1.04)	(1.06)	(1.14)	(0.86)	(1.23)	(1.19)
ROA		-0.73*	-0.38	0.59	1.39	2.68**	2.69***
		(-1.69)	(-1.42)	(0.92)	(1.56)	(2.14)	(2.75)
ROE		0.27	0.02	-0.05	-0.27	-0.36	-0.35
		(1.18)	(0.24)	(-0.38)	(-0.83)	(-0.93)	(-1.53)
Turnover		-0.04	0.02	0.04**	0.03	0.07	-0.02
		(-0.79)	(1.16)	(2.10)	(0.87)	(1.24)	(-0.30)
Age		-0.01	0.00	0.03	0.00	0.01	0.01
		(-1.45)	(0.42)	(0.87)	(0.19)	(0.97)	(0.87)
COGS			-0.04	-0.06	-0.06	-0.13	-0.08
			(-1.23)	(-0.87)	(-0.52)	(-0.71)	(-0.60)
RC			-0.00	-0.01	0.00	-0.02	-0.03
			(-1.09)	(-1.22)	(0.28)	(-1.13)	(-1.07)
STKCO			-0.00	-0.03	-0.01	-0.01	-0.01
			(-0.92)	(-1.06)	(-0.93)	(-0.97)	(-0.84)
SG&A			0.00	0.05	-0.09	0.02	-0.08
			(0.13)	(0.56)	(-1.32)	(0.22)	(-0.65)
R&D			0.00	0.01	0.01	0.01	0.00
			(1.04)	(1.02)	(1.09)	(1.07)	(0.96)
Topics x ES							
Customer Service			-0.14	-0.28	0.53	0.64	0.25
			(-0.36)	(-0.57)	(0.47)	(0.50)	(0.21)
Career Opportunity			0.14	0.07	0.27	0.25	0.14
			(0.37)	(0.12)	(0.19)	(0.16)	(0.09)
Organisational Strategy			0.10	-0.31	0.63	1.33	-0.05
			(0.19)	(-0.61)	(0.48)	(0.81)	(-0.02)
Compensation and Benefits			0.12	-0.00	0.45	0.31	0.11
			(0.33)	(-0.00)	(0.37)	(0.22)	(0.07)
Work Environment			-0.06	0.12	1.23	0.74	0.76
			(-0.16)	(0.25)	(0.94)	(0.54)	(0.51)
Communication			-0.25	-0.29	0.69	0.63	1.14
			(-0.77)	(-0.46)	(0.47)	(0.40)	(0.71)
Team Dynamics			0.21	-0.06	0.89	0.73	0.09
			(0.42)	(-0.11)	(0.65)	(0.51)	(0.06)
Leadership			0.06	0.17	0.65	0.91	-0.08
			(0.12)	(0.29)	(0.50)	(0.66)	(-0.06)
Management			0.25	0.20	-0.28	-0.26	-0.78
			(0.72)	(0.38)	(-0.46)	(-0.33)	(-0.93)
Working Hours			-0.03	-0.32	0.49	0.80	0.16
			(-0.08)	(-0.46)	(0.34)	(0.55)	(0.12)
Job Security			-0.20	0.02	0.75	0.86	0.61
			(-0.62)	(0.04)	(0.62)	(0.61)	(0.42)
Work-Life Balance			-0.82*	-1.35	-1.33	-1.77	-0.51
			(-1.76)	(-1.38)	(-0.65)	(-0.63)	(-0.25)
Technology			-0.02	-0.08	0.65	1.36	1.58
			(-0.06)	(-0.14)	(0.54)	(0.97)	(1.00)
Observations	42,773	42,773	42,773	42,773	42,773	42,773	42,773
R-squared	0.01	0.05	0.11	0.11	0.11	0.11	0.11
Num of groups	111	111	111	111	111	111	111
Pseudo-R2(%)	0.007	0.0534	0.108	0.108	0.107	0.108	0.108

Notes: This table reports the results of Fama-MacBeth regressions for 1, 3, 6, 9, and 12 months-ahead excess return forecasts using BERT (Panel A), LM (Panel B), the overall star rating (Panel C) as employee sentiment measures. The topics are a set of dummy variables at the review level, in the regressions, they are aggregated into topic weights and interacted with the ES for each company in each month. T-statistics with Newey-West adjustments are provided in parentheses. The asterisks *, **, and ***, respectively, indicate the significance at the 10%, 5%, and 1% levels.

Table 4.13: Distribution of Employee-related Costs (%) per Industry

Industry	Total Expenses	COGS	RC	STKCO	SG&A	R&D
Energy	100.00	83.11	1.99	2.28	2.07	10.55
Materials	100.00	77.86	1.72	0.83	4.06	15.54
Industrials	100.00	76.38	1.18	0.84	4.03	17.56
Consumer Discretionary	100.00	69.18	1.10	0.84	2.01	26.87
Consumer Staples	100.00	68.13	1.13	0.70	3.37	26.67
Health Care	100.00	52.69	1.38	1.78	10.29	33.87
Financials	100.00	70.78	1.22	3.39	0.56	24.05
Information Technology	100.00	48.41	1.50	3.31	12.43	34.35
Communication Services	100.00	55.50	1.20	2.94	9.12	31.24
Utilities	100.00	96.60	1.77	1.53	0.00	0.10
Real Estate	100.00	83.98	2.54	2.27	0.02	11.18

Notes: The table reports the employee-related costs for the 11 Global Industry Classification Standard (GICS) industries. It includes the total expenses (the sum of the following components), cost of goods sold (COGS), restructuring cost (RC), stock compensation (STKCO), selling, general, and administrative (SG&A) expenses and research and development (R&D) expenses as percentages of the total expenses for each industry.

Table 4.14: Effects of BERT Employee Sentiment and Industry-specific Costs on Long-short Portfolios

		Value-weighted portfolios High-Low		Equal-weighted portfolios High-Low	
		Average excess return	4-Factor alpha	Average excess return	4-Factor alpha
COGS	low	0.43***	0.30	0.25	0.30*
		(2.76)	(1.57)	(1.40)	(1.68)
	high	-0.25	-0.25	-0.50	-0.43**
		(-1.18)	(-1.18)	(-2.32)	(-2.35)
RC	low	0.30	0.25*	0.15	0.18
		(1.86)	(1.78)	(0.73)	(1.01)
	high	-0.08	0.06	-0.05	0.15
		(-0.27)	(0.22)	(-0.17)	(0.89)
STKCO	low	0.16	0.18	0.21	0.26
		(1.08)	(1.08)	(1.65)	(1.35)
	high	0.42	0.39*	-0.09	0.02
		(1.70)	(1.69)	(-0.44)	(0.10)
SG&A	low	-0.13	-0.17	-0.21	-0.23
		(-0.77)	(-1.11)	(-1.10)	(-1.31)
	high	0.49	0.54**	0.39	0.51***
		(2.36)	(2.33)	(1.87)	(2.69)
R&D	low	0.09	0.17	-0.12	-0.04
		(0.48)	(0.74)	(-0.54)	(-0.21)
	high	0.38	0.10	0.38***	0.44***
		(1.74)	(0.54)	(2.94)	(3.03)

Notes: This table reports the average returns and Fama-French-Carhart (FFC) four-factor estimated alphas for tercile portfolios sorted by employee satisfaction and industry-specific employee costs. Both value-weighted and equal-weighted long-short portfolios are presented in the table. The stocks in each portfolio are categorised into a low and a high group based on industry-level values of Cost of Goods Sold (COGS), Restructuring Costs (RC), Stock Compensation Expenses (STKCO), Selling, General and Administrative Expenses (SG&A), and Research and Development Expenses (R&D), and then sorted into tercile portfolios based on employee satisfaction scores measured by ES_BERT. T-statistics with Newey-West adjustments are provided in parentheses. The asterisks *, **, and ***, respectively, indicate the significance at the 10%, 5%, and 1% levels.

Table 4.15: Fama-Macbeth Regression for Return Predictability Across Industries

GICS Code	10	15	20	25	30	35	40	45	50	55	60
ES_BERT	0.45	-1.34	0.06	0.03	3.39	0.14	0.49*	0.22	0.40	-1.12	-0.02
	(0.62)	(-0.53)	(0.20)	(0.16)	(1.40)	(0.65)	(1.90)	(0.86)	(0.96)	(-0.53)	(-1.03)
log(BM)	0.94	-4.90	-0.13	-0.02	-0.28	-0.23	0.09*	-0.28*	0.07	2.00*	-0.04
	(1.24)	(-1.02)	(-0.45)	(-0.15)	(-0.62)	(-1.40)	(1.80)	(-1.87)	(0.40)	(1.92)	(-1.31)
log(Size)	0.20	3.18	-0.20	-0.06	-0.36	-0.01	0.07	-0.39***	0.00	1.76**	0.10**
	(1.39)	(0.89)	(-0.71)	(-0.29)	(-0.93)	(-0.04)	(1.22)	(-3.07)	(0.01)	(2.27)	(2.18)
COGS	-0.55	0.46	-0.01	-0.15	-1.27	-0.16	0.09	-0.05	-0.02	-3.52**	-0.08
	(-1.09)	(0.81)	(-0.08)	(-1.32)	(-1.34)	(-1.53)	(1.08)	(-0.60)	(-0.09)	(-2.33)	(-0.94)
RC	0.02	0.59	0.00	-0.00	0.04	0.00	-0.03	-0.00	-0.31	-0.18	0.11
	(0.19)	(1.07)	(0.64)	(-1.59)	(1.07)	(0.54)	(-0.48)	(-0.05)	(-1.52)	(-0.76)	(1.49)
STKCO	0.02	0.22	0.01	0.00	0.07	-0.00	0.01	0.00	0.05	-0.08	-0.04
	(0.81)	(0.53)	(0.59)	(0.01)	(0.78)	(-0.09)	(0.09)	(0.14)	(0.31)	(-0.63)	(-0.48)
SG&A	0.13	-1.65	0.06	0.09	-0.49	0.21	-0.03	0.40**	-0.06	0.55*	-0.22
	(0.12)	(-0.92)	(0.50)	(0.72)	(-0.79)	(1.01)	(-1.10)	(2.48)	(-0.98)	(1.77)	(-1.54)
R&D	-0.01	0.02	-0.00	-0.01	0.01*	-0.00	0.03	0.00	0.01	0.16	-0.02
	(-0.73)	(0.41)	(-0.44)	(-1.00)	(1.96)	(-0.90)	(0.42)	(1.42)	(0.47)	(0.92)	(-1.04)
Observations	581	1,438	3,120	3,658	1,653	4,078	234	4,281	401	1,640	216
R-squared	0.97	0.66	0.33	0.27	0.59	0.26	0.99	0.25	0.98	0.52	0.99
Num of groups	106	106	108	110	106	107	103	110	108	106	106
Pseudo-R2(%)	0.329	0.327	0.114	0.114	0.298	0.134	0.137	0.136	0.577	0.385	0.416

Notes: This table reports the results of Fama-MacBeth regressions of one month-ahead excess return forecasts for each of the 11 Global Industry Classification Standard (GICS) industries. The GICS classification is as follows. 10: Energy, 15: Materials, 20: Industrials, 25: Consumer Discretionary, 30: Consumer Staples, 35: Health Care, 40: Financials, 45: Information Technology, 50: Communication Services, 55: Utilities, 60: Real Estate. The regression is controlled for employee sentiment as measured by BERT, the book-to-market ratio, company size, and employee-related cost including Cost of Goods Sold (COGS), Restructuring Costs (RC), Stock Compensation Expenses (STKCO), Selling, General and Administrative Expenses (SG&A), and Research and Development (R&D) Expenses. T-statistics with Newey-West adjustments are provided in parentheses. The asterisks *, **, and ***, respectively, indicate the significance at the 10%, 5%, and 1% levels.

C Appendices for Chapter 4

C.1 Variable Description

Table C16: Variable Description

Panel A: Employee Reviews Characteristics	
Employee Ratings	The ratings consist of an overall evaluation and five subcategories: Work-Life Balance, Culture & Values, Senior Leadership, Career Opportunities, Compensation & Benefits. These values are in the range of 1 to 5 stars given by the employees, where 5 is the highest rating and 1 is the lowest. We aggregate the ratings on a monthly basis for each company.
Recommend To Friend	A dummy variable with the value 1 if the reviewer is willing to recommend the company to a friend, 0 otherwise.
CEO Approval	A dummy variable with the value 1 if the reviewer approves of the company's CEO, 0 if they have no opinion, and -1 if they disapprove.
Business Outlook	A dummy variable with the value 1 if the reviewer considers the business outlook of the company to be positive, 0 neutral, and -1 negative.
Employment Length	The length of the reviewer's employment in ranges of 0, 1, 5, 8, 10, or more years.
Current Employee	A dummy variable with the value 1 if the reviewer is a current employee, 0 otherwise.
Full-time Employee	A dummy variable with the value 1 if the reviewer is a full-time employee, 0 otherwise.

Table C16: Variable description - Continued

Panel B: Firm-level Characteristics	
β^{MKT}	The market beta is estimated by regressing the excess returns of individual stocks on the value-weighted market excess return.
Size	Measured by the market equity at the end of the previous fiscal year.
BM	The book-to-market ratio measured at the end of the previous fiscal year.
ROA	The return on asset ratio is calculated as the net income over total assets at the end of the previous fiscal year.
ROE	The return on equity ratio is calculated as the net income over shareholders' equity at the end of the previous fiscal year.
Age	The age of a company is defined as the year when the company's data became available in Compustat and CRSP datasets.
Illiquidity	The Amihud illiquidity measure is calculated on a monthly basis by dividing the absolute monthly return by the average monthly trading volume.
Turnover	The turnover is calculated as trading volume over average shares outstanding.
COGS	The cost of goods sold represents the costs directly associated with the operation, including the cost of materials, labour, and other expenses directly related to the production or acquisition of the goods.
SG&A	Selling, general and administrative expenses represent indirect expenses incurred by a company to support its overall operations, such as marketing, salaries, rent, utilities, and other overhead costs.
R&D	Research and development expenses represent the costs incurred during the year that relate to the development of new products or services, including expenses related to research, design, testing, and experimentation.
STKCO	Stock compensation expenses represent compensation in the form of company stock on a pre-tax basis, including stock bonuses, deferred compensation, amortisation of deferred compensation, and non-cash compensation expense.
RC	Restructuring costs represent the pretax expenses incurred by a company when it undergoes significant changes to its operations, including costs associated with employee severance, asset impairments, facility closures, and other restructuring activities.

C.2 Sentiment Analysis Accuracy of LM and BERT

Table C17: Accuracy of LM and BERT

Panel A: Topic level accuracy			
Topic Label	Num of Reviews	LM Acc.	BERT Acc.
Customer Service	179,264	60.19	88.00
Career Opportunity	148,600	72.71	95.22
Organisational Strategy	71,688	62.80	90.34
Compensation and Benefits	217,806	61.04	92.37
Work Environment	68,928	62.83	92.96
Communication	97,379	61.81	84.19
Team Dynamics	69,117	63.91	95.93
Leadership	57,448	64.35	87.90
Management	107,151	61.99	85.83
Working Hours	84,660	58.56	90.40
Job Security	133,247	62.79	87.24
Work-Life Balance	114,412	59.53	95.05
Technology	79,754	57.78	90.97
Panel B: Industry level accuracy			
GICS Industry	Num of Reviews	LM Acc.	BERT Acc.
Energy	31,084	61.56	90.48
Materials	41,153	63.45	89.75
Industrials	164,885	63.76	89.92
Consumer Discretionary	314,667	61.99	89.08
Consumer Staples	93,150	62.62	88.99
Health Care	146,544	63.83	88.63
Financials	189,370	63.80	89.71
Information Technology	337,238	59.26	91.04
Communication Services	84,605	65.96	90.02
Utilities	12,216	68.96	91.93
Real Estate	14,542	70.05	90.95

Notes: This table reports the accuracy of sentiment analysis at the topic (Panel A) and industry (Panel B) levels.

C.3 Correlation Matrix of Variables

Table C18: Correlation Matrix

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
(1)	ES_BERT	1.00																			
(2)	ES_LM	0.27	1.00																		
(3)	ES_Rating	0.38	0.28	1.00																	
(4)	Overall Rating	0.41	0.33	0.84	1.00																
(5)	Work-Life Balance Rating	0.22	0.23	0.38	0.45	1.00															
(6)	Culture & Values Rating	0.18	0.12	0.33	0.43	0.15	1.00														
(7)	Senior Leadership Rating	0.27	0.29	0.51	0.62	0.71	0.28	1.00													
(8)	Career Opportunities Rating	0.25	0.25	0.48	0.59	0.62	0.31	0.76	1.00												
(9)	Compensation & Benefits Rating	0.21	0.21	0.30	0.40	0.62	0.24	0.63	0.67	1.00											
(10)	Recommend To Friend	0.37	0.32	0.74	0.79	0.46	0.37	0.62	0.57	0.41	1.00										
(11)	CEO Approval	0.27	0.27	0.51	0.58	0.37	0.27	0.55	0.47	0.35	0.59	1.00									
(12)	Business Outlook	0.23	0.23	0.42	0.53	0.25	0.50	0.42	0.40	0.29	0.51	0.46	1.00								
(13)	Employment Length	0.02	0.08	0.14	0.17	0.17	0.00	0.19	0.17	0.10	0.18	0.18	0.15	1.00							
(14)	Current Employee	0.02	0.00	0.05	0.07	-0.07	0.46	-0.05	0.02	0.06	0.06	0.01	0.10	-0.03	1.00						
(15)	Full-time Employee	0.05	0.00	0.10	0.15	-0.14	0.70	-0.07	0.00	0.01	0.10	0.08	0.24	-0.03	0.56	1.00					
(16)	COGS	-0.01	-0.03	0.04	0.01	-0.06	0.00	-0.03	0.04	-0.04	0.01	-0.04	-0.04	0.00	0.08	-0.02	1.00				
(17)	RC	0.01	0.02	-0.01	0.00	0.01	-0.03	0.03	0.00	0.02	0.01	0.03	0.02	0.00	-0.05	-0.03	-0.07	1.00			
(18)	STKCO	0.07	0.02	0.11	0.13	0.04	0.06	0.05	0.09	0.07	0.11	0.07	0.07	0.04	0.04	0.04	0.35	0.06	1.00		
(19)	SG&A	0.01	-0.04	0.06	0.03	0.01	-0.01	0.00	0.01	-0.03	0.02	-0.01	-0.03	0.00	-0.02	-0.05	0.29	-0.08	0.34	1.00	
(20)	R&D	0.05	0.02	0.11	0.12	0.06	0.02	0.04	0.07	0.07	0.09	0.05	0.04	0.05	0.04	0.02	0.27	-0.19	0.59	0.42	1.00

Notes: This table reports the correlation between the variables. The row header corresponds to the variable names in the first column.

C.4 Grouping Industry-specific Employee Costs

Table C19: Industry-specific Employee Costs

	GICS Code	COGS	GICS Code	RC	GICS Code	STKCO	GICS Code	SG&A	GICS Code	R&D
Low	45	52.81	45	-4.64	30	0.72	55	0.10	55	0.00
	35	53.87	15	-1.73	15	0.86	10	10.91	60	0.03
	50	56.91	35	-1.53	25	0.86	60	11.34	40	0.59
	30	69.57	10	-1.48	20	0.87	15	15.95	25	2.09
	25	70.54	20	-1.31	55	0.95	20	18.04	10	2.33
	40	71.71	30	-1.27	35	1.83	40	25.02	30	3.50
High	20	78.30	50	-1.03	60	2.31	30	27.48	20	4.11
	15	80.32	25	-1.03	10	2.43	25	27.54	15	4.61
	55	82.17	40	-0.79	50	2.98	50	31.97	50	9.18
	60	85.43	60	0.90	45	3.42	35	35.09	35	10.74
	10	85.82	55	16.78	40	3.47	45	35.58	45	12.83

Notes: The stocks in each portfolio are categorised into a low and a high group based on the values of Cost of Goods Sold (COGS), Restructuring Costs (RC), Stock Compensation Expenses (STKCO), Selling, General and Administrative Expenses (SG&A), and Research and Development Expenses (R&D) at the industry level. The Global Industry Classification Standard (GICS) classification is as follows. 10: Energy 15: Materials, 20: Industrials, 25: Consumer Discretionary, 30: Consumer Staples, 35: Health Care, 40: Financials, 45: Information Technology, 50: Communication Services, 55: Utilities, 60: Real Estate.

Chapter 5

Conclusion and Future Research

5.1 Summary and Conclusions

In summary, this thesis studies sentiment analysis in finance, with a particular emphasis on the analysis of Glassdoor employee reviews. It makes significant contributions to the current literature in several ways. Firstly, in Chapter 2 we provide a more comprehensive literature review of sentiment analysis in finance and propose a framework for classifying the most frequently-used sentiment analysis methods. The framework consists of the Pure Conceptual Approach, the Hybrid Approach, and the Pure Empirical Approach, depending on whether the sentiment is evaluated by human annotators, predefined rules, machine learning tools or the market itself.

We primarily focus on the Hybrid approach, which includes methods such as lexicon-based approaches, machine learning algorithms and pre-trained models. By empirically evaluating 31 sentiment analysis methods using text comments extracted from Glassdoor employee reviews. We compare their relative performance measured against the label provided by the user rating and identify optimal approaches, guiding future researchers in selecting the optimal approach for their work. To the best of our knowledge, our work is the first to assess sentiment analysis methods using this specific dataset. The findings indicate that BERT models consistently yield the best outcomes in both binary and three-class sentiment classification tasks. However, it should be noted that the BERT models are more complex and require a higher level of computational power.

On the other hand, machine learning methods, such as LR, SVM, and XBG classifiers, demonstrate robust predictive capabilities in sentiment analysis. However, the choice of word embedding techniques plays a crucial role in determining the performance of sentiment analysis models. For instance, we observe that contextual word embeddings such as Word2Vec and GloVe do not exhibit a clear advantage over traditional Bag-of-Words (BOW) and TF-IDF approaches. This lack of advantage is mainly attributed to the high dimensionality of contextual word embeddings. Our research also reveals that lexicon-based approaches perform the worst among the methods considered. These findings have significant implications for previous research, as lexicon-based approaches have been extensively used in sentiment analysis. While machine learning methods demonstrate promising results, it is essential not to underestimate the significance of the dimensionality and efficiency of word embeddings.

In Chapter 2, we also assess the impact of task complexity and label ambiguity on the performance of sentiment analysis models. The results reveal a decline in model performance as the complexity of the task and the ambiguity of the labels increase. Specifically, all methods perform well in identifying positive sentiment from the *pros* section and negative sentiment from the *cons* section of reviews. However, when the overall star rating of the reviews is used as the expected sentiment, the performance of all methods decreases. This decline is particularly evident in three-star ratings, which tend to contain more ambiguous expressions of sentiment.

Secondly, our exploration of multilingual sentiment analysis in Chapter 3 adds a valuable dimension to the field. By examining the challenges and implications of analysing sentiment across multiple languages, we acknowledge the global nature of finance and highlight the importance of considering language diversity when analysing sentiment in financial texts. In Chapter 3, we explore multilingual sentiment analysis using three pre-trained NLP models: DistilmBERT, mBERT, and

XLM-R. The dataset consists of Glassdoor employee reviews in German, French, Portuguese, and Spanish, as well as their translations into English. The overall accuracy of all models exceeds 94%, but a statistically significant decline of 0.8% to 1.5% is observed when using translated texts, a finding consistent with foreign-to-English translations.

Regarding specific languages, the DistilmBERT model demonstrates a 0.7% improvement in accuracy after translating Portuguese to English. The model faced difficulties in identifying positive sentiment initially, leading to low recall. However, translation enhanced the detection of positive sentiment. In the case of French-English translation, BERT exhibited lower recall and failed to identify false negatives, while XLM-R demonstrated lower specificity and failed to identify false positives.

We also demonstrate the zero-shot capability of DistilmBERT, mBERT, and XLM-R. These cross-lingual models are valuable when dealing with languages that lack labelled data or have limited resources. Zero-shot learning, a well-established technique, allows knowledge transfer between languages, enabling practical applications for processing multilingual texts. Our results confirm that even when fine-tuned using only English reviews, these models still provide relatively accurate predictions for sentiment in German, French, Portuguese, and Spanish reviews. The study also explores the similarity between English and other languages and the vocabulary overlaps of the models, finding that zero-shot transfer of mBERT and XLM-R correlates highly with syntactic language similarity rather than vocabulary overlaps.

Finally, in Chapter 4, we present a novel application of BERT for sentiment analysis on employee reviews and investigate the relationship between employee sentiment and stock returns. This is the first work that applies BERT to the intersection of sentiment analysis and financial markets using Glassdoor data. We also use the overall star rating and LM dictionary on text comments to measure employee sentiment. We sort portfolios by employee sentiment measured by BERT, LM and star ratings, and show that portfolios with medium to high sentiment tend to outperform the market benchmark according to BERT and star ratings, while the low sentiment portfolio outperforms based on LM.

Our portfolio analysis also reports the value-added nature of BERT in investment decision-making. We demonstrate that BERT outperforms LM in high employee sentiment portfolios, highlighting the advantages of using a superior sentiment measure. However, when employee sentiment is at a low to moderate level, either using BERT or LM to assess the text seems to be more empirically informative than relying on overall ratings alone. We also examine topic-sentiment interactions for stock prediction and find topics such as customer service, work environment, and working hours expressed in positive sentiment using BERT can predict increases in stock returns. However, LM and overall ratings do not share meaningful topic-sentiment relations or substantial evidence for predicting stock returns. Finally, this chapter investigates the impact of employee-related costs on both employee satisfaction and stock returns, revealing the significant combined influence of employee satisfaction and employee-related costs in industries with lower production costs but higher operational and management expenses.

Overall, this thesis highlights the significance of using text comments to assess employee sentiment, challenging the conventional reliance on overall star ratings. By emphasising the depth and specificity offered by text comments, we advocate for a more nuanced approach to understanding and evaluating employee sentiment. This research contributes to the growing body of work that recognises the value of qualitative textual data in sentiment analysis and its implications for evaluating corporate performance and employee satisfaction.

5.2 Future Research

While this thesis has many substantial contributions to the field of sentiment analysis in Glassdoor employee reviews, it is important to acknowledge the limitations of our research. One limitation is that our work is based on sentiment analysis in the context of Glassdoor employee reviews. While Glassdoor provides a valuable source of employee feedback, it is essential to consider that employees' sentiments and experiences may extend beyond what is expressed on this platform. Future research could explore sentiment analysis in other sources, such as customer feedback, financial reports, and employee surveys, to gain a more comprehensive understanding of the dynamics between employee satisfaction and corporate performance. It would also be beneficial to investigate alternative types of sentiment indicators that could provide valuable insights beyond employee sentiment. For instance, examining sentiment derived from earnings conference calls, financial reports or other financial documents could offer a different perspective on sentiment within a company. This could help provide a more comprehensive view of sentiment within an organisation and its potential impact on various aspects of corporate performance.

In Chapter 4, we empirically study employee sentiment in relation to stock returns, future work can expand the application of sentiment indicators to predict variables beyond stock returns. One such variable of interest is Environmental, Social, and Governance (ESG) ratings. Investigating the relationship between sentiment and ESG ratings could shed light on how sentiment within an organisation relates to its broader sustainability and ethical practices. Future researchers could also consider continuously monitoring employee sentiment within the company. The real-time sentiment monitoring systems can offer more up-to-date and regular ESG performance feedback for investors and stakeholders interested in responsible and sustainable investing.

Another area of future research in the field of sentiment analysis is the empirical application of multilingual sentiment analysis in international business and finance. Researchers can expand the scope of analysis to include more languages than those explored in this thesis, thereby capturing a broader range of linguistic and cultural nuances. This empirical exploration of multilingual sentiment analysis can offer valuable insights into the complex interplay between employee sentiment, cultural factors, and financial performance in the global business environment.

Furthermore, while we demonstrate the effectiveness of various NLP models, including lexicon-based approaches and more advanced models such as BERT, there have been rapid developments in sentiment analysis models since BERT. Numerous large language models such as the Generative Pre-trained Transformer 3 (GPT-3) (Brown et al. 2020) and GPT-4 (OpenAI 2023) have emerged, each with its own strengths and limitations. Future research can empirically apply these more recent and larger models to stay up-to-date with the latest advancements in the field. In addition, our research relies on text-based sentiment analysis methods but there may be other dimensions of sentiment, such as tone of voice that are not captured in our analysis. Exploring multi-modal sentiment analysis techniques that incorporate audio or video data could provide a more holistic understanding of employee sentiment and enrich the information gained from textual analysis.

Lastly, as sentiment analysis continues to evolve, future research should address the ethical implications associated with its use in employee reviews. The considerations of data privacy, consent, and potential biases in sentiment analysis algorithms demand careful examination to ensure responsible practices. It is also crucial to be mindful of the potential for market manipulation if sentiment measures become a standard element in valuing companies. There may be an incentive for malicious actors to manipulate sentiment scores, such as by submitting fake reviews, with the intention of influencing market reactions and trading on those reactions. To ensure the responsible and fair use of sentiment analysis methodologies in corporate and financial contexts, future studies should not only explore the ethical dimensions but also develop robust frameworks and guidelines to address these challenges effectively. By doing so, we can foster a more transparent and trustworthy application of sentiment analysis, while guarding against potential biases, market manipulation, and the diminishing predictive power of sentiment-based indicators in investment decisions.

References

Abbasi, A., Li, J., Adjeroh, D., Abate, M. & Zheng, W. (2019), 'Don't mention it? Analyzing user-generated content signals for early adverse event warnings', *Information Systems Research* **30**(3), 1007–1028.

Ahuja, K., Kumar, S., Dandapat, S. & Choudhury, M. (2022), Multi task learning for zero shot performance prediction of multilingual models, *in* 'Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Dublin, Ireland, pp. 5454–5467.

Amihud, Y. (2002), 'Illiquidity and stock returns: cross-section and time-series effects', *Journal of Financial Markets* **5**(1), 31–56.

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

Araci, D. (2019), 'FinBERT: Financial sentiment analysis with pre-trained language models', *arXiv preprint arXiv:1908.10063* .

Artetxe, M., Ruder, S. & Yogatama, D. (2020), On the cross-lingual transferability of monolingual representations, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics.

Au, S., Dong, M. & Tremblay, A. (2021), 'Employee flexibility, exogenous risk, and firm value', *Journal of Financial and Quantitative Analysis* **56**(3), 853–884.

Baccianella, S., Esuli, A. & Sebastiani, F. (2010), Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, *in* 'Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)'.

Bingler, J. A., Kraus, M., Leippold, M. & Webersinke, N. (2022), 'Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures', *Finance Research Letters* **47**, 102776.

Bochkay, K., Brown, S. V., Leone, A. J. & Tucker, J. W. (2023), 'Textual analysis in accounting: What's next?', *Contemporary Accounting Research* **40**(2), 765–805.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017), 'Enriching word vectors with subword information', *Transactions of the Association for Computational Linguistics* **5**, 135–146.

Boughorbel, S., Jarray, F. & El-Anbari, M. (2017), 'Optimal classifier for imbalanced data using matthews correlation coefficient metric', *PloS One* **12**(6), e0177678.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Brown, G. W. & Cliff, M. T. (2004), 'Investor sentiment and the near-term stock market', *Journal of empirical finance* **11**(1), 1–27.

Brown, G. W. & Cliff, M. T. (2005), 'Investor sentiment and asset valuation', *The Journal of Business* **78**(2), 405–440.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020), 'Language models are few-shot learners', *Advances in neural information processing systems* **33**, 1877–1901.

Campbell, D. W. & Shang, R. (2022), 'Tone at the bottom: Measuring corporate misconduct risk from the text of employee reviews', *Management Science* **68**(9), 7034–7053.

Canning, E. A., Murphy, M. C., Emerson, K. T., Chatman, J. A., Dweck, C. S. & Kray, L. J. (2020), 'Cultures of genius at work: Organizational mindsets predict cultural norms, trust, and commitment', *Personality and Social Psychology Bulletin* **46**(4), 626–642.

Chen, J., Tang, G., Yao, J. & Zhou, G. (2023), 'Employee sentiment and stock returns', *Journal of Economic Dynamics and Control* **149**, 104636.

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining', pp. 785–794.

Chicco, D. & Jurman, G. (2020), 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics* **21**(1), 6.

Chicco, D., Tötsch, N. & Jurman, G. (2021), 'The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation', *BioData Mining* **14**(1), 1–22.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2020), Unsupervised cross-lingual representation learning at scale, *in* 'Association for Computational Linguistics'.

Corritore, M., Goldberg, A. & Srivastava, S. B. (2020), 'Duality in diversity: How intrapersonal and interpersonal cultural heterogeneity relate to firm performance', *Administrative Science Quarterly* **65**(2), 359–394.

Cortes, C. & Vapnik, V. (1995), ‘Support-vector networks’, *Machine Learning* **20**(3), 273–297.

Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S. & Davis, B. (2017), SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news, *in* ‘Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)’, Association for Computational Linguistics, Vancouver, Canada, pp. 519–535.

Creek, S. A., Kuhn, K. M. & Sahaym, A. (2019), ‘Board diversity and employee satisfaction: The mediating role of progressive programs’, *Group & Organization Management* **44**(3), 521–548.

Dabirian, A., Kietzmann, J. & Diba, H. (2017), ‘A great place to work!? Understanding crowdsourced employer branding’, *Business Horizons* **60**(2), 197–205.

Dabirian, A., Paschen, J. & Kietzmann, J. (2019), ‘Employer branding: Understanding employer attractiveness of it companies’, *IT Professional* **21**(1), 82–89.

Dai, A. M. & Le, Q. V. (2015), ‘Semi-supervised sequence learning’, *Advances in Neural Information Processing Systems* **28**.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018), ‘BERT: Pre-training of deep bidirectional transformers for language understanding’, *arXiv preprint arXiv:1810.04805* .

Edmans, A. (2011), ‘Does the stock market fully value intangibles? Employee satisfaction and equity prices’, *Journal of Financial Economics* **101**(3), 621–640.

Engelberg, J. E., Reed, A. V. & Ringgenberg, M. C. (2012), ‘How are shorts informed?: Short sellers, news, and information processing’, *Journal of Financial Economics* **105**(2), 260–278.

Esuli, A. & Sebastiani, F. (2006), Sentiwordnet: A publicly available lexical resource for opinion mining, *in* 'Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)'.

Feldman, R., Govindaraj, S., Livnat, J. & Segal, B. (2010), 'Management's tone change, post earnings announcement drift and accruals', *Review of Accounting Studies* **15**(4), 915–953.

Feuerriegel, S. & Gordon, J. (2019), 'News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions', *European Journal of Operational Research* **272**(1), 162–175.

Frankel, R., Jennings, J. & Lee, J. (2022), 'Disclosure Sentiment: Machine Learning vs. Dictionary Methods', *Management Science* **68**(7), 5514–5532.

Godbole, N., Srinivasaiah, M. & Skiena, S. (2007), 'Large-scale sentiment analysis for news and blogs.', *The International AAAI Conference on Web and Social Media (ICWSM)* **7**(21), 219–222.

González-Carvajal, S. & Garrido-Merchán, E. C. (2020), 'Comparing BERT against traditional machine learning text classification', *arXiv preprint arXiv:2005.13012* .

Green, T. C., Huang, R., Wen, Q. & Zhou, D. (2019), 'Crowdsourced employer reviews and stock returns', *Journal of Financial Economics* **134**(1), 236–251.

Hagenau, M., Liebmann, M. & Neumann, D. (2013), 'Automated news reading: Stock price prediction based on financial news using context-capturing features', *Decision Support Systems* **55**(3), 685–697.

Hales, J., Moon Jr, J. R. & Swenson, L. A. (2018), 'A new era of voluntary disclosure? empirical evidence on how employee postings on social media relate to future corporate disclosures', *Accounting, Organizations and Society* **68**, 88–108.

Hickman, L., Thapa, S., Tay, L., Cao, M. & Srinivasan, P. (2022), 'Text preprocessing for text mining in organizational research: Review and recommendations', *Organizational Research Methods* **25**(1), 114–146.

Huang, A. H., Wang, H. & Yang, Y. (2023), 'FinBERT: A large language model for extracting information from financial text', *Contemporary Accounting Research* **40**(2), 806–841.

Huang, A. H., Zang, A. Y. & Zheng, R. (2014), 'Evidence on the information content of text in analyst reports', *The Accounting Review* **89**(6), 2151–2180.

Huang, K., Li, M. & Markov, S. (2020), 'What do employees know? Evidence from a social media platform', *The Accounting Review* **95**(2), 199–226.

Huang, M., Li, P., Meschke, F. & Guthrie, J. P. (2015), 'Family firms, employee satisfaction, and corporate performance', *Journal of Corporate Finance* **34**, 108–127.

Hutto, C. & Gilbert, E. (2014), VADER: A parsimonious rule-based model for sentiment analysis of social media text, *in* 'Proceedings of the International AAAI Conference on Web and Social Media', Vol. 8, pp. 216–225.

Jaggi, M., Mandal, P., Narang, S., Naseem, U. & Khushi, M. (2021), 'Text mining of stocktwits data for predicting stock prices', *Applied System Innovation* **4**(1), 13.

Jing, C., Keasey, K., Lim, I. & Xu, B. (2019), 'Financial constraints and employee satisfaction', *Economics Letters* **183**, 108599.

Karthikeyan, K., Wang, Z., Mayhew, S. & Roth, D. (2020), Cross-lingual ability of multilingual BERT: An empirical study, *in* 'Proceedings of International Conference on Learning Representations 2020'.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihok, G. & Den Hartog, D. N. (2018), 'Text classification for organizational researchers: A tutorial', *Organizational Research Methods* **21**(3), 766–799.

Kothari, S. P., Li, X. & Short, J. E. (2009), 'The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis', *The Accounting Review* **84**(5), 1639–1670.

Kriebel, J. & Stitz, L. (2021), 'Credit default prediction from user-generated text in peer-to-peer lending using deep learning.', *European Journal of Operational Research* **302**(1), 309–323.

Lample, G. & Conneau, A. (2019), 'Cross-lingual language model pretraining', *arXiv preprint arXiv:1901.07291* .

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. & Soricut, R. (2019), 'ALBERT: A Lite BERT for self-supervised learning of language representations', *arXiv preprint arXiv:1909.11942* .

Lauscher, A., Ravishankar, V., Vulic, I. & Glavas, G. (2020), From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers, *in* 'Proceedings of Empirical methods in natural language processing 2020'.

Leippold, M. (2023), 'Sentiment spin: Attacking financial sentiment with gpt-3', *Finance Research Letters* **55**, 103957.

Li, F. (2010), 'The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach', *Journal of Accounting Research* **48**(5), 1049–1102.

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C. & Levin, L. (2017), URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors, *in* 'Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers', Association for Computational Linguistics, Valencia, Spain, pp. 8–14.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019), 'Roberta: A robustly optimized BERT pretraining approach', *arXiv preprint arXiv:1907.11692* .

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

Lommel, A., Uszkoreit, H. & Burchardt, A. (2014), 'Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics', *Tradumàtica* (12), 0455–463.

Loughran, T. & McDonald, B. (2011), 'When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks', *The Journal of Finance* **66**(1), 35–65.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J. & Takala, P. (2014), 'Good debt or bad debt: Detecting semantic orientations in economic texts', *Journal of the Association for Information Science and Technology* **65**(4), 782–796.

Matthews, B. W. (1975), 'Comparison of the predicted and observed secondary structure of t4 phage lysozyme', *Biochimica et Biophysica Acta (BBA)-Protein Structure* **405**(2), 442–451.

Melián-González, S., Bulchand-Gidumal, J. & González López-Valcárcel, B. (2015), 'New evidence of the relationship between employee satisfaction and firm economic performance', *Personnel Review* **44**(6), 906–929.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781* .

Miller, G. A. (1995), 'WordNet: A lexical database for english', *Communications of the ACM* **38**(11), 39–41.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T. & Trajanov, D. (2020), 'Evaluation of sentiment analysis in finance: from lexicons to transformers', *IEEE Access* **8**, 131662–131682.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y. & Ngo, D. C. L. (2014), 'Text mining for market prediction: A systematic review', *Expert Systems with Applications* **41**(16), 7653–7670.

Newey, W. K. & West, K. D. (1987), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix', *Econometrica* **55**(3), 703–708.

Ni, J., Jin, Z., Wang, Q., Sachan, M. & Leippold, M. (2023), 'When does aggregating multiple skills with multi-task learning work? a case study in financial nlp'.

OpenAI (2023), 'GPT-4 technical report', *arXiv preprint arXiv:2303.08774* .

Pandey, S. & Pandey, S. K. (2019), 'Applying natural language processing capabilities in computerized textual analysis to measure organizational culture', *Organizational Research Methods* **22**(3), 765–797.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002), Bleu: A method for automatic evaluation of machine translation, *in* 'Proceedings of the 40th Annual Meeting on Association for Computational Linguistics', ACL '02, Association for Computational Linguistics, USA, p. 311–318.

Pennington, J., Socher, R. & Manning, C. D. (2014), GloVe: Global vectors for word representation, *in* 'Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)', pp. 1532–1543.

Peters, M. E., Ammar, W., Bhagavatula, C. & Power, R. (2017), Semi-supervised sequence tagging with bidirectional language models, *in* 'Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Vancouver, Canada, pp. 1756–1765.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018), Deep contextualized word representations, *in* 'Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', Association for Computational Linguistics, New Orleans, Louisiana, pp. 2227–2237.

Pires, T., Schlinger, E. & Garrette, D. (2019), How multilingual is multilingual BERT?, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 4996–5001.

Poncelas, A., Lohar, P., Way, A. & Hadley, J. (2020), The impact of indirect machine translation on sentiment classification, *in* 'Association for Machine Translation in the Americas (AMTA)'.

Price, S. M., Doran, J. S., Peterson, D. R. & Bliss, B. A. (2012), 'Earnings conference calls and stock returns: The incremental informativeness of textual tone', *Journal of Banking & Finance* **36**(4), 992–1011.

Quinlan, J. R. (1986), 'Induction of decision trees', *Machine Learning* **1**(1), 81–106.

Ramachandran, P., Liu, P. & Le, Q. (2017), Unsupervised pretraining for sequence to sequence learning, *in* 'Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Copenhagen, Denmark, pp. 383–391.

Renault, T. (2020), 'Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages', *Digital Finance* **2**(1), 1–13.

Rezaee, K., Loureiro, D., Camacho-Collados, J. & Pilehvar, M. T. (2021), On the cross-lingual transferability of contextualized sense embeddings, *in* 'Proceedings of the 1st Workshop on Multilingual Representation Learning', Association for Computational Linguistics, Punta Cana, Dominican Republic, pp. 107–115.

Rish, I. et al. (2001), An empirical study of the naive bayes classifier, *in* 'IJCAI 2001 workshop on empirical methods in artificial intelligence', Vol. 3, pp. 41–46.

Robertson, J., Lord Ferguson, S., Eriksson, T. & Näppä, A. (2019), 'The brand personality dimensions of business-to-business firms: a content analysis of employer reviews on social media', *Journal of Business-to-business Marketing* **26**(2), 109–124.

Rogers, J. L., Van Buskirk, A. & Zechman, S. L. (2011), 'Disclosure tone and shareholder litigation', *The Accounting Review* **86**(6), 2155–2183.

Sanh, V., Debut, L., Chaumond, J. & Wolf, T. (2019), 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter', *arXiv preprint arXiv:1910.01108* .

Schmiedel, T., M¨uller, O. & vom Brocke, J. (2019), 'Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture', *Organizational Research Methods* **22**(4), 941–968.

Sennrich, R., Haddow, B. & Birch, A. (2016), Neural machine translation of rare words with subword units, *in* 'Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Berlin, Germany, pp. 1715–1725.

Sohangir, S., Petty, N. & Wang, D. (2018), Financial sentiment lexicon analysis, *in* '2018 IEEE 12th International Conference on Semantic Computing (ICSC)', IEEE, pp. 286–289.

Sousa, M. G., Sakiyama, K., de Souza Rodrigues, L., Moraes, P. H., Fernandes, E. R. & Matsubara, E. T. (2019), BERT for stock market sentiment analysis, *in* '2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)', IEEE, pp. 1597–1601.

Stamolampros, P., Korfiatis, N., Chalvatzis, K. & Buhalis, D. (2019), 'Job satisfaction and employee turnover determinants in high contact services: Insights from employees' online reviews', *Tourism Management* **75**, 130–147.

Stevenson, M., Mues, C. & Bravo, C. (2021), 'The value of text for small business default prediction: A deep learning approach', *European Journal of Operational Research* **295**(2), 758–771.

Stone, P. J. & Hunt, E. B. (1968), 'The general inquirer: A computer approach to content analysis.', *American Journal of Sociology* **73**(5), 634–635.

Storer, A. & Reich, A. (2021), ''Losing My Raise': minimum wage increases, status loss and job satisfaction among low-wage employees', *Socio-Economic Review* **19**(2), 681–709.

Sull, D., Sull, C. & Zweig, B. (2022), 'Toxic culture is driving the great resignation', *MIT Sloan Management Review* **63**(2), 1–9.

Symitsi, E., Stamolampros, P. & Daskalakis, G. (2018), 'Employees' online reviews and equity prices', *Economics Letters* **162**, 53–55.

Symitsi, E., Stamolampros, P., Daskalakis, G. & Korfiatis, N. (2021), 'The informational value of employee online reviews', *European Journal of Operational Research* **288**(2), 605–619.

Tambe, P., Ye, X. & Cappelli, P. (2020), 'Paying to program? Engineering brand and high-tech wages', *Management Science* **66**(7), 3010–3028.

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

Tetlock, P. C., Saar-Tsechansky, M. & Macskassy, S. (2008), 'More than words: Quantifying language to measure firms' fundamentals', *The Journal of Finance* **63**(3), 1437–1467.

Tsai, M.-F. & Wang, C.-J. (2017), 'On the risk prediction and analysis of soft information in finance reports', *European Journal of Operational Research* **257**(1), 243–250.

Van der Heijden, H. (2022), 'Predicting industry sectors from financial statements: An illustration of machine learning in accounting research', *The British Accounting Review* **54**(5), 101096.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017), 'Attention is all you need', *Advances in Neural Information Processing Systems* **30**, 5998–6008.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A. & Grave, E. (2020), CCNet: Extracting high quality monolingual datasets from web crawl data, *in* 'Proceedings of the Twelfth Language Resources and Evaluation Conference', European Language Resources Association, Marseille, France, pp. 4003–4012.

Wolter, J. S., Bock, D., Mackey, J., Xu, P. & Smith, J. S. (2019), 'Employee satisfaction trajectories and their effect on customer satisfaction and repatronage intentions', *Journal of the Academy of Marketing Science* **47**(5), 815–836.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. et al. (2016*a*), 'Google's neural machine translation system: Bridging the gap between human and machine translation', *arXiv preprint arXiv:1609.08144* .

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. et al. (2016*b*), 'Google's neural machine translation system: Bridging the gap between human and machine translation', *arXiv preprint arXiv:1609.08144* .

Xu, Y., Armony, M. & Ghose, A. (2021), ‘The interplay between online reviews and physician demand: An empirical investigation’, *Management Science* **67**(12), 7344–7361.

Yang, Y., Uy, M. C. S. & Huang, A. (2020), ‘FinBERT: A pretrained language model for financial communications’, *arXiv preprint arXiv:2006.08097* .

Yao, J. & Shepperd, M. (2020), Assessing software defection prediction performance: Why using the matthews correlation coefficient matters, *in* ‘Proceedings of the Evaluation and Assessment in Software Engineering’, EASE ’20, Association for Computing Machinery, New York, NY, USA, p. 120–129.

Yekini, L. S., Wisniewski, T. P. & Millo, Y. (2016), ‘Market reaction to the positiveness of annual report narratives’, *The British Accounting Review* **48**(4), 415–430.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M. & Liu, Q. (2019), ERNIE: Enhanced language representation with informative entities, *in* ‘Proceedings of ACL 2019’.

Zhao, L., Li, L., Zheng, X. & Zhang, J. (2021), A BERT based sentiment analysis and key entity detection approach for online financial texts, *in* ‘2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)’, IEEE, pp. 1233–1238.

Zhu, Y., Hoepner, A. G., Moore, T. K. & Urquhart, A. (2022), ‘Sentiment analysis methods: Survey and evaluation’, *Available at SSRN 4191581* .