Erasmus University Rotterdam

Master Thesis Data Science and Marketing Analytics

Enhancing Recruitment Through Advanced NLP

A Novel Framework for ABSA Feature Extraction and Predictive Modeling

Haithem Habib Aoudia

501172

Supervisor

Dr. Nuno Camacho

July 06, 2024

Abstract

This research introduces a novel framework for extracting features from Aspect-Based Sentiment Analysis (ABSA) to enhance text classification and predictive modeling tasks. Utilizing the advanced Llama-3-70b architecture, we introduced and compared three distinct methods: Sentiment Scoring, Binary Aspect-Sentiment Encoding, and Aspect Sentiment Embedding. Our findings highlight the Binary Aspect-Sentiment Encoding method's superior interpretability and strong predictive performance. Additionally, we identified key strategies for improving candidate experience and talent acquisition, emphasizing the importance of structured interviews, prompt communication, and alignment with candidates' personal goals. This study not only advances ABSA methodologies that can be used across multiple domains but also provides actionable insights for businesses aiming to enhance their candidate experiences and talent acquisition strategies.

The Application of AI in This Study

ChatGPT and Grammerly were used to assist with editing and improving the quality of writing in terms of grammar, spelling, and sentence structure. ChatGPT was also used to debug errors in the code during the development of the web scraper script and the results collection scripts. Llama 3-70b was incorporated into the research methodology within the scope of this study for the ABSA task. I independently carried out research tasks myself, including the literature review, developing the ideas behind the new methods introduced in this study, data analysis, interpretation of results, and insights and conclusions derived. This use of AI tools complies with the policy of the Erasmus School of Economics (ESE).

Contents

1 Introduction **5**
- 1.1 Research Question . . . 5
- 1.2 Academic Relevance . . . 7
- 1.3 Industry Relevance . . . 7
- 1.4 Societal Relevance . . . 8
- 1.5 Outline . . . 8

2 Literature Review **8**
- 2.1 Candidate Experience and Strategic Talent Acquisition . . . 8
  - 2.1.1 Talent Recruitment . . . 8
  - 2.1.2 Candidate Experience . . . 10
  - 2.1.3 Employer Branding . . . 11
- 2.2 Sentiment Analysis and Feature Extraction . . . 13
  - 2.2.1 Aspect Based Sentiment Analysis . . . 13
    - 2.2.1.1 Aspect Term Extraction . . . 13
    - 2.2.1.2 Aspect Polarity Classification . . . 14
  - 2.2.2 ML Applications of ABSA . . . 14
  - 2.2.3 ABSA in Recruitment Domain . . . 15
  - 2.2.4 Generative LLM applications of ABSA . . . 16
  - 2.2.5 ABSA Feature Extraction . . . 16

3 Methodology **17**
- 3.1 3.1 Llama 3-70b Large Language Model . . . 18
  - 3.1.1 Transformer Neural Networks Architecture . . . 18
  - 3.1.2 Llama 3 ABSA Application . . . 21
- 3.2 End-to-end Feature Extraction Pipeline . . . 22
  - 3.2.1 Sentiment Scoring . . . 23
  - 3.2.2 Binary Aspect Sentiment Encoding . . . 24
  - 3.2.3 Aspect Sentiment Embeddings . . . 26
  - 3.2.4 Text Embedding . . . 26
  - 3.2.5 K Means – Aspect Aggregation . . . 27
- 3.3 Logistic Model . . . 27
- 3.4 Evaluation Metrics . . . 28

4 Data **29**
- 4.1 Data Acquisition . . . 29
- 4.2 Dataset . . . 30
- 4.3 Data Processing and Transformation . . . 31
- 4.4 ABSA Dataset . . . 32
- 4.5 Descriptive Statistics . . . 32

5 Results **36**
- 5.1 Aspect Based Sentiment Analysis . . . 36
- 5.2 Embedding Clusters . . . 40
- 5.3 Feature Extraction and Final Dataset Development . . . 41
- 5.4 Candidate Experience Prediction . . . 45

5.5 Recruitment Outcome Prediction 48
5.6 Comparison with Baseline Models 52

6 Discussion **54**
6.1 Technical Insights 54
6.2 Recruitment Insights 55

7 Conclusion **57**
7.1 Conclusion 57
7.2 Limitations 58
7.3 Future Research 59

8 References **60**

9 Appendix **65**

1 Introduction

The fierce competition for attracting exceptionally skilled professionals in the current job market presents a pivotal challenge for companies. According to McKinsey's "war for talent" report, only 23 percent of 6,000 executives surveyed strongly agreed that their companies successfully attract top-tier talent (Michaels et al., 2001). To address this challenge, companies are increasingly prioritizing optimizing the candidate experience as a strategic measure to cultivate a strong employer image and remain competitive in the employment landscape. Building a superior employee value proposition has proven to be a key differentiator for companies, significantly enhancing their ability to attract top talent. As such, developing a recruitment process that conveys a strong employee value proposition for potential hires is paramount (Michaels et al., 2001). Hence, the escalating war for talent is increasingly pressuring companies to adopt more data-driven techniques to optimize recruitment strategies and marketing techniques for employer branding.

Venturing into a new opportunity involves a significant degree of risk for talented candidates, prompting them to dedicate substantial time and effort to search for information about potential employers. In this quest for information, many turn to resources like Glassdoor reviews, which provide insights into the experiences of employees and candidates, offering a transparent look at the company's true employer brand. In job marketing signaling (Spence, 1978), individuals aim to avoid poor decisions by depending on signals that help assess quality. Hence, organizations must leverage their corporate brand to transmit the right signals in the recruitment process to enhance the likelihood that candidates will attribute a competitive advantage to the brand (Michaels et al., 2001).

Following the intense "war for talent" and the challenges companies experience as a result, much academic literature has shed light on this challenge and how companies can react. Such research has primarily focused on how employer branding can be leveraged strategically to attract talent (Collins and Han, 2004). However, in reality, the experience a candidate receives in the recruitment process is often neglected and is a significant driver in the company's ability to attract talent. However, current academic literature examining data-driven methods to enhance the recruitment processes using genuine feedback from candidates is extremely scarce. Moreover, online review analysis has predominantly been within the scope of sentiment analysis, topic modeling, and review classification. Recently, more fine-grained approaches have been introduced, namely, Aspect Based Sentiment Analysis (ABSA), which involves both aspect extraction and sentiment classification tasks (W. Zhang et al., 2023). However, such developments have predominately been done within the scope of introducing new frameworks and methods to increase the accuracy of the aspect extraction and sentiment classification tasks with very little focus on broader applications of such developments and their potential to be used in out-of-domain data to improve business outcomes. In light of this, there is a significant need to explore how fine-grained sentiment analysis methods such as ABSA can be utilized on out-of-domain data for downstream analysis and tasks that have the potential to provide business-relevant insights.

1.1 Research Question

Recognizing the significance of adopting a data-driven approach to building great candidate experiences to attract the best talent, this research aims to highlight how aspect-

based sentiment analysis can be used for downstream prediction tasks to identify opportunities for improving candidate experience and talent acquisition. As such, this research will focus on answering the following research question:

How can features be extracted from aspect-based sentiment analysis for downstream prediction tasks to identify opportunities for improving candidate experiences and talent acquisition?

Candidate experience encompasses the perceptions, emotions, and interactions a candidate experiences throughout their journey with a company throughout the recruitment process. These experiences are pivotal in shaping how candidates perceive a company as an employer, influencing their decision to accept a job offer and their subsequent engagement and loyalty to the organization.

While text mining and sentiment analysis are growing in popularity and application, very limited research has been done on using these techniques for analyzing candidate sentiment towards recruitment processes. In particular, the use of both aspect extraction and sentiment polarity analysis has been growing but has not been applied in the context of candidate reviews. The relevance of aspect and sentiment analysis together for improving candidate experiences lies in their ability to uncover underlying patterns, sentiments, and pain points that individuals may not explicitly express. This study further extends this application by exploring how aspect sentiment pairs retrieved can be used to predict if the candidate's experience was positive and if they will accept or reject a company's offer.

The application of large language models will form the basis of the aspect-based sentiment analysis (ABSA) task of this study. Since the release of BERT in 2018 by Devlin et al. (2018), it has been widely adopted for sentiment analysis and aspect extraction tasks. In contrast, in the current academic landscape, there is limited research into the application of generative large language models (LLM) for aspect-based sentiment analysis. This study will focus on using the latest state-of-the-art LLM Lama-3-70b model for performing ABSA. This has not been done in the past and will be pivotal in propelling forward the technical discourse on sentiment analysis methods.

In addressing the aforementioned challenges and opportunities related to employer branding and talent acquisition, the empirical approach aims to utilize ABSA using the Lama-3-70b model to predict the candidate's experience and if they will accept or reject the job offer using various feature extraction methods as the central research design. The Dutch technology industry is used as the primary case study given its importance to the global technology landscape and the extreme competition it experiences for talented technology professionals. In the Netherlands, demand for tech talents doubled in 2023, reaching 26 job vacancy openings per available tech worker and 5 out of 6 tech job openings going unfulfilled (Amsterdam Economic Board, 2017). Hence, candidate reviews from all technology divisions in the Netherlands are scraped from Glassdoor. In the context of this study, aspect-based sentiment analysis extends the concept of traditional sentiment analysis by not only categorizing sentiments as positive, negative, or neutral but also linking these sentiments to specific aspects or attributes mentioned in the reviews. The retrieved aspect sentiment pairs contain valuable contextual and semantic information, capturing nuanced sentiments associated with specific aspects of the candidate's experience. This study aims to take advantage of this valuable information in downstream tasks by introducing three novel feature extraction methods to transform the aspect sentiment pairs

into predictive features. This will be used to predict whether a candidate's experience is positive or negative and if they accept or reject a job offer using a logistic regression model. This approach will allow us to identify how ABSA can be used in downstream prediction tasks and also determine which aspects of the recruitment process are most important in shaping the candidate's experience and their decision to join the company.

1.2 Academic Relevance

Only recently has a growing body of literature started focusing on the importance of candidate experience as the war for tech talent ensues. Current research has not extended the application of advanced aspect-based sentiment analysis to understand candidate experience in a way that can aid recruiters and human resource professionals in improving their recruitment processes and positioning their employer value proposition more effectively. Moreover, the use of aspect extraction and sentiment analysis has been predominantly focused on customer reviews; extending its application to candidate reviews will be a valuable addition to the growing field of review mining for multi-domain analysis. In addition to this, the majority of studies have tackled the tasks of aspect term extraction and aspect sentiment classification independently. This research will focus on creating a multitask learning framework to handle these two tasks together. Akhtar et al. (2020).

Most studies focused on predicting outcomes from textual reviews utilize topics, word embedding, or bag of words models as features (HaCohen-Kerner et al., 2020;Stein et al., 2019), however this study will extend this application by establishing multiple frameworks for extracting features from aspect sentiment pairs to predict the outcomes in the recruitment process. No previous research has addressed the challenge of utilizing ABSA results for downstream analysis by evaluating different methods to determine the most optimal approach. Furthermore, the utilization of aspect-based sentiment analysis has predominantly been focused on benchmark datasets, where researchers introduce novel methods to outperform past methods without a specific focus on practical applications. Therefore, the application of ASBA on out-of-domain unlabeled datasets is very limited, given its challenges. Moreover, in the current academic landscape, there is very limited research into the application of generative large language models (LLM) for aspect-based sentiment analysis; this is still an emerging area that has not been fully addressed as most studies have focused on the utilization of BERT and LSTMs based models with self-attention mechanisms for ABSA tasks.

1.3 Industry Relevance

From an industry perspective, the research is highly relevant as it addresses a pressing concern of attracting technology professionals faced by organizations across the global technology sector. Given the large volume of public candidate review data online, it is difficult by nature for companies to translate this feedback into large-scale insights to improve candidate experiences and employer branding. Hence, companies are yet to adopt a robust model to transform this vast amount of public candidate review data into actionable insights that can drive meaningful improvements in candidate experiences. By leveraging advanced sentiment analysis techniques, this study aims to provide actionable insights that can empower human resource professionals, both large corporations and startups, to improve their recruitment processes, foster positive work environments, and

ultimately contribute to the overall growth and competitiveness of the global technology industry.

1.4 Societal Relevance

At a broader societal level, this research holds significant social relevance by tackling multifaceted issues pertinent to employment opportunities, diversity, and inclusion, as well as the cultivation of better candidate and company interactions. Centering on the experiences of candidates, this research acknowledges the critical role that workforce satisfaction plays in driving innovation, talent attraction, and economic prosperity, thereby aligning with broader societal goals of fostering a thriving and sustainable technology ecosystem. Moreover, the findings of this research will also be valuable to students, graduates and job seekers seeking to work in the Dutch technology sector. Offering insights into the current challenges of the job search landscape will empower them to make well-informed career choices, providing a clearer understanding of the industry's realities.

1.5 Outline

In the following sections, a theoretical framework is created through a detailed review of the current literature concerning talent recruitment, candidate experience, aspect-based sentiment analysis, and embedding feature extraction. Then, the generative LLMs chosen, Llama-3-70b for ABSA, the feature extraction methods, and the predictive model, logistic regression, will be explained in depth as part of the methodology. Moreover, the dataset scraped from Glassdoor will be described, and some explanatory data analysis will be conducted. Afterward, the application of the models, their results, and the comparative performance of different feature extraction methods in terms of predictive performance and interpretability will be discussed. Finally, the conclusions derived will be highlighted along with the limitations and recommendations for future research in this area.

2 Literature Review

2.1 Candidate Experience and Strategic Talent Acquisition

2.1.1 Talent Recruitment

Talent recruitment refers to the process an organization takes to generate applicant pools, maintain viable applicants, and encourage desired candidates to join those organizations Dineen and Soltis (2011). While Ployhart et al. (2017) defines recruitment in a broader sense, referring to a wide set of activities that connect applicants to organizations and their jobs and further differentiates between internal and external recruitment. In the context of this study, external recruitment will be the main focus. This typically stems from an organizational need or challenge, prompting companies to seek out the most capable and competent individuals externally who can effectively tackle these issues and drive the organization forward. This challenge has become more prominent in the current recruitment technology landscape given the increase in specialized roles due to advancements in niche technology areas. Hence, defining effective recruitment practices is essential, as the methods an employer uses to recruit can significantly influence the level of talent

attracted and whether the organization succeeds in hiring the talent it needs Breaugh (2013).

Despite the significant research conducted on recruitment, no general theory has been developed that explains how different recruitment variables, job applicant attributes, and organizational attributes interact to impact recruitment outcomes Breaugh (2013). Current theoretical frameworks of recruitment pertain to specific areas of recruitment in isolation from other activities. Hence, despite the extensive research on recruitment, the existing literature has often been criticized for lacking concrete conclusions on recruitment activities and their implications Breaugh and Starke (2000). Acikgoz (2019) seeks to bridge this gap by developing an integrative model that highlights the interplay between organizational-level and individual-level factors in shaping the outcomes of employee recruitment and job search activities. Given companies' unique business objectives, Phillips and Gully (2015) differentiates the concepts of strategic recruitment from traditional recruitment. They postulate that strategic recruitment aligns the firm's strategy and context with its recruitment processes and activities, highlighting this alignment as both crucial and a largely unexplored area in research. Despite this, such research does not take into account the heterogeneous nature of firms' size when crafting recommendations and frameworks for recruitment practices. Barber et al. (1999) further expands on this concept by establishing that larger firms tend to adopt more bureaucracy and formalities compared to smaller firms, which in turn influences job seekers' behaviors. Thus, the candidate pools for corporations and startups often differ in terms of talent. To effectively attract talent from these varied pools and enhance job offer acceptance rates, it is crucial for employers to be proactive with candidates. Becker et al. (2010) finds that quicker offers post interviews and assessments are most likely to be accepted and do not have adverse effects on turnover performance.

In the context of the fierce war for talent, digitalization and artificial intelligence have played a major role in shaping the current recruitment practices companies adopt. This has had a large impact on streamlining and enhancing recruitment processes and causing a shift towards skills-based hiring to reduce bias and improve hire quality. A growing number of research have recently captured this important shift, highlighting the need for companies to adopt more technology-driven recruitment processes. Van Esch et al. (2019) express the crucial need for HR and business executives to increase their efforts to adopt AI as part of their recruitment processes. This encompasses the entire recruitment process, from utilizing AI-based tools for candidate screening to employing gamification and AI-enabled interviews for assessments. A notable example is Unilever's partnership with AI HR service providers Pymetrics and HireVue, which revolutionized their screening and assessment processes Feloni (2017). As a result, applications more than doubled, greatly enhancing the diversity of the candidate pool. AI-enabled interviews and evaluations streamlined the selection of over 45,000 internship applicants down to just 300. Similar successes are observed across the technology industry, where AI-enabled recruitment has not only streamlined hiring processes but also markedly improved the quality of candidates.

While AI and data-driven recruitment strategies have significantly enhanced operational efficiencies, companies must prioritize cultivating a positive candidate experience to truly attract top talent rather than merely expanding their application pool and filtering it with AI. Black and van Esch (2020) argue that given the significant volume of applications

that AI-enabled outreach generates, most companies are in the rejecting business rather than the hiring business. Thus, it is in the self-interest of companies to ensure a positive experience for all candidates, especially the ones rejected.

In the cultivation of a recruitment strategy that is not only efficient for the company but also generates a positive candidate experience, it is key to set the right metrics to evaluate the success of the recruitment process. Traditional metrics companies commonly use include quality of hire, cost per hire, time to hire, and application drop-off rate (Hirebee, 2023; Skeeled, 2019; Recruitee, 2021). However, this does not paint the full picture of how successful the process is. The lack of proper feedback channels to assess the candidate experience and how they perceive it remains to be an issue most companies have not resolved. According to a study by Board (2017) 41.3% of companies do not survey or gather data from job candidates about their experiences. This is crucial considering that 77% of candidates who have a positive experience, and 61% of those with a negative one, share their feedback with friends and family. Without understanding these experiences, a company cannot effectively capitalize on the positive or address the negative to improve Black and van Esch (2020). This gap further establishes the necessity for this study, which aims to adopt a more fine-grained, data-driven approach to dissect what drives candidate experiences and how companies can improve to attract top talent.

2.1.2 Candidate Experience

Given the intense rise of competition for talent, candidate experience has become an integral part of a company's competitiveness and ability to attract highly skilled employees. The candidate experience encapsulates the interaction between job seekers and prospective employers, spanning various touchpoints from the initial application to interviews and feedback reception. This process was outlined by Schwab et al. (1987) depicting search, evaluation, and outcome as the major elements of the job search process. From the candidate's perspective, this is a very lengthy and intensive process with the ultimate aim of securing employment that aligns with their personal objectives. This entails thoroughly researching the job's requirements, company research, crafting CVs, preparing for interviews, and completing assessments from one decision point to another.

From the perspective of a highly talented candidate seeking employment, amidst navigating numerous job opportunities and offers, they are more inclined to steer clear of companies with negative experiences in the recruitment process. However, recent studies have shown that the impact of negative candidate experience does not solely pertain to the candidate's employment decision but also the overall brand image of the company. An annual survey conducted by Board (2016) revealed that 41% of candidates worldwide who encountered and reported a negative candidate experience expressed their intention to discontinue their affiliations, product purchases, or relationships with the respective organization. An internet service provider and television company, Virgin Media, quantified the cost of its poor candidate experience to be $6 million in lost revenue annually Adams (2016). This is primarily attributed to both the recruitment process itself and the employees responsible for interacting with job candidates, making them critical components of the employer brand Russell and Brannan (2016). This is further reinforced by Miles and McCamey (2018) as they emphasize the crucial role employer branding plays in business outcomes. They find that enhancing the employer brand through improvements in the recruitment process can foster stronger connections with customers,

investors, referrals from acquaintances, and future applicants to the company.

Candidate experience remains to be a very niche area with a limited number of literature studying it. A few studies have attempted to dissect the main determinants of a good candidate experience. Miles and McCamey (2018) finds communication to be the most important element for a good candidate experience. This encompasses the nature of the communications, the timeline for the process, acknowledgment of application materials, and candidate selection or rejection notifications. In essence, such activities between the candidate and the company allow the potential talent to directly experience elements of the organization's brand and its employees Miles and Mangold (2004). However, given the heterogeneity of applicants in a vacancy, the influence of certain factors in the recruitment stage may vary depending on the candidate. For instance, the content of job postings has a greater influence on experienced job seekers than inexperienced job seekers in terms of their organizational attitudes Walker et al. (2008). While Becker et al. (2010) establishes that proactiveness in providing feedback post-interviews plays a key role in shaping the candidate's experience and their decision to accept an offer. They find that quicker offers post interviews and assessments are most likely to be accepted and do not have adverse effects on turnover performance.

Given the prominent role Artificial Intelligence (AI) has played in recruitment, it has significantly impacted the candidate experience. Van Esch et al. (2019) suggest candidates should be actively engaged to complete AI-enabled recruitment processes and that HR and business executives should increase their efforts to adopt AI as part of their recruitment processes. On the other hand, Black and van Esch (2020) emphasizes the importance of positive experiences for all candidates, particularly those not selected, highlighting three key benefits. Black and van Esch (2020) establishes three main reasons for this. The first is that candidates who have a positive rejection experience are more likely to consider future opportunities with the company. This openness to reapply can be crucial as their suitability for roles may change over time. Secondly, the experiences of rejected candidates often translate into word-of-mouth that can significantly influence a company's reputation. Positive feedback from these candidates can significantly enhance the company's employer brand image and attract future applicants. Thirdly and most importantly in the current context, a positive recruitment experience increases the likelihood that selected candidates will accept job offers. Since companies can only hire candidates who accept their offers, creating a positive impression during the recruitment process is key to securing top talent.

Overall, the importance of fostering a positive candidate experience emphasizes the necessity of utilizing text-mining techniques on candidate reviews, such as aspect-based sentiment analysis. The adoption of such fine-grained techniques in analyzing candidate reviews has the potential to unveil nuanced details regarding the candidate experience and provide insights into areas for improvement, a realm largely unexplored in current literature.

2.1.3 Employer Branding

The notions of candidate experiences discussed are closely tied to employer branding. The candidate's experiences can be viewed as one of the main constructs of a company's image as an employer. It is also key to acknowledge that employer branding is a subset of the overall company's brand. Different stakeholder groups can adopt different views

toward a single entity. For instance, customers may perceive a company in a positive way due to the quality of their service. However, employees may perceive the company in a negative light due to adverse internal circumstances within the organization. Lievens and Slaughter (2016) elaborates on this by defining employer image as a mix of mental representations of specific aspects of a company as an employer held by individuals. These perceptions are held by individuals, can change over time, focus on specific aspects rather than an overall impression, and are cognitive in nature (Lievens and Slaughter, 2016).

A large number of studies have focused on the significance of maintaining a strong employer brand in an organization. Balmer and Gray (2003) advocates that a strong and favorable brand is a powerful indicator to a variety of stakeholders, which includes existing employees, shareholders, and also potential employees. This emphasizes the need for effective corporate brand management both internally and externally. In essence, employer branding enables firms to differentiate themselves from other employers in the market who are competing for the same talent. This is in aim to attract applicants that ideally possess identical values to that of the organization Backhaus and Tikoo (2004). This effect is observed in a study by Collins and Han (2004) as they find that strong and positive employer branding can increase the applicant quantity and quality. This in turn can increase organizational performance as well Fulmer et al. (2003). An example of this is the 100 Best Places to Work At list that is released annually. Fulmer et al. (2003) finds that companies on this list benefit not only from stable and positive employee attitudes, but also form superior performance over the broad market.

In the context of candidate experience, a key part of building a strong brand image is forming a strong employer value proposition for potential employees to join. Dabirian et al. (2019) attempts to shed light on this by establishing eight value propositions for employer branding: Social value, interest value, application value, development value, economic value, management value, work-life balance, and brand image. They found that interest value, in particular, which refers to the interest value of the work, was a main determinant in attracting employees and was often lacking once employees joined the company. This brings forward another issue, which is the gap between the internal employer image viewed by employees and the external brand image viewed by candidates. Lievens and Slaughter (2016) emphasizes this by making a distinction between the external brand image viewed by candidates and the internal brand image perceived by employees. Dabirian et al. (2019) explains this phenomenon where companies would brand themselves as innovative and cool, but such expectations were not met once employees joined. This results in a disparity between brand image and brand identity, specifically, the difference between how IT firms present their employer brand to the outside world to attract talent and the reality of how these promises are perceived internally by new hires. IT firms are recommended to focus significantly on synchronizing their external and internal branding efforts to close this gap and effectively retain new talent (Dabirian et al., 2019).

Overall, employer branding is an integral concept that companies must seek to prioritize as part of their strategic long-term vision. The need to align corporate branding with employer branding becomes even more imperative when considering the dynamic nature of stakeholder identities (Knox and Freeman, 2006). For instance, potential employees and candidates may also be the customers of an organization; both are key stakeholder groups that have a significant impact on the corporate brand (Knox and Freeman, 2006).

2.2 Sentiment Analysis and Feature Extraction

2.2.1 Aspect Based Sentiment Analysis

Given the large volume of opinionated text in online reviews, it is by nature difficult for humans to decipher and summarize the information and opinions in them. Moreover, given the subjectivity of human text analysis, considerable bias is bound to be present. Hence, automated opinion mining systems are necessary to extract relevant insights (Liu and Zhang, 2012). Sentiment analysis is a text-mining computational treatment of opinions, sentiments, and subjectivity of text, such as reviews, feedback, or comments (Medhat et al., 2014). It categorizes the text into positive, negative, or neutral sentiments, providing valuable insights into how individuals feel about specific aspects of their experiences. Hence, it is specifically geared towards understanding people's opinions, attitudes, and emotions towards specific entities (Pang and Lee, 2008). Such emotions include delight, joy, and satisfaction for positive sentiment and anger, fear, guilt, sadness, dissatisfaction, and frustration for negative sentiment (Balahur et al., 2012). As such, despite opinions and sentiments being very interrelated, it is key to recognize them as distinctly separate concepts. Opinion mining extracts and analyzes people's opinions about an entity, while sentiment analysis is more geared toward identifying the feelings and emotions people express Medhat et al. (2014).

Despite sentiment analysis becoming a very popular method due to the increasing number of large opinionated texts on social media, standard sentiment analysis methods lack the ability to capture nuanced sentiments directed towards specific topics and entities mentioned within the text, thereby limiting its application in truly understanding the sentiments behind certain product or service attributes. A recent development that addresses this is aspect-based sentiment analysis (ABSA). ABSA does not focus on the overall sentiment of the text but instead works by first extracting aspects within the text. This, for instance, could be a particular attribute or characteristic of an entity, such as services offered by a company or product features. Once the aspects are identified, the sentiment polarity associated with each aspect in the text is identified. This can tremendously improve our understanding of experiences towards processes as opinions of candidates towards certain areas of a company's recruitment can be extracted and analyzed on a larger scale. Hence, this empowers companies with the ability to convert a vast number of textual candidate reviews into actionable insights, enhancing opportunities to improve candidate experiences and optimize talent acquisition strategies. As such, ABSA can be divided into two subtasks: aspect term extraction (ATE) and aspect polarity classification (APC) (G. Zhao et al., 2023).

2.2.1.1 Aspect Term Extraction

The first task, ATE, aims to identify entities in a text that represent specific attributes and characteristics of a service in a review (G. Zhao et al., 2023). This task was first studied by Hu and Liu (2004) and further distinguished between implicit and explicit tasks. Since then, multiple methods have been proposed for this task with the aim of addressing the manual rule-based and time-consuming method that traditional ATE techniques encompassed. Poria et al. (2014) proposed a novel approach for this task through a rule-based approach that takes advantage of common sense knowledge and sentence dependency trees to identify both explicit and implicit aspects in a review. However, with the development of advanced machine learning and deep learning models,

the aspect extraction task can be done automatically from a corpus of text using both supervised and unsupervised methods. The supervised approach can be carried out using K nearest neighbors (Shah et al., 2020), for instance, while an unsupervised approach can be using pre-trained text models such as BERT (G. Zhao et al., 2023). However, in recent years, deep learning methods have mainly been adopted for ATE, given its good performance. Poria et al. (2016) utilizes a deep convolutional neural network to tag each word in a sentence as either an aspect or non aspect word. This method achieved higher accuracy than state-of-the-art methods.

2.2.1.2 Aspect Polarity Classification

The second sub-task behind ABSA is aspect polarity classification (APC). This aims to predict the sentiment polarity of each aspect. In recent years, this task has primarily been done through conventional machine learning and deep learning methods, with deep learning methods in particular being more widely adopted (G. Zhao et al., 2023). Li et al. (2018) proposed the use of transformation networks using a CNN layer. Moreover, an attempt to utilize attention mechanisms introduced by transformer neural networks has also been introduced (Vaswani et al., 2017). D. Ma et al. (2017) utilizes an interactive attention network to learn and represent targets and contexts separately to model target-context relationships for sentiment classification. However, recently, pre-trained language models have increasingly been used for aspect polarity classification tasks. Given the advanced transformer architecture and the pretraining done over large texts, promising results have been shown. Song et al. (2019) also applies attention-based encoders for the modeling between the context and target but employs BERT for the task, yielding high accuracy. To further enhance the understanding of the context, BERT can also incorporate external knowledge to improve its understanding of niche contexts. A. Zhao and Yu (2021) implements this through a knowledge-enabled BERT to obtain better embedding vectors for the aspect sentiment analysis task.

2.2.2 ML Applications of ABSA

Recently, many studies have adopted a multitask model for ATE and APC via aspect-based sentiment analysis for online customer reviews. Most prior methods utilize long short-term memory (LSTM) and attention mechanisms to predict the sentiment polarity of the specified targets. Y. Ma et al. (2018) employs this approach as they utilize a two-step attentive neural architecture and LSTM model for ABSA to achieve an improved performance on extracting aspect categories and sentiment polarity towards entities. However, such methods are often very complex and are computationally intensive. As such, a model based on convolutional neural network and gating mechanisms is proposed, resulting in more accurate sentiment predictions more efficiently (Xue and Li, 2018). The application of ABSA using transformer neural network architecture is later introduced as Hoang et al. (2019) utilizes BERT with a fine-tuning method to solve out-of-domain ABSA, which outperformed other state-of-the-art methods. Further developments in ASBA were introduced by incorporating domain knowledge into the model. A. Zhao and Yu (2021) achieved this by introducing a knowledge-enabled BERT for ASPA. This achieved superior performance with a relatively small amount of training data. Another line of research for ABSA is targeted sentiment analysis, which classifies the polarity of opinions about a certain target entity mentioned in sentences under scrutiny (Vo and Zhang, 2015). These studies collectively demonstrate the potential of

ABSA in understanding and extracting sentiment from online reviews.

2.2.3 ABSA in Recruitment Domain

The development and application of ABSA in current research have been predominantly focused on customer reviews. In particular, ABSA has been a recurring task in Semantic Evaluation challenges over the past decade, focusing on introducing novel ABSA models that achieve better predictive performance on relatively small customer feedback datasets by Mitchell et al. (2013) and Pontiki et al. (2014, 2015, 2016) such as restaurant review, laptop review and Twitter domains. The application of ABSA in other domains is very scarce, and when it comes to the recruitment domain, there is no literature that utilizes candidate reviews for ABSA tasks.

A key challenge of this, as a result, is that there are no post-trained models on candidate reviews. As a result, the application of ABSA on any language model will result in the training data having a different data distribution than the test data. This concept is key to ABSA. To adopt a fine-grained approach to sentiment analysis through the Aspect Term Extraction (ATE) and Aspect Polarity Classification (APC) tasks, it is important for the language model to be trained on sentences and words that are very similar to the data during inference. This presents a unique challenge for this study.

One approach to address this challenge would be to create a labeled dataset for the candidate reviews by setting a ground truth for aspect sentiment terms for the model to be trained on. This would involve manually annotating a substantial number of candidate reviews to identify and classify the aspects and sentiments expressed within them. However, this approach is very challenging due to the intensive nature of data labeling tasks and the difficulty in establishing a ground truth. It would require human experts in the recruitment domain and a sufficiently long time frame to build a large enough labeled dataset, which falls outside the scope of this study.

Another approach is to use a language model pre-trained on a sufficiently large corpus to generalize across various domains, thereby understanding contextual information and nuances in candidate reviews. This method leverages the model's ability to comprehend a wide range of contexts and vocabularies, which can be beneficial for accurately interpreting candidate reviews. This approach involves an out-of-domain task for ABSA, a crucial area requiring further exploration and research. Out-of-domain ABSA aims to develop models that can effectively analyze sentiment in domains not represented in the training data, significantly enhancing the applicability and robustness of ABSA models across various fields.

Previous approaches have typically trained models on every domain, leading to significant computational and resource costs. Performing multi-domain ABSA using traditional methods is very challenging and has proven to yield limited performance (Luo et al., 2022). Recent work in cross-domain ABSA has begun exploring the use of multiple LLMs for this task, yielding very promising results. Varia et al. (2022) showcased the ability to generalize across multiple ABSA subtasks with minimal examples by employing supervised fine-tuning and multi-task learning. Meanwhile, Fei et al. (2023) developed a multi-turn chain of thought (CoT) prompting approach using multiple Generative LLMs to comprehend implicit sentiments and opinions. They also tested multiple LLMs and found that the larger the LLM model, the better the performance in the sentiment analysis

task. These studies offer preliminary evidence of the significant potential of LLMs in ABSA tasks and tackling cross-domain challenges with ATE and APC tasks.

Recognizing the significant opportunity this presents, this research will focus on utilizing state-of-the-art large language models for the ABSA task, leveraging their strong cross-domain generalization capabilities to perform ABSA effectively.

2.2.4 Generative LLM applications of ABSA

Despite the frequent adoption of non-generative large language models like BERT for Aspect-Based Sentiment Analysis (ABSA) tasks, only a few studies have explored the potential of using generative LLMs in this area. Like BERT, generative LLMs are pre-trained on extensive text corpora from the internet to attain a robust general understanding of language. The latest generative models, such as GPT-4 and Llama-3, undergo an extremely intensive optimization process during pre-training, resulting in a highly sophisticated grasp of nuanced language use. LLMs like ChatGPT and LLaMA have only recently been utilized for ABSA tasks, where they have achieved significant success. As LLMs continue to scale up, new techniques like In-Context Learning (ICL) (Ye et al., 2024) and Chain of Thought (CoT) (Wei et al., 2022) have been developed. ICL shows that including detailed instructions and examples in task prompts can greatly improve task performance, both in zero-shot inference and supervised training scenarios(Yang et al., 2024).

W. Zhang et al. (2023) investigates this by applying state-of-the-art generative LLMs, from basic sentiment analysis to advanced ABSA. They find that even non fine-tuned zero-shot LLMs are capable of accurately predicting sentiment polarity for basic tasks but still fall short of specialized fine-tuned models for complex tasks such as ABSA. Furthermore, another key finding of their research is that large LLM models do not always guarantee better performance. Using prompt tuning on smaller LLMs such as Flan-UL2 suffice for practical sentiment analysis are comparable to large LLMs such as GPT-3.5 in performance (W. Zhang et al., 2023). Only a few initial attempts were made to compare how generative LLMs perform against BERT for sentiment analysis. An example of this is the study by Zhong et al. (2023) that finds that the zero-shot performance of LLMs is in line with a fine-tuned BERT model. However, the scope of the study is still very limited, and no comprehensive study has explored how generative LLMs can be employed for complex sentiment analysis tasks such as ABSA.

2.2.5 ABSA Feature Extraction

Extracting features from unstructured data is a crucial area in NLP that involves transforming raw text into meaningful representations that can be used for various downstream tasks. An important application of feature extraction is review classification, where techniques like sentiment analysis, topic modeling, and named entity recognition are employed. Predictive tasks in review text mining are primarily geared towards predicting ratings or sentiment classification tasks of reviews. This can be summarized as classifying a review as recommended or not recommended (Turney, 2002). Both supervised and unsupervised learning methods using machine learning and lexicon-based approaches have been frequently utilized in this task. Pang et al. (2002) use support vector machines to classify reviews as positive or negative, while they also use Naïve Bayesian to train a predictive model to calculate the sentiment polarity of a non-rated review. However, such

studies have not performed too well relative to topic-based categorization approaches. Hence, unsupervised approaches have also been introduced where the rating of the review can be predicted by calculating the average semantic orientation of sentences in a review(Turney, 2002). Lexicon based approaches were also used frequently given their simplicity. Thelwall et al. (2012) explores this using the SentiStrength algorithm. They find it robust enough to be applied to a wide variety of different social web contexts. However, this approach generally suffers from the fact that even positive words may have a different semantic meaning depending on the sentence it is used and the context.

In aim to incorporate semantic and contextual meaning as features for review classification tasks, multiple studies have explored the use of aspect, topics, or aspect sentiment pairs as predictive features. Qiu et al. (2018) develops a logit model that utilizes aspects and their sentiment in a review to predict the rating of a review. They find that this predictive framework is feasible and effective at predicting the rating of reviews. On the other hand, many studies focusing on review classification tasks utilize topic modeling techniques as features with the aim of addressing the high dimensionality of text data. Onan et al. (2016) uses a Latent Dirichlet allocation (LDA) topic modeling approach as features for numerous prediction algorithms. To assess the quality and effectiveness of different topic modeling approaches, Hong and Davison (2010) propose several training schemes and find that training a topic model on aggregated text can yield higher performance in the Twitter classification task. The main issue with using topic models as features is their lack of concrete interpretability in terms of whether the topic was mentioned in a positive or negative context. Therefore, this study will aim to move from using traditional topic modeling and aspect extraction techniques to state-of-the-art LLMs to obtain fine-grained aspect sentiment terms that can be easily interpreted.

3 Methodology

This chapter will explore the empirical research by diving into the technical concepts and processes of the study's methodology. Initially, the Llama-3-70b architecture will be examined and its application to the ABSA task. To address one of the main goals of this research, of identifying the most optimal way of extracting features from ABSA for downstream prediction tasks, three different approaches will be explored. The first approach is the Sentiment Scoring method inspired by the research of Binder et al. (2019) where they use aspect-based sentiments as independent variables to explain online review star ratings. This method involves extracting aspects from the text and based on the sentiment associated, it assigns a score to the aspect for each review. The second approach is the Binary Aspect-Sentiment Encoding approach inspired by the use of topics from topic extraction tasks as features in text classification prediction models (Büschken and Allenby, 2016;Korfiatis et al., 2019). Under this method, the aspect sentiment pairs extracted from all the reviews will be transformed into binary features, where if a review contains the aspect sentiment pair, it takes the value 1 and 0 otherwise. The third approach is the Aspect Sentiment Embedding approach. The idea of this approach stems from the numerous studies utilizing embedding for text classification tasks Stein et al. (2019). The embedding of the full aspect sentiment pairs of each review will be used as features in the prediction task. The three above approaches will be utilized which will result in the creation of three uniquely different datasets that all stem from the same ABSA done initially. An end-to-end feature extraction pipeline for all three approaches

is designed and will be outlined and detailed to develop the final datasets. Finally, the logistic regression model and evaluation metrics used will be thoroughly explained.

3.1 3.1 Llama 3-70b Large Language Model

The latest and most advanced open source state-of-the-art large language model to date is the Llama 3 model introduced by Meta on April 18th, 2024. This is a pretrained and instruction-tuned language model released in an 8 billion parameter model and a 70 billion parameter model. The model used in this study is the 70 billion parameter model. This model represents a significant improvement in terms of accuracy and contextual understanding and outperforms previous models in terms of semantic understanding of textual user-generated reviews.

3.1.1 Transformer Neural Networks Architecture

In order to grasp a solid understanding of how the Llama 3 model works, it is important to first outline the transformer neural network architecture. Transformer neural networks were first introduced by Vaswani et al. (2017) in the infamous paper "Attention is all you need". This was a pivotal study that revolutionized the field of natural language processing by introducing a mechanism called self-attention. Unlike previous architectures such as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), which processed sequential data in a linear fashion, transformers allowed for the parallel processing of data, drastically improving efficiency and performance.

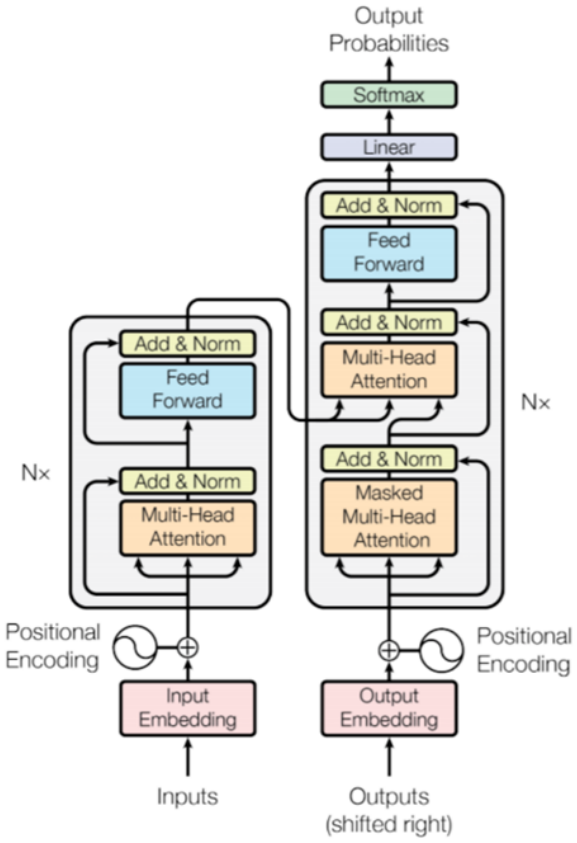


Figure 1: Transformer Encoder-Decoder Architecture from Vaswani et al. 2017

The transformer architecture consists of an encoder and decoder component. Each of these components can contain multiple layers which can be stacked to create deep networks capable of capturing complex patterns in data. The encoder component is responsible for processing input tokens through a process of tokenization, embedding, and self-attention to create a meaningful representation of text in a high-dimensional space that captures semantic information. The decoder component takes these encoded representations and generates the output sequence using attention mechanisms, feed-forward networks, and a softmax activation function. A common application of the transformer architecture is translation tasks. For example, if a text is being translated from Dutch to English, the encoder component would process the Dutch input sequence and convert it to a high-dimensional embedding vector. The decoder would then use this high-dimensional embedding vector to generate an output sequence in English. As Llama 3 is a decoder-only model consisting of only multiple decoder components, the following section will focus only on the decoder component.

Decoder-only transformers are a specialized variant of the standard transformer architecture, utilizing solely the decoder component. To cover the decoder block comprehensively, six important components of the decoder will be focused on: tokenization, embedding, self-attention mechanism, Root Mean Squared Normalization normalization, feed-forward networks, and the output layer. As the scope of this thesis is primarily focused on using Llama 3 for inference, only the model architecture for inference will be discussed without the components that are used for training, such as cross-attention and loss functions.

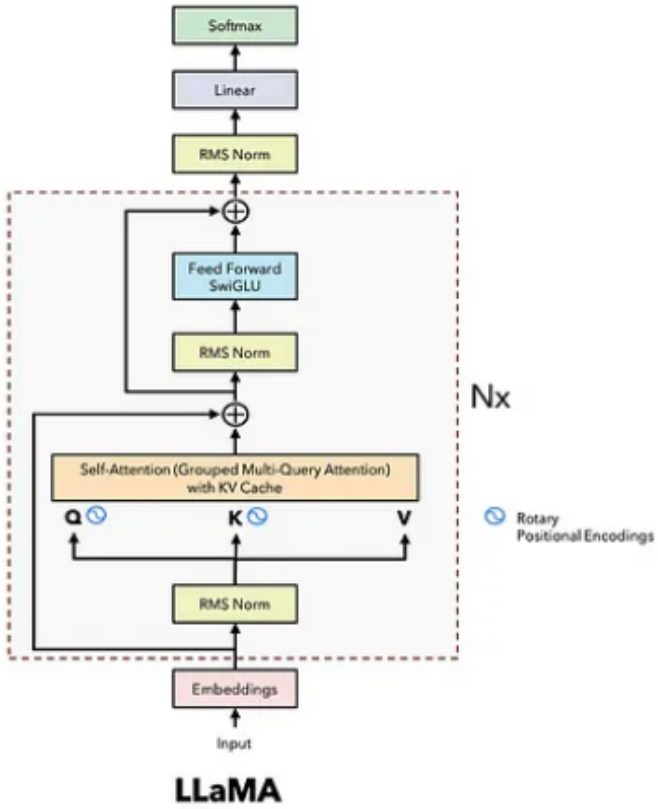


Figure 2: Llama 3 Decoder Architecture Illustration from Jamil (2023)

The initial stage in processing a text sequence through the decoder model is tokenization. This process involves breaking down the text into smaller units, which may vary from words to sub-words or even characters, depending on the model's level of granularity. The primary goal of this process is to convert the continuous stream of text into units that the model can process individually. Llama 3 has a vocabulary of over 128000 tokens and was trained on more than 15 trillion tokens.

Once the tokenization of the text is complete, the tokens are fed into an embedding layer. The primary goal of the embedding layer is to represent the tokens in a high-dimensional vector that carries both the semantic information of the tokens and its contextual information within the text. A critical part of the embedding process is positional embedding. Positional embedding is key to the transformer architecture as it allows the model to capture the position of each token in a sentence. While plain transformers consist of a standard positional embedding using cosine similarity, the Llama 3 architecture utilizes Rotary Positional Encoding (RoPE). RoPE is a recent development by Su et al. (2024) that improves the positional embedding of transformers by taking into account both absolute positional encoding and relative positional encoding. It applies a rotational computation to the vectors in addition to the fixed embedding. This enhances the model's attention calculation allowing the embedding to have a more flexible representation (Su et al. (2024)).

A key component of processing embeddings in the transformer architecture is self-attention. As certain words have different meaning depending on how they are used and which words it is combined with, it is paramount for the model to be able to distinguish between such meanings. Self-attention is a very important concept of transformers that was introduced to solve this challenge. The idea behind self-attention is to weigh the importance of each word in the sequence relative to the current word. By doing so, the model learns the similarity between each word in the sequence. For example, if a word is referring to another word in a sentence, then the similarity score will reflect that by impacting how the transformer encodes it. This is done by converting every token into a query (Q), a key (K), and a value (V). The attention scores are computed by calculating the dot product of the query with all keys (Vaswani et al., 2017). A softmax activation function is then applied to obtain the relative weights of each token. These weights are then used to compute a weighted sum using the value key. Hence, every token in the sequence attends to every other token, requiring the computation of attention scores for all pairs of tokens. The main drawback of this process is the high computational intensity and memory-intensive requirements. As such, to mitigate this issue, Llama 3 utilizes the grouped query attention instead introduced by Ainslie et al. (2023).

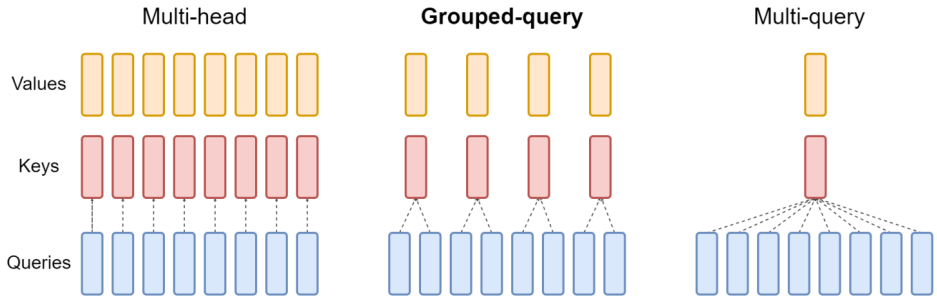


Figure 3: Grouped Query Attention Model from Ainslie et al. (2023)

Instead of each token independently querying all other tokens to calculate the similarity scores, the queries are partitioned into multiple groups based on similarity or token position. Within each group, the attention computation is performed and aggregated across all groups. This method is proven to achieve similar quality to multi-head attention but in significantly less time and computational complexity (Ainslie et al., 2023).

Transformers consist of very deep layers with many activation functions throughout. Hence, the scale of activation functions applied can grow exponentially through the many layers, resulting in gradient explosions, and conversely vanish if activations shrink too much. Hence, normalization layers are incorporated throughout the transformer architecture to stabilize the hidden states and substantially accelerate convergence (Ba et al., 2016). Llama 3 uses Root Mean Squared Normalization. It is a relatively novel approach introduced by B. Zhang and Sennrich (2019) that yields better performance in both training and inference.

Following the attention and normalization layers, the feedforward block plays a key role in the model's learning capabilities. It introduces non-linearity through an additional network of weights, biases, and activation networks to model the complex relationships generated by the embedding and attention layers (Geva et al., 2020). Llama 3 uses the SwiGLU activation function introduced by Shazeer (2020) and is a variation of Gated Linear Unit that combines gating mechanisms with nonlinear transformations yielding better results than the widely used ReLU activations (Ramachandran et al., 2017).

The final component is the output layer, which applies a softmax activation function to the final hidden state, transforming it into probabilities over the vocabulary to generate the prediction of the next word.

3.1.2 Llama 3 ABSA Application

Current ABSA models that achieve satisfactory performance hold a common assumption of the training and testing data coming from the same domain. When the distribution of the data between the training and testing changes, re-training the ABSA model is needed to guarantee satisfactory performance (W. Zhang et al., 2023). A significant challenge of this is that for out-of-domain data sets, an extensive data labeling task and the collection of the large text of that specific domain is required, which is often not feasible and too expensive (W. Zhang et al., 2023). The introduction of state-of-the-art generative large language models has been pivotal in addressing this issue. Llama 3 has been trained on 15 million tokens with text from a wide range of different domains from the internet. Hence, this comprehensive training has equipped Llama 3 with a more robust understanding of diverse topics and language use, making it capable of generalizing beyond specific domains and promising to be more capable of being used in ABSA within the recruitment domain.

In order to process over 7000 reviews using Llama 3, an API connection to the Llama 3-70b model is made using Groq via the request package in python. Groq is a high-performance AI platform designed to accelerate large-scale computations and deliver ultra-fast inference speeds. Groq's architecture is optimized for massive parallelism, allowing it to handle the extensive data processing and complex calculations required by large models like Llama 3-70b with minimal latency and maximum efficiency.

The ATE and APC tasks are carried out using the following prompt:

Recognize all aspect terms with their corresponding sentiment polarity in the given review delimited by triple quotes. The aspect terms are nouns or phrases appearing in the review that indicate specific aspects or features of the recruitment process. Determine the sentiment polarity from the options [positive, negative, neutral]. Only answer in the format [aspect, sentiment] without any explanation. If no aspect term exists, then only answer [].

This prompt was used based on a previous implementation of ABSA using Llama 3 by Sreenivasan (2024). The above prompt is then concatenated with the review being analyzed.

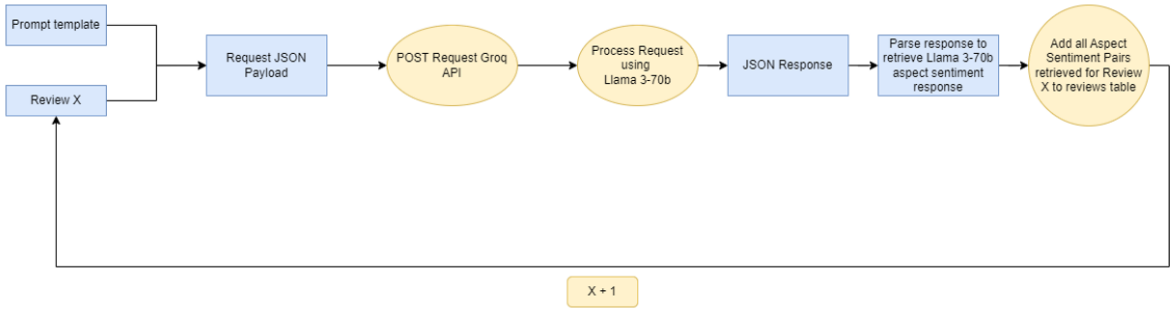


Figure 4: Llama 3 ABSA Pipeline

Figure 4 illustrates the end-to-end process of the ABSA using Llama 3-70b. The prompt template to instruct Llama 3-70b to carry out ABSA is concatenated with the review being analyzed to form a single prompt. This prompt is then fed into the request payload which represents the information that will be sent to the model. Since Llama 3-70b is a very large model that requires significant processing power to run, a post request to the Groq API is set up to leverage Groq's extremely fast inference speeds for interacting with Llama 3-70b. The response is then received in JSON format containing the response from Llama 3-70b which is then parsed and stored in the review data frame in a new separate column. This process is iterated until all reviews have been processed.

3.2 End-to-end Feature Extraction Pipeline

The next stage of the empirical research approach of this study is to extract features from the results of the aspect-based sentiment analysis. Feature extraction is a key area in NLP that refers to the process of identifying and isolating important aspects or characteristics of data, which can then be used to simplify and enhance the performance of data analysis tasks. It involves selecting a subset of features using a structured and effective approach to reduce the dimension of the feature space in the data (Trier et al., 1996). A key component of this is not only to transform the data into better representations but also into measurable features that downstream models and analysis tasks can be done efficiently. Traditional methods such as the Bag of Words, Term Frequency-Inverse Document Frequency (TF-IDF), and word embedding using Word2Vec and GloVe are widely used approaches for extracting features from text to be incorporated in prediction models (HaCohen-Kerner et al., 2020). As the primary goal is to transform ABSA results into features and not from the review text directly, it is key to capture the relationship between aspects and the context in which they were used. As such, since traditional text

feature extraction methods are more focused on word frequency or word co-occurrences, they fall short of capturing the nuanced and contextual information that ABSA provides. Hence, in aim to transform ABSA results into features while preserving as much information as possible, three different approaches are introduced: Sentiment Scoring, Binary Aspect Sentiment Encoding and Aspect Sentiment Embeddings. These will be elaborated in the following sections.

3.2.1 Sentiment Scoring

The first method used to extract features from ABSA is the sentiment scoring method. In the research by Binder et al. (2019), they predefined 6 different aspects and based on the sentiment associated with each aspect from ABSA they assigned a score for each aspect within each review. Their proposed approach is easy to interpret, providing valuable insights for analyzing user reviews. This study further extends this approach by carrying out the aspect term extraction task automatically rather than having a predefined list of aspects. A key challenge to this is the significantly high number of unique aspects extracted. Given the variety of different experiences candidates encounter, every candidate describes their experience in their own terminology. As such, a large number of different aspects of the recruitment process can be retrieved. Moreover, to avoid identical aspects such as 'interview' and 'interviews' being treated differently due to spelling and grammatical differences, it is key to treat these aspects as the same. Therefore, this necessitates the need for an aspect aggregation method to reduce the high dimensionality of the features and simplify the representation of the data such that downstream analysis can be carried out efficiently.

As the number of unique aspects is over 5000, manually going over and merging similar aspects is not feasible. Hence, a method that identifies different themes within aspects and groups the aspects together is necessary to overcome this challenge. As the aspects are retrieved in a textual format, it is first necessary to represent these aspects in a way that clustering algorithms can work with while also preserving the semantic and contextual information of the aspects. To do so, the aspects are transformed into an embedding. The embedding represents a high-dimensional vector matrix that can preserve the semantic and contextual information. Each dimension in the vector space can represent a specific attribute. Hence, similar aspects should be closer to each other in the embedding space, while more different aspects will be farther apart in the embedding space. The details behind the embedding process and embedding models utilized will be further elaborated in section 3.2.4.

To group and merge similar aspects together, a clustering algorithm is applied to the aspect embedding. K means clustering algorithms are commonly used as it provides a simple, effective and robust method of partitioning the data into K distinct clusters based on similarity. Each aspect embedding is assigned to the cluster with the nearest mean, ensuring that similar aspects are grouped together for further analysis. With k means, the choice of K is key to retrieving good and interpretable clusters. The optimal number of clusters are determined using the aid of both the elbow method and silhouette score which is outlined in appendix 1 and 2. Based on that, the optimal number of clusters chosen is 50. This implementation will be further discussed in section 3.2.5.

To effectively extract features from the retrieved aspect clusters, it is crucial to assign representative labels to each cluster that represents its main theme. As mentioned, given

the extremely large number of unique aspects, it is not feasible to manually go over all the aspects in each cluster to generate a representative cluster label. Hence, to generate a representative cluster label, the centroid cluster is used. As K-means clustering groups similar aspects together, the centroid of each cluster effectively represents the average or central aspect of that cluster. The centroid is calculated as the mean of all the embeddings within the cluster and serves as a summary of the cluster's main characteristics. By analyzing the centroid, the key features and common themes can be identified among the aspects in the cluster. This presents a more scalable approach that allows for the automatic generation of meaningful and representative labels for each cluster. Since K means is also prone to generating uninterpretable clusters due to noise in the data, the top 10 aspects nearest to the cluster centroid are also retrieved to validate if the cluster is truly representing one common theme. By doing so, any cluster where the aspects nearest to the centroid do not represent a common theme is removed from the analysis as they do not represent a feature that can be effectively used in downstream prediction models.

The final cluster labels retrieved represent the features that will be utilized in the development of the final dataset. For each review, if the aspect is present within any of the final aspect groups, a sentiment score is assigned to the aspect group. A description of the features in the final dataset is outlined in table 1.

Column Name	**Data Type**	**Description**
Review ID	int	Unique identifier for each review
Experience	int	Overall star rating assigned to the review (e.g., 1 to 5)
Outcome	str	Outcome of the job application process (Accepted Offer, Rejected Offer, No Offer)
Aspect_1	float	Aggregated sentiment score for the first aspect
Aspect_2	float	Aggregated sentiment score for the second aspect
...	float	Aggregated sentiment score for intermediate aspects
Aspect_50	float	Aggregated sentiment score for the fiftieth aspect

Table 1: Sentiment Scoring Feature Extraction Final Dataset

The scoring is based on the following: if it is associated with a positive sentiment, a score of +1 is added; if it is a neutral sentiment, 0 is added, and if it is a negative sentiment, -1 is added to the score. This enables the aggregation of an overall sentiment score for each aspect within a review.

3.2.2 Binary Aspect Sentiment Encoding

The second feature extraction approach developed is using aspect sentiment pairs together. This approach was inspired by Büschken and Allenby (2016) as they utilize topics extracted as features in a text classification task. Similar to the previous sentiment scoring method, an embedding of the text is first generated. However, in this approach, an embedding of both the aspect and sentiment pair together is generated. Similarly, k means is applied to the embedding as a dimensionality reduction method to simplify the representation of the data. As this approach generates an embedding vector for the full aspect sentiment pair, there are two key semantic pieces of information to preserve in the embedding: the aspect and the sentiment. Since the idea is to have more

similar aspect sentiments close together in the embedding space and vice versa, it is key to ensure that the aspect information does not dominate the sentiment information in the vector embedding. For instance, some unique aspects, such as "interviewer's accent," may not be very prevalent in the reviews. Hence, if for example, this occurs twice, once in a positive and the other in a negative context, the embedding may reflect the sentiment difference strongly by assigning "interviewer's accent, negative" closer to "interviewer's accent, positive" than "interviewer, negative" to take into account the unique presence of "interviewer's accent". In aim to address this challenge, different embedding methods will be utilized to evaluate which embedding model will best address this challenge, and also a higher number of groups will be used to cluster the aspect sentiment pairs.

Neutral aspect sentiment are excluded from this method. The main reasons to this is that they do not aid in providing actionable insights and explanatory information of why a candidate would regard their experience as positive or negative, or reject or accept an offer. The second critical reason, neutral aspect sentiments represent the a significantly large number of all the aspect sentiments retrieved adding an extremely significant increase in the computational cost and time to generating the embedding. Hence, to avoid an over complex model using the binary aspect sentiment encoding method, neutral aspect sentiment are removed.

In the implementation of K means, 80 clusters are used. This is also based on the elbow method using inertia and silhouette analysis as outlined in appendix 3 and 4 but also based on the overall interpretability of the clusters and feasibility of choosing cluster labels as well. Similar to the previous aspect scoring method, the method for choosing cluster labels is done using the cluster centroid method and the top 10 nearest aspect sentiment pairs to the centroid as a validation measure. Table 2 outlines the variables and their description in the final dataset developed using the binary aspect sentiment encoding method.

Column Name	**Data Type**	**Description**
Review ID	int	Unique identifier for each review
Experience	int	Overall star rating assigned to the review (e.g., 1 to 5)
Outcome	str	Outcome of the job application process (Accepted Offer, Rejected Offer, No Offer)
Aspect_sentiment_1	binary	Binary indicating if the first aspect sentiment pair is present
Aspect_sentiment_2	binary	Binary indicating if the second aspect sentiment pair is present
...	binary	Binary indicating if the aspect sentiment pair is present
Aspect_sentiment_80	binary	Binary indicating if the eightieth aspect sentiment pair is present

Table 2: Binary Aspect Sentiment Encoding Feature Extraction

In the development of the final dataset using this feature extraction method, the cluster labels for the aspect sentiment pairs are used as binary features. If the review contains the aspect sentiment pairs within the cluster, the cluster label chosen will take the value of 1 for that review, and 0 otherwise.

3.2.3 Aspect Sentiment Embeddings

The third and final feature extraction method directly uses the embeddings as features. The high-dimensional vector embeddings generated contains significant semantic and contextual information. In the research by Stein et al. (2019), they utilize word embeddings for text classification tasks, proving that it yields significantly high accuracy. As such, the utilization of embeddings as features has been widely adopted as a proven method for obtaining high accuracy in text classification. The description of the dataset variable using aspect sentiment embeddings are outlined in table 3.

Column Name	**Data Type**	**Description**
Review ID	int	Unique identifier for each review
Experience	int	Overall star rating assigned to the review (e.g., 1 to 5)
Outcome	str	Outcome of the job application process (Accepted Offer, Rejected Offer, No Offer)
Aspect Sentiment Embedding	float	Embedding array of length d of the entire aspect sentiments

Table 3: Aspect Sentiment Embeddings Feature Extraction Final

To retain the granularity of the data at the review level, the embedding is now generated across the entire aspect sentiment pairs of the review. For instance, if a review has the aspect sentiment pairs "interview, positive" and "assessment, negative", one embedding is generated to take into account both aspect sentiment pairs. As such, the final dataset using this method consists solely of the reviews, the embeddings, and dependent variables.

3.2.4 Text Embedding

Text embedding is a pivotal component of the success of the empirical research in terms of encoding the results of ABSA in a contextually and semantically rich representation. As such, state-of-the-art embedding tuned models are chosen as they are already trained over a large corpus of texts. The models chosen are all-minilm-l6-v2-f32 introduced by Wang et al. (2020)and gte-large-en-v1.5 introduced by Alibaba's NLP team. Despite LLMs themselves generating embeddings, these models are chosen as they are finely tuned for embedding tasks and have been demonstrated to significantly outperform LLMs in creating more effective embeddings. The goal behind choosing two different models is to evaluate whether the use of different embedding models in extracting features will substantially impact the performance of downstream prediction models.

The first embedding model used, all-minilm-l6-v2-f32, is built using the Bidirectional Encoder Representations from Transformers (BERT) architecture but optimized for smaller and more efficient usage. It uses a reduced parameter count of just 22.6 million parameters which significantly enhances its speed and usability without compromising much on performance. The all-minilm-l6-v2-f32 model is particularly efficient for tasks such as semantic search, clustering, and paraphrase identification. It achieves this by condensing

the input text into a fixed-size vector of length 384, enabling easier for tasks to compare text similarity and clustering (Reimers and Gurevych, 2020).

The second embedding model gte-large-en-v1.5 is built using the Transformer++ encoder. This encoder integrates various cutting-edge components, including BERT, RoPE (Rotary Position Embedding), and GLU (Gated Linear Units), which together enhance the model's ability to capture contextual dependencies and semantic information in text (NLP, 2024).

3.2.5 K Means – Aspect Aggregation

K-means is a clustering algorithm that partitions a dataset into K distinct, non-overlapping clusters. It works by iteratively assigning each data point to one of the K clusters where the goal is to minimize the variance within each cluster and maximize the variance between clusters (MacQueen, 1967). The process starts with the random initialization of K centroids representing the central point of the clusters. The algorithm then alternates between two main steps: assignment and update. In the assignment step, each data point is assigned to the nearest centroid based on the Euclidean distance, effectively grouping the data into clusters. In the update step, the centroids are recalculated as the mean of all data points in each cluster, shifting the centroids to new positions that better represent the cluster's center. This iterative process continues until convergence, which typically occurs when the centroids no longer change significantly between iterations, indicating that the clusters have stabilized (Selim and Ismail, 1984).

To obtain meaningful and interpretable clusters, it is paramount to choose the optimal number of clusters K. This is a predefined parameter of the algorithm that establishes how many centroids will be initialized and consequently the number of clusters. To determine K, the elbow rule to minimize the sum of squared distances between each data point and also maximize the silhouette score is used in providing an indication the optimal k. The interia and silhouette scores for all the iterations done are shown in Appendix 1, 2, 3, and 4.

In both the sentiment scoring and binary aspect sentiment encoding methods, k means is applied to the aspect embeddings using all-minilm-l6-v2-f32 and gte-large-en-v1.5. In all the embeddings, there was no distinctly clear optimal K value using the elbow method with the inertia and silhouette scores. Hence 50 is chosen as the optimal K for the aspect scoring method and a k of 80 is chosen for the binary aspect sentiment method as furthering cluster does not yield significantly better results.

3.3 Logistic Model

As the text classification task concerns a binary outcome variable, a logistic regression model is used due to the high level of interpretability it offers for identifying opportunities for improving candidate experiences and talent acquisition. The logistic regression model predicts the probability that a given input belongs to a particular class. It is primarily used for binary outcomes using the sigmoid activation function which maps the output values between 0 and 1. For the model to decide which class an observation belongs to, a probability threshold is used where probabilities above that threshold will classify the observation as 1, and if below the threshold, it will classify it as 0. The probability threshold used in this study is 0.5.

For each of the three feature extraction approaches, two logistic regression models are constructed, one where all-minilm-l6-v2-f32 embedding was used and the second where embedding gte-large-en-v1.5 was used.

In the sentiment scoring method, the following logistic regression model is used:

$$z = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \cdots + \beta_{50} s_{50}$$

The sentiment s_1 to s_{50} represents the sentiment scores for aspect 1 to aspect 50 where s_1 would be the accumulated sentiment score for the first aspect, s_2 for the second aspect and so forth. β_0 is the intercept and $\beta_1, \beta_2, \ldots, \beta_{50}$ are the coefficients corresponding to each aspect sentiment score.

In the binary aspect sentiment encoding approach, the following logistic regression model is used:

$$z = \beta_0 + \beta_1 AS_1 + \beta_2 AS_2 + \cdots + \beta_{80} AS_{80}$$

The aspect sentiment AS_1 to AS_{80} are binary variables representing each aspect sentiment pair extracted and from all 80 clusters. If the aspect sentiment pair is present in the review, it will take the value of 1, otherwise it will take the value of 0. β_0 is the intercept and $\beta_1, \beta_2, \ldots, \beta_{80}$ are the coefficients corresponding to each binary aspect sentiment.

In the aspect sentiment embedding approach the following logistic regression model is used:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d$$

The embedding are represented by a vector $\mathbf{x}$ with d dimensions. Hence, each text input is converted into a feature vector $\mathbf{x} = [x_1, x_2, \ldots, x_d]$ where x_i represents the i-th feature from the embedding. As such, each dimension represents will represent a feature in the logistic regression where x_i is the value of the embedding in dimension i. Similarly, β_0 is the intercept and $\beta_1, \beta_2, \ldots, \beta_d$ are the coefficients corresponding to each dimension of the embedding.

The output from the formulas, z represents the linear combination of input features. The logistic model transforms this value to a probability between 0 and 1 using the sigmoid function.

$$p = \sigma(z) = \frac{1}{1 + e^{-z}}$$

$\sigma(z)$ is the sigmoid function and e is the base of the natural logarithm.

3.4 Evaluation Metrics

In alignment with the research objectives, which aims to evaluate methodologies for extracting features from ABSA results for downstream prediction tasks and identifying opportunities for improving candidate experiences and talent acquisition, two primary criteria will be employed: model prediction performance, goodness of fit and model interpretability.

The first criterion, model prediction performance, is selected for its critical role in ensuring that the extracted features accurately represent the relevant information. This is essential for the model to effectively learn and map the relationships between these features and

the dependent variables, candidate experience and recruitment outcome. To quantitiavely assess this, the accuracy and F1 score will be used.

The second criterion is goodness of fit. Goodness of fit measures how well the logistic regression model describes the observed data. It helps assess whether the model is appropriate for the data, providing further assessment for the best feature extraction methods. High goodness of fit means the model accurately captures the relationship between the independent variables and the dependent variable, leading to reliable predictions and insights. To quantitatively evaluate this, the Pseudo R squared and the Akaike Information Criterion scores are used. The Psuedo R squared indicates the proportion of variance in the dependent variable that is predictable from the independent variables. Higher pseudo R squared values suggest better model fit. Meanwhile, AIC is a measure of the relative quality of statistical models. It assesses the trade-off between the goodness of fit and the complexity of the model. Lower AIC values indicate a better model, balancing fit and complexity to avoid overfitting. AIC assumes the models are trained on the same data. Despite each feature extraction method yielding a different dataset for the logistic regression models, all methods are applied to the same underlying data which is the candidate reviews scraped from Glassdoor. Each method simply results in a different representation of this data. Therefore, the metric remains a valid measure for this study.

The third criterion, interpretability, refers to the ease of which the model's decisions and the relationships it has learned can be understood. This is crucial in this study's context of recruitment. It is not enough for a model to be accurate; stakeholders must also be able to comprehend why certain decisions are made. The interpretability is evaluated by examining to what extent can the features used allow us to identify opportunities for improving candidate experiences and talent acquisition.

4 Data

This section will first provide an overview of the data acquisition process, detailing the methods used for scraping, parsing, transformation, and storage. Next, it outlines the preprocessing steps, including data cleaning and text translation for both prediction tasks. Finally, it presents the finalized dataset used for analysis, along with its descriptive statistics and some explanatory data analysis.

4.1 Data Acquisition

To obtain the required data for this study, reviews from candidates on Glassdoor were scraped. Glassdoor is one of the largest recruiting websites where current and former employees can anonymously review companies and their management. It provides a platform for employees to share insights about their work environment, salaries, and interview processes. Additionally, it offers companies the opportunity to showcase their brand and attract potential talent by engaging with the reviews and providing information about their organization. Due to the intense competition for talented technology professionals, particularly in the Netherlands, the data collection focused exclusively on job vacancies in Dutch technology companies to align the context of this study. This includes any company or department dedicated to providing technology-based products or services, such as software, automation, or applications.

The end-to-end flow of the data acquisition pipeline encompasses data extraction, data parsing, and storage. This process is outlined in figure 5.

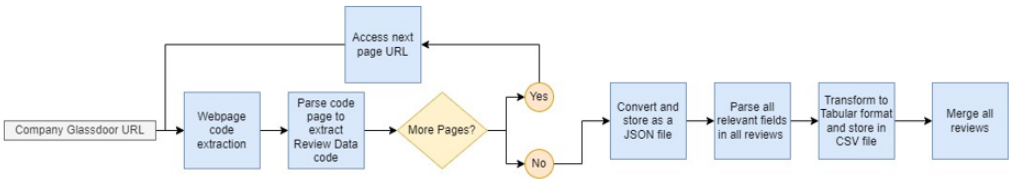


Figure 5: Data Scraping Pipeline Flow

Utilizing the candidate review pages on Glassdoor, the web page's codebase is extracted and parsed to retrieve all user-submitted review data. A scraper bot performs this task across all pages, gathering reviews and saving them as JSON files. A separate script then processes these JSON files, extracting relevant review fields and storing them in a tabular format. This procedure is repeated until all reviews are parsed and stored. Finally, the data is merged into a CSV file, which serves as the final dataset for analysis in this study.

4.2 Dataset

The final dataset collected includes 19 features representing various aspects of the review the user submits. 7436 reviews have been collected for 40 different technology companies in the Netherlands. Only companies and departments with technology-driven products or services are included.

Feature Index	Feature Name	Description
1	Review ID	Unique identifier for each review
2	Review Date	Date when the review was posted
3	Company Name	Name of the company being reviewed
4	Current Job	Whether the reviewer is currently employed at the company
5	Outcome	Outcome of the job application process (Accepted Offer, Rejected Offer, No Offer)
6	Candidate Review	Candidate review of the recruitment process
7	Advice	Advice given by the reviewer
8	Difficulty	Rated difficulty of the interview process (Easy, Medium, Difficult, Very Difficult)
9	Duration Days	Duration of the interview process in days
10	Employer Response	Response from the employer on the review
11	Employer Response Date	Date when the employer responded
12	Experience	Overall experience rating (Positive, Neutral, Negative)
13	Job Title	Job title related to the review
14	Interview Questions	Questions the candidate got asked in the interview

Table 4: Complete Dataset Extracted

Table 4 presents all the review data extracted from Glassdoor. This includes user-submitted textual reviews describing their experiences, rating scores for the candidate experience, and the outcome of the process whether the job offer was accepted if one was extended by the company. Additionally, data on various aspects of the interview process are extracted, such as the perceived difficulty of the interview questions, the length of the process, and the specific interview questions asked, along with any advice the candidate has for future applicants. Furthermore, since employers can reply to reviews, these responses are also included in the dataset.

4.3 Data Processing and Transformation

To meet the data requirements for the analysis of this study multiple processing and transformation steps are implemented. First, since not all the reviews were in English, all textual reviews were translated to English. Google's Cloud Translation API is used to ensure accurate translation. It is an AI-based solution known for its reliability and is used as an enterprise-grade solution. Secondly, non-alphanumeric characters are removed from the text, excluding numbers and punctuation. Moreover, typos are fixed using the autocorrect package in Python. Given the LLM's comprehensive and advanced understanding of textual data, these are the only processing steps applied to the textual reviews. Traditional text data cleaning pipelines, such as removing stop words, stemming,

and lemmatization, are not needed as LLMs inherently handle these aspects efficiently. This streamlined approach ensures that the data remains rich in context and meaning, providing a solid foundation for accurate and insightful analysis. As the central focus of this study is on candidate reviews, only the candidate review will be utilized for the analysis without the use of the additional data scraped.

4.4 ABSA Dataset

As established already, one of the main goals of this study's empirical approach is to utilize ABSA to extract aspect sentiment pairs from the candidate reviews for the prediction task. As such, separate datasets are used for the ABSA task and the prediction task.

Feature Index	Feature Name	Description
1	Review ID	Unique identifier for each review
2	Candidate Review	Candidate review of the recruitment process

Table 5: Dataset for ABSA

The data used in the ABSA task is shown in table 5, the candidate review text field is utilized for the ABSA and the review id to be able to identify each unique review submitted.

4.5 Descriptive Statistics

The descriptive statistics of the 7,436 reviews in terms of the recruitment outcome, candidate experience and process difficulty are presented in Tables 6, 7, and 8.

Outcome	Frequency	Percentage Distribution
No Offer	3820	51%
Accepted Offer	2906	39%
Declined Offer	710	10%

Table 6: Outcome Distribution

Candidate Experience	Frequency	Percentage Distribution
Positive	4433	60%
Negative	1772	24%
Neutral	1172	16%

Table 7: Candidate Experience Distribution

Difficulty of The Process	Frequency	Percentage Distribution
Very Easy	343	5%
Easy	1199	17%
Average	4187	58%
Difficult	1355	19%
Very Difficult	143	2%

Table 8: Process Difficulty Distribution

Most respondents reported a positive experience, with 60% indicating favorable feedback. In contrast, 24% reported a negative experience, and 16% had a neutral experience. In terms of the recruitment outcome variable, 51% of candidates received no offer, 39% accepted offers, and 10% declined offers. However, to align with the candidate-centric focus of this study, the group that received no offers will be excluded from the recruitment outcome prediction task. The majority of candidates found the recruitment process to be of average difficulty (58%), with 5% finding it very easy and 2% finding it very difficult. Overall, while most candidates did not receive offers, those who did generally had positive experiences and found the process to be of average difficulty. Due to class imbalances in the outcome and candidate experience variables, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to increase the minority classes of the training set in both prediction tasks. This method adjusts the data distribution so that the minority class

represents 80% of the number of samples in the majority class, creating a more balanced dataset.

In terms of understanding what drives a candidate to accept or reject an offer, as outlined in the literature review, the experience a candidate receives throughout the recruitment process is a critical factor. The pie charts in Figure 6 and Figure 7 illustrate the distribution of offer responses based on candidate experience, divided into negative and positive outcomes.

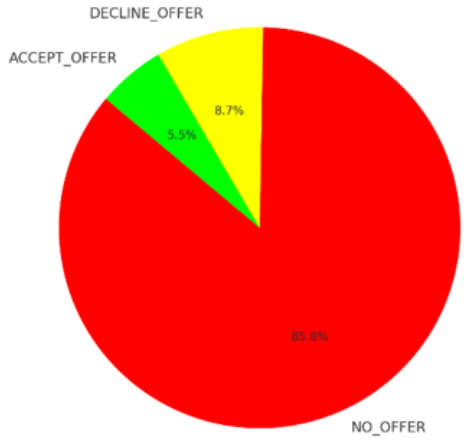


Figure 6: Negative Experience Offer Acceptance Distribution

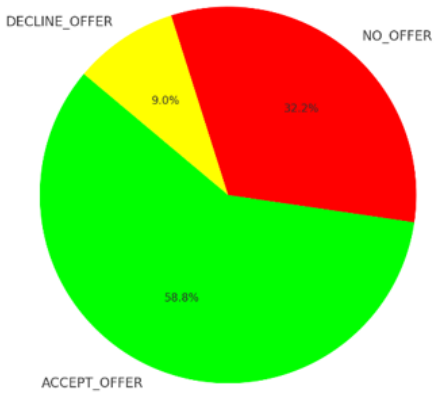


Figure 7: Positive Experience Offer Acceptance Distribution

In the negative outcome chart, 85.8% of candidates did not receive an offer, 5.5% accepted offers, and 8.7% declined offers. In the positive outcome chart, 58.8% accepted offers, 32.2% did not receive offers, and 9.0% declined offers. Comparing these outcomes reveals that candidates with positive experiences are far more likely to accept offers. Additionally, rejected candidates who report a negative experience are significantly higher than those reporting a positive experience. This may suggest that factors such as how candidates are rejected or perhaps ghosting candidates may play an influential role in the final decision of whether to view the experience as positive or negative. The proportion of candidates declining offers is similar across the positive and negative experience groups, suggesting that once candidates are in the final round and get given an offer, their decision is motivated by more than just if they view the experience as positive or negative. Hence, this necessitates a more nuanced understanding of candidate behavior and how they view the company in terms of their interactions with company employees, and the job opportunity itself. As such, this further emphasizes the importance of companies adopting and understanding how a systematic, data-driven approach can be utilized to improve candidate experience and their recruitment efforts.

With the plethora of research done on textual reviews and ratings, it has been established that review length can be an important predictor of the rating. Users with highly positive and highly negative experiences feel more obliged to leave a review about their experience outlining either extremely positive or negative aspects. The box plots in Figures 8 and 9 highlight whether the same can be said about the candidate reviews by exploring the review token length across candidate experiences and recruitment outcomes.

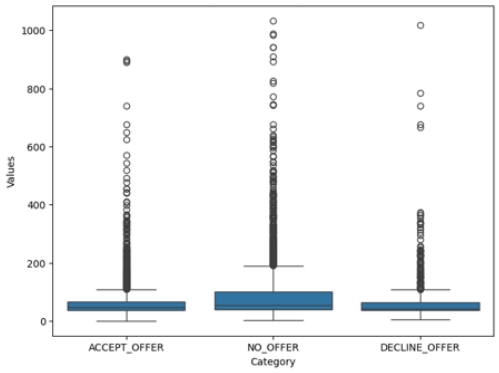


Figure 8: Negative Experience Offer Acceptance Distribution

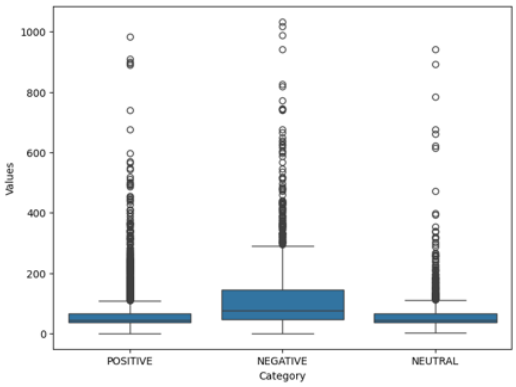


Figure 9: Positive Experience Offer Acceptance Distribution

In the candidate experience box plots, negative sentiment has the highest average token length in reviews, while in the recruitment outcome box plots, candidates who were rejected have the highest average token length in their reviews. This aligns with the idea that negative experiences prompt candidates to leave more detailed and longer reviews about the recruitment process. Both positive and neutral experiences result in similar review token lengths, indicating that candidates with these experiences tend to leave shorter and less detailed reviews compared to those with negative experiences. This implies that ABSA may result in more aspect sentiments retrieved within negative experience reviews and reviews where candidates were rejected. As such, negative feedback and rejection experiences are likely to provide richer data for extracting aspect sentiments, which could offer deeper insights into specific areas for improvement in the recruitment process.

The terminology used by candidates is an important differentiator between negative and positive experiences, offering insights into which areas of the recruitment process are being spoken about in negative and positive reviews. This is visualized in the word clouds for positive and negative reviews in Figures 10 and 11.

Figure 10: Positive Reviews Word Cloud

Figure 11: Negative Reviews Word Cloud

The terms "interview," "recruiter," and "company" are prominent in both positive and negative reviews, suggesting that these areas are focal components of a candidate's experience. This prominence indicates that candidates frequently mention these aspects in their feedback, making them significant targets for aspect extraction tasks in ABSA. In positive reviews, words like "conversation," "team," and "question" appear more frequently. This suggests that positive experiences often stem from the personal interactions

and relational aspects of the recruitment process. Engaging conversations, a supportive team environment, and relevant questions likely contribute to a candidate's favorable perception of the process. While in negative reviews, terms such as "time," "call," and "feedback" are more prevalent. This indicates that negative experiences are often associated with the procedural and operational aspects of the recruitment process. Issues related to the length of the process, the frequency and quality of communication, and the feedback provided to candidates appear to be significant factors contributing to a negative experience.

To truly address the goals of this study and build a systematic data-driven approach to improving candidate experience and talent recruitment, a more contextual understanding is needed of these terms. The application of ABSA will enable us to pinpoint the specific aspects that constitute positive and negative experiences in every review. By employing a comprehensive end-to-end feature extraction approach, the aspect-sentiment pairs can be leveraged to predict candidate experience and offer acceptance, providing a more detailed understanding of candidate feedback and thereby enabling targeted improvements in the recruitment process to build more positive experiences and attract highly talented candidates.

5 Results

In this section, the results from ABSA, feature extraction, and prediction models will be covered comparing each feature extraction method and highlighting the significance of the embedding model chosen. First, the results from the ABSA task will be analyzed by exploring the aspect sentiments retrieved. Secondly, the results from the aspect aggregation task, which includes creating embeddings for the ABSA output and clustering them, will be outlined and discussed. Thereafter, the results of feature extraction and the final data set developed for all three methods will be covered. Lastly, the results of all prediction models are outlined and evaluated in accordance with the research methodology.

5.1 Aspect Based Sentiment Analysis

The ABSA task yields aspect sentiment pairs for each review depending on how many aspects the model is able to detect for each review. To explore these results, the aspect sentiment pair across the whole dataset is analyzed. Table 9 shows the 10 most frequently occurring aspect sentiment pairs across all reviews.

Aspect Sentiment Pair	**Frequency**
interview, neutral	1223
interviews, neutral	674
process, positive	526
recruiter, neutral	503
interview, positive	469
recruiter, positive	468
interview process, positive	422
cv, neutral	416
questions, neutral	406
hr, neutral	349

Table 9: Top 10 Most Frequent Aspect Sentiment Pairs

A significant number of aspect sentiment pairs are neutral, such as “interview, neutral” appearing 1223 times and “interviews, neutral” appearing 674 times. This indicates that many reviews may contain descriptive aspects of the recruitment process rather than opinionated statements. Among the most frequent aspects extracted are ”interview,” ”recruiter,” and ”CV,” representing the most commonly discussed areas in candidate reviews. Additionally, certain aspects like ”process,” ”recruiter,” and ”interviews” are frequently mentioned in a positive context, suggesting that candidates often view these elements favorably when they are handled well.

Aspect Sentiment Pair	Frequency
m&as, neutral	1
code skills, neutral	1
a/b testing approach, neutral	1
html/css knowledge, neutral	1
main tasks, neutral	1
team related questions, neutral	1
team at booking.com, neutral	1
pen and paper, neutral	1
leetcode-like problem, neutral	1
feedback culture, positive	1

Table 10: Top 10 Least Frequent Aspect Sentiment Pairs

The least frequent aspect sentiment pairs from Table 10 reveal that these pairs pertain to very niche areas of the interview experience, often specific to candidates in certain fields. Examples include "html/css knowledge, neutral" and "a/b testing approach, neutral", each mentioned only once across all reviews. However, it is important to note that these niche aspects are part of broader categories within the interview process. For example, "html/css knowledge" aspect is part of the technical knowledge area of the recruitment process, yet because ABSA employs a fine-grained approach that identifies aspects based on noun and subject recognition, it is capable of extracting very specific and detailed aspects. This specificity highlights the comprehensive nature of ABSA, allowing for a deeper and more nuanced understanding of candidate experiences across various disciplines and technical domains. This notion can be further emphasized by the distribution of aspect sentiment frequency in Figure 12.

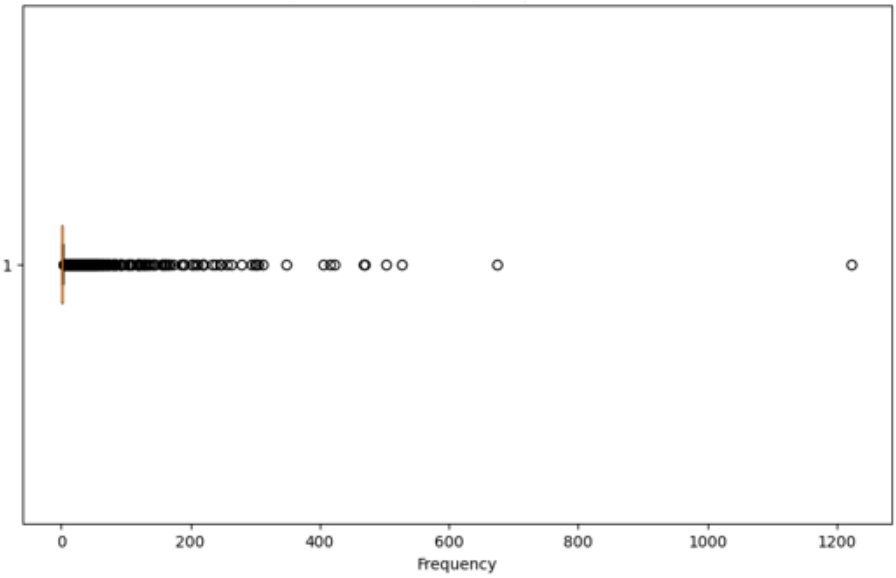


Figure 12: Aspect Sentiment Frequency Distribution

Figure 12 reveals that a large majority of aspects are mentioned only a few times, often just once or twice. This can be attributed to the automated aspect term extraction (ATE) task, which identifies numerous unique aspects. This is particularly heightened

by the fact that reviews are solely from tech companies and departments. Technology is a very broad area that also consists of many niche domains such as web development, backend engineering, data engineering, and many others. Each of which consists of its own specialized terminology and language used to describe specific aspects of each niche domain.

While the broad range of specific aspects extracted offers a detailed and nuanced understanding across multiple domains, this study's focus is more granular than solely analyzing individual aspects within each domain. Instead, it adopts a higher-level approach when interpreting the results from ABSA. This strategy aligns with the primary objective of developing a method for extracting features that can be used in downstream analysis and prediction. Therefore, the abundance of unique, low-frequency aspects necessitates developing a method for aggregating similar aspects together into broader categories while still retaining the sentiment associated with it. By aggregating related aspects, the dimensionality of the data can be significantly reduced, and more interpretable results can be provided for analyzing reviews at scale.

To gain a better understanding of the aspect sentiment pair that are frequently occurring in different candidate experiences and recruitment outcomes, tables 11, 12, 13, and 14 show the 5 most frequently occurring aspect sentiments across all relevant categories.

Aspect Sentiment Pair	**Frequency**
interview, neutral	777
interviews, neutral	468
process, positive	455
interview process, positive	376
interview, positive	362

Table 11: Top 5 Most Frequent Aspect Sentiment Pairs for Positive Experience

The most frequent aspect sentiment pair in positive reviews is "interview, neutral," occurring 777 times. This indicates that descriptive rather than opinionated reviews also constitute a significant proportion of positive feedback. Pairs such as "interviews, neutral" (468 instances), "process, positive" (455 instances), "interview process, positive" (376 instances), and "interview, positive" (362 instances) are also very prevalent highlights the importance of the interview itself and the process in general as key factors in shaping favorable candidate experiences.

Aspect Sentiment Pair	**Frequency**
interview, negative	238
Feedback, negative	221
Recruiter, negative	209
interview process, negative	157
Company, negative	151

Table 12: Top 5 Most Frequent Aspect Sentiment Pairs for Negative Experience

In the negative experiences in Table 12, the pair "interview, negative" appears 238 times, and "interview process, negative" is mentioned 157 times, further emphasizing that inter-

views are also a focal point in negative experiences as well. Additionally, other frequent aspects in negative reviews include "feedback, negative" (221 instances), "recruiter, negative" (209 instances), and "company, negative" (151 instances). As such, the quality of feedback, interactions with recruiters, and the overall perception of the company are frequently mentioned negatively, comprising of a large proportion of the negative reviews.

Aspect Sentiment Pair	**Frequency**
interview, neutral	116
Interviews, neutral	221
Process, positive	209
Questions, neutral	157
Recruiter, neutral	151

Table 13: Top 5 Most Frequent Aspect Sentiment Pairs for Declined Offer

Based on Table 13, the most frequent pair is "interview, neutral," appearing 116 times, followed by "interviews, neutral," with 73 instances. This suggests that many candidates provide descriptive feedback about the interview process rather than strong opinions, even when declining offers. Other common pairs include "process, positive" (58 instances), indicating that some candidates still had positive views of the overall process despite deciding not to accept the offer.

Aspect Sentiment Pair	**Frequency**
interview, neutral	238
Interviews, neutral	221
Process, positive	209
Interview process, positive	157
Recruiter, positive	151

Table 14: Top 5 Most Frequent Aspect Sentiment Pairs for Accepted Offers

While in the group that accepted the job offer, similarly the most frequent pair is "interview, neutral," mentioned 490 times, and "interviews, neutral," with 342 instances further signifying that interviews are a focal point in all reviews. Moreover, "process, positive" is also frequently appearing 325 times, highlighting that a positive perception of the overall recruitment process could significantly influence offer acceptance. Additionally, "interview process, positive" is mentioned 284 times, and "recruiter, positive" appears 231 times, further emphasizing the importance of favorable interactions during interviews and with recruiters.

Aspect Sentiments	**Frequency**
Positive aspects	10633
Neutral aspects	26971
Negative aspects	7479

Table 15: ABSA Sentiment Distribution

The ABSA using Llama 3-70b has resulted in the extraction of a substantial number of neutral aspects, with 26,971 occurrences. This underscores the observation that many sentences in the reviews are descriptive rather than opinionated. As previously established, this could also be due to the fact that llama 3-70b has focused on extracting many nouns and subjects based on the prompt used. Positive aspects are the second most frequent, occurring 10,633 times, followed by negative aspects, which appeared 7,479 times. The llama 3-70b's ability to identify a wide range of aspects, including niche and highly specific ones but also general ones, demonstrates its precision and comprehensiveness in understanding the presence of a wide variety of tasks in the ATE task. This capability is crucial for addressing the main goals of this research, which are enabling the identification of opportunities to improve candidate experience and talent acquisition. However, to achieve this goal, this study aims to do this by developing a feature extraction method for ABSA. As established already, for this to be done, it is not feasible to analyze 45083 aspect sentiments as it does not solve the main problem of the difficulty of analyzing large number of textual reviews. Hence, an aspect aggregation method is implemented which will be further analyzed in the following section.

5.2 Embedding Clusters

When clustering embeddings, it is crucial to form clusters that signify a common theme effectively. This process ensures that the aggregated aspects are meaningful and relevant to the goals of the research. While K-means clustering is particularly suitable for this task where each cluster centroid signifies the average position of all points within the cluster, it is important to acknowledge that clustering may not be perfect due to several factors. As the number of aspects extracted is very large, it may be that random aspects have been extracted due to noise in the text. Their embeddings are likely to be significantly distant from others, making them outliers within the data set. Additionally, some aspects may be too specific, resulting in embeddings that are far removed from more general or commonly occurring aspects. These distant or unique embeddings can lead to the formation of clusters that are not easily interpretable, as they do not align well with common themes in the recruitment process. This is further explored by the cluster results illustrated in Tables 16 and 17.

Model	**Not Interpretable Clusters**	**Final Clusters**
gte-large-en-v1.5	2 (209.0 aspects)	48 (4409 aspects)
minilm-l6-v2-f32	5 (2345 aspects)	45 (2373 aspects)

Table 16: Aspect Clusters Summary – Sentiment Scoring

Table 16 shows the aspect clustering results as part of the sentiment scoring feature extraction method. For the gte-large-en-v1.5 model, 2 clusters were not interpretable containing 209 aspects, resulting in 48 final clusters comprising 4,409 aspects. In contrast, the all-minilm-l6-v2-f32 model produced 5 not interpretable clusters containing 2,345 aspects, with only 45 final clusters encompassing 2,373 aspects. The gte-large-en-v1.5 embedding model results in significantly fewer discarded aspects compared to the all-minilm-l6-v2-f32 model. When many aspects are not interpretable and thus excluded from the final analysis, it can impact downstream tasks by reducing valuable information in the dataset. As such, the effectiveness of the embedding model chosen plays a critical role

in the performance of K-means clustering, influencing the formation of the final clusters that can be used for downstream analysis. Therefore, gte-large-en-v1.5 superiority in clustering aspect embedding over minilm-l6-v2-f32 signifies the importance of selecting an appropriate embedding model to generate meaningful clusters.

Model	**Not Interpretable Clusters**	**Final Clusters**
gte-large-en-v1.5	4 (446 aspect sentiments)	46 (4873 aspects)
minilm-l6-v2-f32	2 (610 aspect sentiments)	48 (4709 aspects)

Table 17: Aspect Sentiment Clusters Summary – Binary Aspect Sentiment Encoding

Table 17 shows the aspect sentiment clustering results as part of the binary aspect sentiment encoding feature extraction method. The gte-large-en-v1.5 model resulted in 4 not interpretable clusters containing 446 aspect sentiments, leading to 46 final clusters with 4,873 aspect sentiments. In contrast, the all-minilm-l6-v2-f32 model produced 2 not interpretable clusters containing 610 aspect sentiments, with 48 final clusters comprising 4,709 aspect sentiments. These statistics indicate that while the gte-large-en-v1.5 model is also superior in clustering aspect sentiments as well, the difference is not very significant. However, when comparing the clustering results of the embedding models, it is important to consider that a large number of aspects may include a significant amount of noise. Many of these aspects could be non-interpretable, too niche, or random. Therefore, disregarding such aspects could be beneficial for downstream prediction tasks, as it helps focus on the most relevant and coherent data, potentially enhancing the overall accuracy and interpretability of the predictions. Investigating whether the discarded aspects and aspect sentiments represent noise or should have been included in a different cluster with a common theme falls outside the scope of this study. However, examining the performance differences in the predictions between the embedding models can provide insights into whether this is indeed an important issue to address.

5.3 Feature Extraction and Final Dataset Development

As previously outlined, for each aspect, if the associated sentiment is positive 1 is added, if neutral sentiment 0 is added, if it is a negative sentiment 1 is subtracted. Hence, a higher average sentiment score indicates that the aspect is mentioned frequently positively, and a low average sentiment score signifies that the aspect is frequently mentioned negatively. Figure 13 shows the average sentiment scores under the sentiment scoring method for all aspects retrieved after aggregating all aspects using the all-minilm-l6-v2-f32 embedding model.

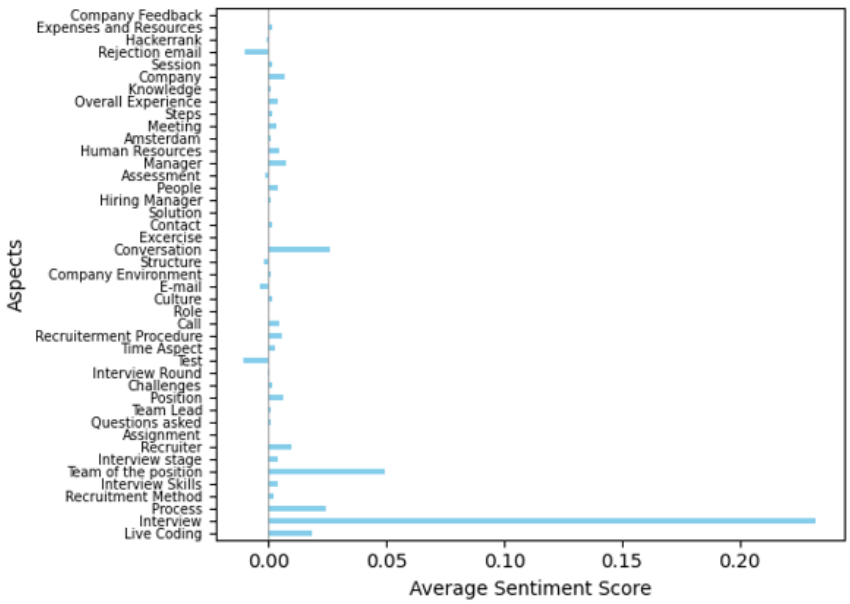


Figure 13: Average Sentiment Scores for all Final Aspects using all-minilm-l6-v2-f32 embedding

Most aspects have average scores close to 0, suggesting that either positive and negative sentiments are balancing each other out or there is a significant loss of aspects in the non-interpretable clusters, leading to a dilution of sentiment intensity. Among all the aspects, "Interview" stands out with a significantly higher positive average sentiment score compared to the rest. "Live Coding," "Team of the Position," and "Conversation" also exhibit highly positive mentions, indicating favorable candidate experiences in these areas. Conversely, "Rejection Email" and "Test" are the most negatively mentioned aspects, highlighting areas of dissatisfaction. This indicates that most negative aspects of the candidate experience are related to the operational side of the process, such as emailing, testing, and assessment, while positive aspects are more associated with personal interactions, such as interviews and conversations.

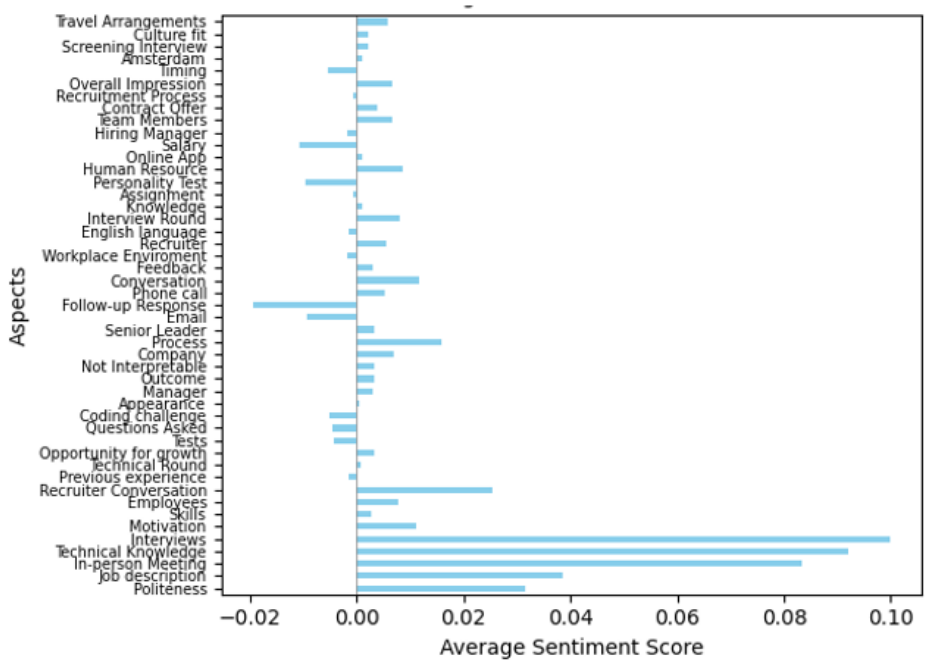


Figure 14: Average Sentiment Scores for all Final Aspects using gte-large-en-v1.5

Contrary to the embedding results from the all-minilm-l6-v2-f32 model, the aspects in Figure 14 using the gte-large-en-v1.5 embedding model are scored further away from 0, either more towards positive or negative rather than neutral. This can be attributed to fewer aspects being discarded in using the gte-large-en-v1.5 embedding model, resulting in a higher sentiment intensity for each final aspect cluster. The aspects "Interview," "Technical Knowledge," "In-Person Meeting" and "Politeness" are mentioned significantly more positively than the rest of the aspects. On the other hand, "Follow-Up Response", "Email", "Salary" and "Personality Test" are the most negatively mentioned aspects. This aligns with the findings under the all-minilm-l6-v2-f32 embedding model, where personal aspects of the candidate experience are more positively rated while operation and practical aspects of the recruitment process are more negatively rated.

These insights are crucial for identifying opportunities to improve candidate experiences. Positive ratings for personal interactions suggest that enhancing elements like interview quality, technical discussions, and face-to-face meetings can significantly boost candidate satisfaction. Conversely, addressing the negative feedback on follow-up responses, email communications, salary discussions, and personality tests can mitigate dissatisfaction. Moreover, comparing the results from both embedding models, it can be established that the embedding model chosen plays a significant role in the development of the final clusters and consequently the sentiment scores in the final dataset.

Under the binary aspect sentiment encoding method, the aspect sentiment pairs are transformed into binaries in order to transform them into features. Tables 18 and 19 present the average values of aspect sentiment pairs transformed into binary variables using the gte-large-en-v1.5 embedding model, where the presence of an aspect sentiment pair in a review is assigned a value of 1, and its absence is assigned a value of 0. Hence, these average values, ranging from 0 to 1, indicate the proportion of reviews in which each aspect sentiment pair appears across all reviews.

Aspect Sentiment	Average Value
interview, positive	0.353416
Hiring process, negative	0.111888
Recruiter, negative	0.105971
Technical knowledge, negative	0.101399
Conversation, positive	0.084857
Recruiter, positive	0.074771
Atmosphere, positive	0.072485
Job position, positive	0.069527
Interview process, negative	0.053389
Phone call, negative	0.047741

Table 18: Top 10 Most Frequent Aspect Sentiments using gte-large-en-v1.5 embedding

Aspect Sentiment	Average Value
Amsterdam, positive	0.003362
Room, negative	0.004303
Developer, negative	0.004841
Case, negative	0.005379
Interview Round, positive	0.005514
Assignment, positive	0.005648
Questions, positive	0.005917
Office, positive	0.005917
Interview rounds, negative	0.006321
Website, positive	0.006590

Table 19: Top 10 Least Frequent Aspect Sentiments using gte-large-en-v1.5 embedding

Among the 10 most frequent aspect sentiments, "interview, positive" has the highest average value of 0.353416, indicating its presence in approximately 35% of the reviews. This significantly higher presence compared to other aspect sentiments is consistent with previous findings, highlighting the positive perception of the interview process. On the other hand, negative sentiments related to the "hiring process" (0.111888), "recruiter" (0.105971), and "technical knowledge" (0.101399) are also among the most frequently mentioned, highlighting them as common negative areas of the recruitment process. The hiring process and recruiter are key areas of any recruitment process, and thus, it can be expected that bad candidate experiences are likely to be stemming from such areas. Technical knowledge is an interesting aspect as it most likely represents the technical knowledge of the interviewer or recruiter interacting with the candidate. The frequent mention of this aspect negatively suggests that interviewers may not have exhibited or presented themselves as well versed in the technical domain of the job opportunity as the candidate. Generally, highly talented candidates are very well-versed in their field and, therefore, critical of perceived gaps in technical expertise during the interview process.

Aspect Sentiment	Average Value
interview, positive	0.227542
Recruiter, positive	0.153443
Interviewer, negative	0.085530
conversation, positive	0.080016
company image, positive	0.078806
extra help, positive	0.078268
Evaluation process, negative	0.059844
Interview skills, negative	0.047203
investment on people, positive	0.045992
feedback, positive	0.043572

Table 20: Top 10 Most Frequent Aspect Sentiments using all-minilm-l6-v2-f32 embedding

Aspect Sentiment	Average Value
position matching, positive	0.001076
requirements, negative	0.001345
Recruitment Agency, negative	0.001345
recruitment phase, negative	0.001345
room, negative	0.001614
Waiting period, negative	0.002555
future growth, positive	0.002690
interview stage, positive	0.002824
test review, negative	0.002824
compensation package, positive	0.003362

Table 21: Top 10 Least Frequent Aspect Sentiments using all-minilm-l6-v2-f32 embedding

Tables 20 and 21 show the average values of aspect sentiment pairs using the all-minilm-l6-v2-f32 model. Among the 10 most frequent aspect sentiments, "interview, positive" is also the most frequent aspect sentiment with it being present in over 22% of the reviews. Additionally, "recruiter, positive" is also prevalent, appearing in 15% of reviews. For the negative sentiments, "interviewer, negative" (0.085530), "evaluation process, negative" (0.059844), and "interview skills, negative" (0.047203) are among the highest in

frequency. These are consist with the previous results, further highlighting the interviewer and their interview skills as negative areas of the candidate experience. On the other hand, "Compensation package, positive" (0.003362) and "future growth, positive" (0.002690) are among the least prevalent aspect sentiments across all reviews, indicating that positive sentiments regarding compensation and future growth opportunities are rarely mentioned by candidates.

5.4 Candidate Experience Prediction

The results of the candidate experience binary classification task across all feature extraction methods are outlined in Table 22.

	SENTIMENT SCORING		BINARY ASPECT SENTIMENT ENCODING		ASPECT SENTIMENT EMBEDDING	
	gte-large-en-v1	all-minilm-l6-v2-f32	gte-large-en-v1.5	all-minilm-l6-v2-f32	gte-large-en-v1.5	all-minilm-l6-v2-f32
ACCURACY	89%	89%	90%	89%	90%	87%
F1 SCORE – NEGATIVE CLASS	79%	79%	82%	80%	84%	77%
F1 SCORE – POSITIVE CLASS	93%	92%	93%	92%	93%	91%

Table 22: Logistic Regression Results Predicting Candidate Experience

In the Sentiment Scoring method, the gte-large-en-v1.5 and all-minilm-l6-v2-f32 models both achieved an accuracy of 89%. The F1 scores for the negative class were 79% for both models, while the F1 scores for the positive class were 93% for the gte-large-en-v1.5 model and 92% for the all-minilm-l6-v2-f32 model. These results indicate that both embedding models perform similarly using the Sentiment Scoring method, with stronger performance in predicting positive candidate experiences but slightly lower performance in predicting negative experiences.

For the Binary Aspect Sentiment Encoding, using the gte-large-en-v1.5 embedding model, it achieves the highest overall accuracy of 90%, with F1 scores of 82% for the negative class and 93% for the positive class. The all-minilm-l6-v2-f32 model achieved slightly lower accuracy of 89% and similar F1 scores to the gte-large-en-v1.5 embedding model. This method shows a slight improvement in predicting negative experiences compared to the Sentiment Scoring method, particularly with the gte-large-en-v1.5 model, which also maintained high performance in predicting positive experiences.

Using the Aspect Sentiment Embedding method, the gte-large-en-v1.5 model achieved the highest accuracy of 90%, whereas the all-minilm-l6-v2-f32 model achieved an accuracy of 87%. This method has the highest performance in predicting negative experiences with the gte-large-en-v1.5 model with an F1 score of 84%, indicating its effectiveness in capturing nuanced sentiment details with negative reviews. However, the all-minilm-l6-v2-f32 model showed a slight decrease in performance, particularly in predicting negative experiences.

Overall, the performance of all models is relatively similar, with no very significant difference or notable bad performing models. All feature extraction methods are able to effectively capture the semantic and contextual information for the model to understand the relationships between the features and the candidate experience. However, the Binary Aspect Sentiment Encoding and Aspect Sentiment Embedding methods, in particular, provided the highest accuracy and F1 scores across both positive and negative classes, particularly when using the gte-large-en-v1.5 model. The Aspect Sentiment Embedding

method, in particular, showed superior performance in predicting negative experiences with the gte-large-en-v1.5 model, achieving an F1 score of 84%. The Sentiment Scoring method, while effective, showed slightly lower performance compared to the other two methods.

To evaluate the models in terms of the goodness of fit, the Psuedo R squared and the AIC scores are used and are outlined in Table 23.

Model	Pseudo R squared Score	AIC
Sentiment Scoring - gte-large-en-v1.5	0.36	4872
Sentiment Scoring - all-minilm-l6-v2-f32	0.35	4882
Binary Aspect Sentiment Scoring Method - gte-large-en-v1.5	0.40	4603
Binary Aspect Sentiment Scoring Method - all-minilm-l6-v2-f32	0.34	5029
Aspect Sentiment Embedding - gte-large-en-v1.5	0.77	3689
Aspect Sentiment Embedding - all-minilm-l6-v2-f32	0.58	3778

Table 23: Pseudo R squared and AIC of logistic models

The Sentiment Scoring models using gte-large-en-v1.5 and all-minilm-l6-v2-f32 have pseudo R squared scores of 0.36 and 0.35, respectively, suggesting that they explain approximately 35-36% of the variability in the candidate experience variable. The Binary Aspect, Sentiment Scoring Method models, have more diverged scores between the embedding models, with the thegte-large-en-v1.5variant achieving a score of 0.40, indicating a higher explanatory power than the all-minilm-l6-v2-f32 variant, which has a score of 0.34. The Aspect Sentiment Embedding models have significantly higher pseudo R squared scores than all the other models. The gte-large-en-v1.5 variant scores 0.77, while the all-minilm-l6-v2-f32 variant scores 0.58.

In terms of the AIC, the Sentiment Scoring models have AIC scores of 4872 for gte-large-en-v1.5 and 4882 for Minilm. These scores are relatively high compared to the other models. The Binary Aspect Sentiment Scoring Method models have AIC scores of 4603 for gte-large-en-v1.5 and 5029 for Minilm, indicating that the gte-large-en-v1.5 variant provides a better fit than the all-minilm-l6-v2-f32 variant. The Aspect Sentiment Embedding models have the lowest AIC scores, with the gte-large-en-v1.5 variant scoring 3689 and the all-minilm-l6-v2-f32 variant scoring 3778.

Overall, the gte-large-en-v1.5 embedding model explains the variability in the candidate experiences more effectively than the all-minilm-l6-v2-f32 embedding model. The Aspect Sentiment Embedding feature extraction method using the get embedding model yields the highest pseudo R squared score and the lowest AIC score, indicating it may be the best-performing model in terms of model fitting the data. This is consistent with the notion of the embedding preserving the highest level of semantic and contextual information from the ABSA compared to the other feature extraction methods that require a series of aggregation transformations that may not preserve all the information from ABSA.

To explore the impact and statistical significance of each feature on the prediction of candidate experience via the logistic regression model, the coefficients from the logistic regression models are outlined and analyzed through the figures below. As the aspect sentiment embedding model utilizes a large number of abstract dimensions as features, it is very challenging to interpret an individual feature's impact on the model. Hence, the

focus will be on the coefficients from the sentiment scoring and binary aspect sentiment encoding features.

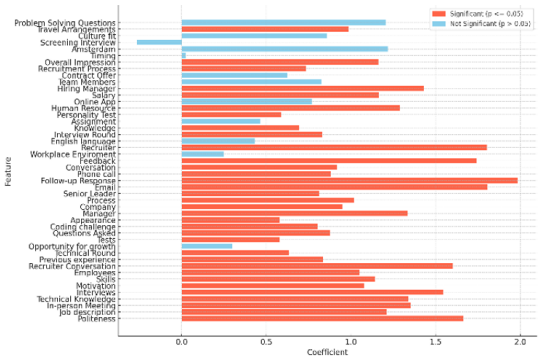


Figure 15: Logistic regression coefficients - gte-large-en-v1.5

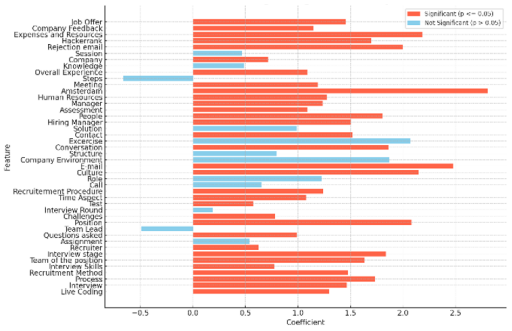


Figure 16: Logistic regression coefficients - all-minilm-l6-v2-f32

Figures 15 and 16 present the coefficients from the logistic regression models for each embedding model, utilizing a significance level of 5%. Variables with a p-value below 0.05 are deemed significant and are highlighted in red, while those that are not significant are highlighted in light blue.

In Figure 15, which uses the gte-large-en-v1.5 embedding model, "Follow-up Response" and "Email" have the highest positive coefficients, indicating the presence of these aspects in a review will significantly increase the odds of the review being about a positive candidate experience. This emphasizes the critical role of effective communication and timely follow-ups in fostering a positive candidate experience. "Politeness" also has a high positive coefficient, highlighting the importance of the personal side of interacting and speaking with candidates. Additionally, "Workplace Environment" is also among the highest significant coefficients, further showcasing the importance of having an attractive workplace in generating a positive candidate experience. The aspect "Screening Interview" has the only negative coefficient, but it is not statistically significant in this model.

In Figure 16, which uses the all-minilm-l6-v2-f32 embedding model, the aspect "Amsterdam" has the highest significant coefficient. Given that the reviews are from the Dutch technology sector, Amsterdam's frequent mention in reviews was expected. However, its high positive coefficient suggests that the mention of location in a review significantly increases the odds of the review being on a positive candidate experience. Similar to the gte-large-en-v1.5 model, aspects such as "Email," "Company Feedback," and "Company Environment" are significant and have high positive coefficients, indicating their strong influence in creating a positive candidate experience.

Comparing the two models, both emphasize the importance of providing good feedback, maintaining effective communication with candidates, and showcasing an attractive company environment. These aspects are consistently highlighted as critical factors in cultivating a positive candidate experience across both embedding models.

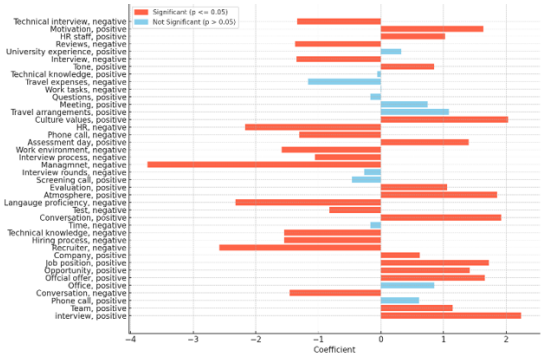


Figure 17: Logistic regression coefficients - gte-large-en-v1.5

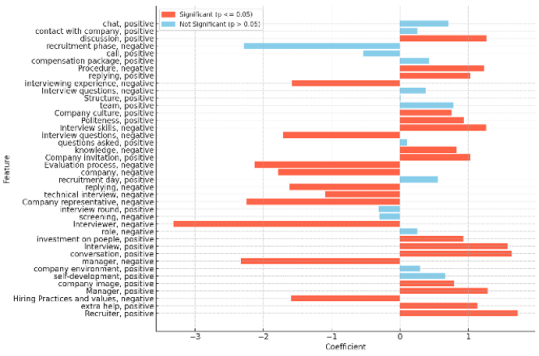


Figure 18: Logistic regression coefficients - all-minilm-l6-v2-f32

Figures 17 and 18 showcase the coefficients of the logistic regression models using the binary aspect scoring method to extract features. These features represent both aspects and sentiments, with interpretations considering the sentiment context, adding an additional layer of analysis.

In the model using the gte-large-en-v1.5 embedding, "management, negative" has the highest negative coefficient, indicating that negative mentions of management significantly decrease the odds of a positive candidate experience. This is expected, as poor management is a common cause of employee dissatisfaction, which candidates aim to avoid. Conversely, "interviews, positive" has the highest significant positive impact on the likelihood of a positive candidate experience, aligning with previous findings. Additionally, "Cultural value, positive" shows a significantly positive coefficient, emphasizing the importance of candidates aligning their values with the company's values.

In the logistic regression model using the all-minilm-l6-v2-f32 embedding, "interviewer, negative" has the highest negative impact on the likelihood of a positive candidate experience, consistent with the notion that the interviewer is a critical aspect of candidate experiences. Furthermore, the "evaluation process, negative" is a significant area of the recruitment process, evident by its notably negative coefficient.

An interesting finding from both tables is that negative aspect sentiments are associated with negative coefficients, while positive aspect sentiments are associated with positive coefficients. This demonstrates the effectiveness of combining ABSA with the binary aspect sentiment encoding method in accurately capturing the positive and negative aspects of textual candidate reviews.

5.5 Recruitment Outcome Prediction

Table 22 presents the results of logistic regression models used to predict whether a candidate will accept or reject a job offer. The models are based on the three different feature extraction methods from Aspect-Based Sentiment Analysis (ABSA) results.

	SENTIMENT SCORING		BINARY ASPECT SENTIMENT ENCODING		ASPECT SENTIMENT EMBEDDING	
	gte-large-en-v1	all-minilm-l6-v2-f32	gte-large-en-v1.5	all-minilm-l6-v2-f32	gte-large-en-v1.5	all-minilm-l6-v2-f32
ACCURACY	80%	80%	77%	78%	75%	73%
F1 SCORE – NEGATIVE CLASS	12%	17%	38%	39%	38%	36%
F1 SCORE – POSITIVE CLASS	89%	89%	86%	86%	84%	83%

Table 24: Logistic Regression Results Predicting Recruitment Outcome

Both embedding models using the sentiment scoring method achieve an accuracy of 80%. This is consistent with performance in predicting accepted offers as both models achieve an F1 score of 89%. However, while the method performs well in predicting accepted offers, the performance in predicting declined offers is very poor with F1 scores of 12% and 17%. Despite the relatively low testing sample size of candidates that declined their offers, the F1 scores are still significantly low, signifying the model's inability to correctly learn the relationships between the features of the declined offer instances.

In the binary aspect sentiment encoding method, an accuracy of 77% is achieved with the gte-large-en-v1.5 model and 78% with the all-minilm-l6-v2-f32 model. The F1 scores for the declined offer class are significantly higher with this method, at 38% and 39% respectively, however is still relatively low. For the accepted offer class, the F1 scores are 86% for both models, showing consistent strong performance predicting accepted offers.

The aspect sentiment embedding method achieves the lowest overall accuracy, with 75% for the gte-large-en-v1.5 model and 73% for the all-minilm-l6-v2-f32 model. The F1 scores for the declined offer class are 38% and 36% respectively, which is comparable to the Binary Aspect Sentiment Encoding method. For the accepted offer class, the F1 scores are slightly higher at 84% and 83%.

The Sentiment Scoring method achieves the highest overall accuracy but performs poorly in predicting declined offers, as evidenced by the low F1 scores for that class. The Binary Aspect Sentiment Encoding method provides a better balance, with improved F1 scores for the declined offer class and strong performance for the accepted offer class. The Aspect Sentiment Embedding method, while effective, shows the lowest overall accuracy and slightly lower F1 scores for both classes. The performance differences between the embedding models (gte-large-en-v1.5 and all-minilm-l6-v2-f32) are relatively minor, indicating that the choice of embedding model has less impact than the feature extraction method itself.

In terms of assessing how well the recruitment outcome prediction models fit the data, table 24 illustrates all the results for Psuedo R squared and AIC.

Model	Pseudo R squared Score	AIC
Sentiment Scoring - gte-large-en-v1.5	-0.10	4025
Sentiment Scoring - all-minilm-l6-v2-f32	-0.10	4063
Binary Aspect Sentiment Scoring Method - gte-large-en-v1.5	-0.06	3962
Binary Aspect Sentiment Scoring Method - all-minilm-l6-v2-f32	-0.08	4040
Aspect Sentiment Embedding - gte-large-en-v1.5	0.53	3719
Aspect Sentiment Embedding - all-minilm-l6-v2-f32	0.21	3533

Table 25: Pseudo R squared and AIC of logistic models

The Sentiment Scoring models using gte-large-en-v1.5 and all-minilm-l6-v2-f32 have pseudo R squared scores of -0.10 and -0.10, respectively. These negative scores suggest that these

models perform poorly in explaining the variability in the candidate experience variable. The Binary Aspect Sentiment Scoring Method models have similar scores, with the gte-large-en-v1.5 variant achieving a score of -0.06 and the all-minilm-l6-v2-f32 variant scoring -0.08, indicating marginally better explanatory power compared to the Sentiment Scoring models.

The Aspect Sentiment Embedding models demonstrate significantly higher pseudo R squared scores than the other models. The gte-large-en-v1.5 variant scores 0.53, while the all-minilm-l6-v2-f32 variant scores 0.21. These scores suggest that the gte-large-en-v1.5 embedding model explains a much more substantial portion of the variability in the candidate experience variable, with the all-minilm-l6-v2-f32 embedding model also showing reasonable explanatory power.

In terms of the AIC, the Sentiment Scoring models have AIC scores of 4025 for gte-large-en-v1.5 and 4063 for Minilm. The Binary Aspect Sentiment Scoring Method models have AIC scores of 3962 for gte-large-en-v1.5 and 4040 for Minilm, with the gte-large-en-v1.5 variant providing a better fit than the all-minilm-l6-v2-f32 variant. The Aspect Sentiment Embedding models have the lowest AIC scores, with the gte-large-en-v1.5 variant scoring 3719 and the all-minilm-l6-v2-f32 variant scoring 3533, indicating these models achieve a better balance between fit and complexity.

Overall, similar to candidate experience prediction tasks, the gte-large-en-v1.5 embedding model explains the variability in the candidate experience more effectively compared to the all-minilm-l6-v2-f32 embedding model. The Aspect Sentiment Embedding feature extraction method using the gte-large-en-v1.5 embedding model yields the highest pseudo R squared score and the lowest AIC score, suggesting it may be the best-performing model in terms of fitting the data. These results are consistent with the candidate experience prediction model, further supporting the idea that embedding-based approaches preserve a higher level of semantic and contextual information from aspect-based sentiment analysis (ABSA) than the other feature extraction methods that may lose information through aggregation transformations.

To evaluate the significance and impact of each feature on the prediction of the recruitment outcome, the coefficients are presented and analyzed below. Similar to the candidate experience prediction models, the focus will be on the coefficients from the sentiment scoring and binary aspect sentiment encoding features.

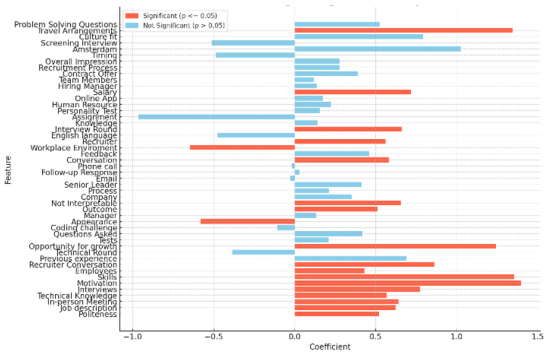


Figure 19: Logistic regression coefficients - gte-large-en-v1.5

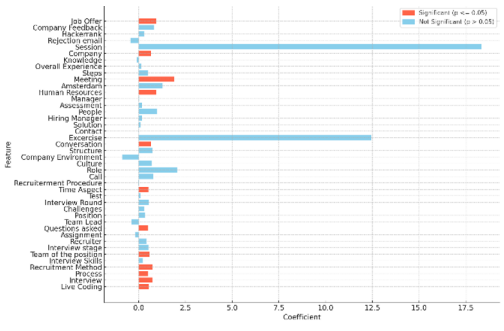


Figure 20: Logistic regression coefficients - all-minilm-l6-v2-f32

In the model using gte-large-en-v1.5 in Figure 19, travel arrangements has the highest significant positive coefficient, signifying the highest impact in increasing the likelihood of the candidate accepting the job offer. This is plausible given that travel arrangements is an aspect that is usually concerning final round candidate that are usually invited to the office rather than early stage candidates in the screening phase. Opportunities for growth and motivation are also important features and is consistent with the fact that in the final evaluation stage, candidates heavily weight future growth and personal motivation as key deciding factors for accepting an offer. Workplace environment and appearance have significantly negative coefficients aligning with the findings in the candidate experience prediction of a bad workplace environment being a significant deterrent for a candidate to join a company. In the model using all-minilm-l6-v2-f32 in Figure 20, meeting, human resources, and job offers have the highest positive coefficients that are statistically significant.

In comparing model results between both embedding models, the all-minilm-l6-v2-f32 embedding yields mostly statically insignificant features, with only 11 features being significant. Moreover, four anomalies with variables of extremely high coefficients are yielded, these are email, rejection email, resources and expenses and exercise. However, all the features are also statistically insignificant, and thus, no concrete conclusion can be derived from them. While the model using gte-large-en-v1.5 embeddings has evenly distributed coefficient magnitudes with no extreme values and mostly statistically significant features.

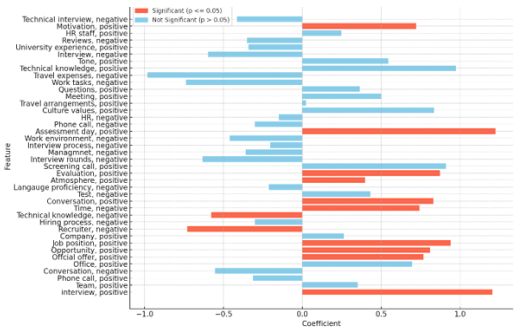


Figure 21: Logistic regression coefficients - gte-large-en-v1.5

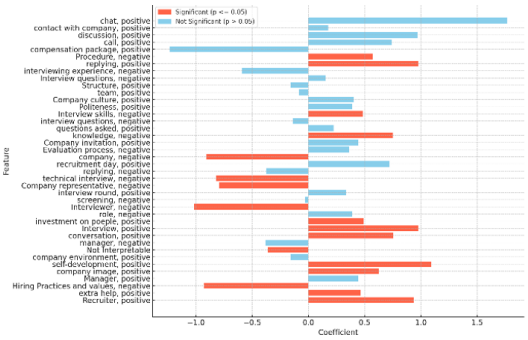


Figure 22: Logistic regression coefficients - all-minilm-l6-v2-f32

Based on the results from Figure 21 in the model using gte-large-en-v1.5 embedding, "assessment day, positive" is the highest statistically positive coefficient. The assessment day typically represents an advanced stage within the recruitment process, and a positive outcome in it for both the candidate and the company is highly likely to result in an offer given by the company and an offer being accepted by the candidate. This is also consistent with "interview, positive" and "Evaluation, positive" also having a high statistically positive coefficient. On the other hand, "recruiter, negative" and "technical knowledge, negative" are statistically significant and have negative coefficients with the highest magnitude. This is also consistent with findings in the candidate experience prediction model, as the recruiter and the technical abilities of the interviews seem to be important factors that deter candidates from accepting offers.

In the results from Figure 22, "recruiter, positive", "company environment" and "replying, positive" are the highest statistically significant positive coefficients and are similar

to the candidate experience model. An interesting aspect sentiment that achieves a highly significant coefficient is "self-development, positive". Self-development is typically a personal aspect to the candidate, and seeing it amongst the most important factors in a candidate accepting an offer further highlights the importance of learning and growth opportunities relative to monetary aspects such as salary and benefits. In the contrary, "interview, negative", "Hiring practices and values, negative" and "company, negative" are the highest impacting significant features that decrease the likelihood of a candidate accepting an offer.

5.6 Comparison with Baseline Models

To evaluate the accuracy of the final logistic regression models, we have so far focused on the novel feature extraction methods developed in this research. To provide a comprehensive assessment of these methods, it is essential to compare their predictive performance with traditional text vectorization techniques that are simpler and more intuitive, such as the bag of words (BoW) model and TF-IDF. This comparison aims to position the novel feature extraction methods within the context of established vectorization techniques, offering a clear understanding of their relative effectiveness. The goal of this comparison is to gain a deeper understanding of the added value the proposed feature extraction methods offer in terms of predictive accuracy. Therefore, a text classification analysis was conducted using logistic regression with the same parameters but using BoW, Binary BoW and TF-IDF applied directly to the candidate reviews as a basic feature extraction technique.

	Bag of Words	**Binary Bag of Words**	**TF-IDF**
Accuracy	76%	76%	78%
F1 Score – Negative Class	31%	34%	29%
F1 Score – Positive Class	86%	85%	87%

Table 26: Recruitment Outcome Prediction using BoW and TF-IDF

For the recruitment outcome prediction, an accuracy of 76% and 78% is achieved. An F1 score for the negative class ranges from 29% to 34%, and for the positive class, it ranges from 85% to 87%. These results are similar to the performance of the feature extraction methods from this study, exhibiting slightly worse accuracy than the sentiment scoring in terms of accuracy and F1 score for the positive class. As such, the improvement in accuracy using the ABSA feature extraction methods is only marginal and cannot be considered to be significant for this text classification task.

In terms of the candidate experience prediction, the highest accuracy achieved is 90% by the TF-IDF vectorizer. An F1 score for the negative class ranges from 74% to 80%, and for the positive class, it ranges from 89% to 93%. The TF-IDF vectorization on the reviews is able to achieve the same highest accuracy achieved by the binary aspect sentiment encoding and the aspect sentiment embedding approach. Hence, the ABSA feature extraction methods only marginally outperforms the BoW model and achieves the same accuracy as the TF-IDF vectorizer on candidate reviews.

	Bag of Words	Binary Bag of Words	TF-IDF
Accuracy	87%	85%	90%
F1 Score – Negative Class	77%	74%	80%
F1 Score – Positive Class	91%	89%	93%

Table 27: Candidate Experience Prediction using BoW and TF-IDF

In drawing comparisons between the feature extraction methods introduced by this study and traditional text feature extraction techniques, it can be established that the accuracy gains in text classification are marginal, if any. This outcome is positive, as the feature extraction methods introduced derive features from ABSA results instead of the raw reviews, yet still achieve comparable or better accuracy. The primary goal of introducing the feature extraction methods in this study is not based on achieving higher predictive accuracy. The significance of this lies in the high level of interpretability the ABSA feature extraction methods offer compared to BoW, TF-IDF, and other traditional text vectorization techniques. While interpretability and accuracy are common trade-offs in text classification tasks, the ABSA feature extraction methods introduced provide improved interpretability without sacrificing accuracy. This enhanced interpretability allows for a more nuanced understanding of the contributing factors in classification decisions, thereby facilitating better insights and actionable strategies, particularly in the context of recruitment and candidate experience optimization. By bridging the gap between detailed sentiment analysis and predictive modeling, these methods not only maintain robust performance but also offer a clearer, more transparent view of text classification tasks.

6 Discussion

This section will discuss the main insights from all the results. To address the main research question, we will discuss both technical insights into ABSA, features extraction, and predictive modeling and non-technical business-relevant insights into identifying opportunities to improve candidate experience and talent acquisition.

6.1 Technical Insights

ABSA was effective not only in capturing opinionated statements but also in identifying general statements, resulting in a large number of neutral aspect sentiments. However, this broad scope of extraction led to a highly dimensional dataset that required reduction through aggregation. Moreover, to align with the focus of this study, neutral aspect sentiment was not utilized in the sentiment scoring and binary aspect sentiment encoding since it would have resulted in extremely high dimensional data that would be very computationally intensive to reduce its dimensionality.

To manage the high dimensionality of the data, two embedding models, gte-large-en-v1.5 and all-minilm-l6-v2-f32, were used to assess the significance of the embedding method in the overall approach. K-means clustering was employed to aggregate the aspects; however, not all clusters were interpretable. Such aspects contained unrelated aspects that did not yield a common theme for creating an aspect label using the cluster centroid method. This occurred due to the presence of highly specific or niche aspects from ABSA and also random text and characters that were extracted as aspects that did not cluster well with other aspects. The all-minilm-l6-v2-f32 model resulted in significantly more non-interpretable clusters compared to gte-large-en-v1.5, highlighting the importance of selecting an effective embedding model in the clustering stage.

In the candidate experience prediction models, all feature extraction methods demonstrated good predictive power for candidate experience with the accuracy not going lower than 85%. The gte-large-en-v1.5 model performs slightly better using the binary aspect sentiment encoding and aspect sentiment embeddings, yielding an accuracy of 90%. For predicting whether a candidate will accept or reject a job offer, all methods again showed decent accuracy, although the performance was not as strong as in predicting candidate experience. This is especially the case with using embedding directly as features to predict if a candidate will accept the offer with an accuracy going as low as 73%. Notably, predicting if a candidate will decline an offer was particularly challenging, with lower F1 scores in this class. There was no significant difference between the embedding models for this task. The contrast between the candidate prediction performance and the recruitment outcome model performance may be attributed to the fact that candidate experience is more well-defined in terms of what areas would generally result in a bad experience. However, the decision to accept or reject an offer is very much dependent on the person and can be due to a multitude of reasons that data-driven methods cannot capture. As such, the decision to accept or reject a job offer is influenced by a multitude of factors beyond those captured in candidate reviews. For example, personal circumstances, competing job offers, compensation expectations, and long-term career goals can all play significant roles, making it more challenging to predict acceptance or rejection based solely on sentiment analysis.

Overall, based on the pseudo R squared and AIC scores, the gte-large-en-v1.5 embedding model explains the variability in the candidate experience more effectively compared to the all-minilm-l6-v2-f32 embedding model. The Aspect Sentiment Embedding feature extraction method using the gte-large-en-v1.5 embedding model yields the best-performing model in terms of goodness of fit. This may support the idea that embedding-based approaches preserve a richer, more nuanced representation of the ABSA results that captures semantic and contextual information more effectively than the other feature extraction methods that may lose information through aggregation transformations.

Furthermore, by highlighting the comparison with traditional text vectorization methods, BoW, Binary BoW, and TF-IDF, it can be concluded that the methods proposed in this study do not significantly increase predictive accuracy over simple text vectorization methods. However, as mentioned in the previous section, the aim of introducing the feature extraction methods in this study is not solely to achieve higher predictive accuracy. The importance of these methods lies in their superior interpretability compared to traditional text vectorization techniques like BoW and TF-IDF.

Lastly, while sentiment scoring and binary aspect sentiment encoding provided interpretable features, the binary aspect sentiment encoding method was particularly effective. It preserved the context in which the aspect was mentioned, making the features more informative and interpretable. In terms of aspect sentiment embedding, while it offered the highest accuracy as well in the candidate experience text classification, it does not yield interpretative features as the features consist of hundreds of dimensions in an embedding space providing the least interpretable model.

6.2 Recruitment Insights

Interview-related aspects emerged as focal points in all analyses, indicating that companies should heavily prioritize the conduct and experience of interviews. Ensuring that interviews are well-structured and positive can significantly enhance overall candidate satisfaction. This can involve ensuring interviews are well-structured, engaging, and time efficient, for instance. Well-structured interviews provide clarity and direction, helping candidates understand the process and what is expected of them. While engaging interviews, where candidates feel heard and valued, can significantly enhance overall satisfaction. Overall, it is paramount to ensure recruiters and interviewers are well-trained to conduct interviews and are knowledgeable in the domain of the position being interviewed for. This is key to ensure the candidate is able to grasp the requirements of the role and deeply understands the nature of the role itself.

Operational and communication-related aspects, such as prompt responses and quality feedback, are also critical. Companies need to be proactive in engaging with candidates and providing clear, constructive feedback to foster a positive experience. Additionally, personal aspects like motivation, self-development, and future career goals are essential. Ensuring that the job opportunity aligns with the candidates' personal goals and motivations can influence their decision to accept or reject a job offer.

Moreover, advanced-stage recruitment activities, such as assessment days and office visits, play a more significant role in candidates' decision-making processes than recruitment activities early on in the process. Companies should ensure these activities are well-organized and reflect positively on the company culture and work environment, as they

are likely to be decisive in the candidate's decision.

Overall, these insights are aimed to provide a comprehensive understanding of the key factors that influence candidate experiences and their decisions. By addressing identifying these as areas of improvement, companies can build a more targeted approach at improving their recruitment processes and enhancing their ability to attract and retain top talent.

7 Conclusion

This section will conclude the study by discussing the main results and answering the research question. Secondly, the limitations will discussed to highlight key areas where this research could be improved on. Lastly, recommendations for future research will be given.

7.1 Conclusion

This consisted of two main research goals, the first was to develop a framework for extracting features from ABSA results for it be used in text classification tasks and possibly other downstream analysis tasks. The second goal is more business and recruitment focused aiming to identify opportunities to improve candidate experience and talent acquisition. As such, this research aimed to answer the following research question:

How can features be extracted from aspect-based sentiment analysis for downstream prediction tasks to identify opportunities for improving candidate experiences and talent acquisition?

Three different methods were proposed to address the first goal. The sentiment scoring method, the binary aspect sentiment encoding method, and the aspect sentiment embedding method. The use of embedding models, clustering algorithms and text and data manipulation techniques were used in the development of these methods. These methods are evaluated based on final prediction model accuracy, goodness of fit and interpretability.

The best-performing method in terms of predictive accuracy for candidate experience was the binary aspect sentiment encoding using gte-large-en-v1 and the aspect sentiment embedding using gte-large-en-v1.5 as well. In terms of predicting the recruitment outcome the best performing method was the sentiment scoring method. However, the accuracy between all methods is not significantly different, and all yield good predictive performance. However, in evaluating the models in terms of interpretability, the binary aspect sentiment encoding provides the highest level of understanding as the contribution of an aspect to the prediction can be interpreted with the sentiment it was mentioned. Furthermore, comparing our methods with traditional text vectorization techniques like BoW, Binary BoW, and TF-IDF shows that while they don't significantly boost predictive accuracy, they greatly enhance the interpretability of text classification tasks in large review datasets.

In addressing the second research business and recruitment-focused goal, the results from the aspect-based sentiment analysis, cluster and embedding results, and logistic regression coefficients were used. Amongst the key findings is that companies should prioritize well-structured and engaging interviews, ensuring recruiters are well-trained and knowledgeable in the relevant domain to help candidates understand the role thoroughly. Secondly, poor operational and communication appeared as key aspects that could play a strong role in creating negative candidate experiences. Hence, companies should ensure prompt responses and quality feedback. Moreover, aligning job opportunities with candidates' personal goals and motivations is also essential. Advanced-stage recruitment activities, like assessment days and office visits, significantly influence candidates' final decisions and should be well-organized to positively reflect the company culture and work

environment.

Collectively, these insights answer the research question by demonstrating how features extracted from ABSA can be used to identify actionable opportunities for enhancing candidate experiences and optimizing talent acquisition processes. This research provides invaluable insights by contributing novel methods for feature extraction for ABSA and bridging the gap between academic contributions in ABSA and the development of business-relevant actionable insights.

Furthermore, this study provides valuable insights and contributions to other domains as well. The feature extraction methods developed in this study can be adapted to various industries and use cases where organizations can leverage the methodologies used on any textual data to enhance their processes and operations that are outcome-orientated. To mention a few examples, this could be used in the healthcare industry to analyze patient feedback for patient comfort, analyze current employee feedback to reduce turnover, or customer feedback to reduce churn on financial products. In summary, the feature extraction methods developed in this study provide a versatile and powerful toolkit for leveraging textual data across various industries to analyze aspect sentiment at scale for more granular insights to make more informed strategic decisions.

7.2 Limitations

Despite the promising results and insights from this research, several limitations must be acknowledged. First, the accuracy of the Aspect-Based Sentiment Analysis itself was not measured, which could affect the reliability of the extracted features. This is mainly due to the intensive data labeling process required and the difficulty in establishing a ground truth for aspect sentiments for candidate reviews. Hence, the creation of a supervised ABSA task was out of the scope of this research. Additionally, the clustering of unrelated aspects posed a challenge. The process of reducing very large dimensional data to 50 clusters has led to clustering unrelated aspects together, potentially diluting the specificity and relevance of the insights.

Another limitation is the removal of uninterpretable aspects. While necessary for clarity, this step resulted in the loss of potentially valuable information. Furthermore, the methodology does not account for company-specific factors. The types of candidates large corporations and startups attract are different in nature in terms of job expectations and personal motivations. This aspect was not taken into account and is and crucial for providing relevant and actionable insights that companies can action on based on their own needs.

Furthermore, the computational complexity associated with running Llama 3, and especially the embedding models, is significant, which may limit the scalability and practical application of this approach in real-world settings. Additionally, given the empirical approach involves multiple steps of using different models as components of a larger process, there is a need for concrete evaluation metrics for each step of the process, rather than relying solely on the final prediction performance. This would ensure a more thorough assessment of each component's contribution to the overall outcomes and the ability to troubleshoot the correct component if needed.

7.3 Future Research

As the goal of this research was to introduce a novel approach to extracting features from ABSA, it is crucial to consider how this work could shape future research directions in ABSA, NLP, and text review analysis. Therefore, several recommendations for future research are proposed to build upon these findings and drive further advancements in these fields.

Firstly, future research should focus on domain adaptation studies to evaluate the effectiveness of feature extraction methods across different domains. This involves testing the ABSA models and feature extraction techniques on datasets from various domains, such as customer reviews, employee reviews and potentially text from industries such as healthcare, finance, and retail. By identifying which domains these methods work best in, researchers can generalize the findings and enhance the practical applicability of the research. This research could also explore the challenges and solutions for transferring knowledge from one domain to another, thereby improving the versatility and robustness of the feature extraction methods and ABSA models.

Secondly, as this study has performed zero-shot ABSA without providing any learning instances for the model, another important area for future research is the exploration of few-shot learning approaches in the context of ABSA. Zero-shot learning involves evaluating the model's ability to generalize to unseen tasks without any task-specific training, while few-shot learning involves providing the model with a very small amount of task-specific data. By comparing the performance of ABSA models under zero-shot and few-shot conditions, researchers can determine the most effective strategies for deploying these models in real-world applications with limited data availability. This research could also explore the potential of transfer learning techniques to improve the models' performance where the initial training is done on a large, diverse dataset, and the model is then fine-tuned on a smaller, domain-specific dataset.

Thirdly, another avenue for future research is carrying out a comparative analysis of different ABSA models, with a particular focus on models like BERT. This analysis would involve evaluating non generative LLM BERT against generative LLMs on the same dataset to assess their performance in terms of accuracy, interpretability, and computational efficiency. BERT and its variants have shown significant promise in many natural language processing tasks, and comparing them with the generative LLMs, such as the Llama-3-70b model, could reveal valuable insights into conducting out-of-domain ABSA. On the other hand, exploring the impact of fine-tuning these models on domain-specific data and evaluating their impact could provide an interesting study on evaluating which language models require domain fine-tuning to perform reasonably in ABSA.

8 References

Acikgoz, Y. (2019). Employee recruitment and job search: Towards a multi-level integration. *Human resource management review*, *29*(1), 1–13.

Adams, B. (2016). 6 simple steps to revitalizing your candidate experience [Accessed: 2023-07-04]. https://www.ere.net/5-simple-steps-to-revitalizing-your-candidate-experience/

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., & Sanghai, S. (2023). Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*.

Akhtar, M. S., Garg, T., & Ekbal, A. (2020). Multi-task learning for aspect term extraction and aspect sentiment classification. *Neurocomputing*, *398*, 247–256.

Amsterdam Economic Board. (2017). *Fighting the odds: Amsterdam's approach to achieving a strong digital economy* (tech. rep.). https://www.amsterdameconomicboard.com/app/uploads/2017/09/Fighting-the-odds.pdf

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Backhaus, K., & Tikoo, S. (2004). Conceptualizing and researching employer branding. *Career Development International*, *9*(5), 501–517.

Balahur, A., Hermida, J. M., & Montoyo, A. (2012). Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, *53*(4), 742–753.

Balmer, J. M., & Gray, E. R. (2003). Corporate brands: What are they? what of them? *European Journal of Marketing*, *37*(7/8), 972–997.

Barber, A. E., Wesson, M. J., Roberson, Q. M., & Taylor, M. S. (1999). A tale of two job markets: Organizational size and its effects on hiring practices and job search behavior. *Personnel psychology*, *52*(4), 841–868.

Becker, W. J., Connolly, T., & Slaughter, J. E. (2010). The effect of job offer timing on offer acceptance, performance, and turnover. *Personnel Psychology*, *63*(1), 223–241.

Binder, M., Heinrich, B., Klier, M., Obermeier, A., & Schiller, A. P. R. (2019). Explaining the stars: Aspect-based sentiment analysis of online customer reviews.

Black, J. S., & van Esch, P. (2020). Ai-enabled recruiting: What is it and how should a manager use it? *Business Horizons*, *63*(2), 215–226.

Board, T. (2016). 2016 candidate experience report [Retrieved from http://www.thetalentboard.org/cand awards/cande-results-2016/]. http://www.thetalentboard.org/cande-awards/cande-results-2016/

Board, T. (2017). 2017 candidate experience report [Retrieved from https://www.sparcstart.com/wp-content/uploads/2018/03/2017-CandE-Report.pdf]. https://www.sparcstart.com/wp-content/uploads/2018/03/2017-CandE-Report.pdf

Breaugh, J. A. (2013). Employee recruitment. *Annual Review of Psychology*, *64*, 389–416.

Breaugh, J. A., & Starke, M. (2000). Research on employee recruitment: So many studies, so many remaining questions. *Journal of Management*, *26*(3), 405–434.

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, *35*(6), 953–975.

Collins, C. J., & Han, J. (2004). Exploring applicant pool quantity and quality: The effects of early recruitment practices, corporate advertising, and firm reputation. *Personnel Psychology*, *57*, 685–717.

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

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*.

Dineen, B. R., & Soltis, S. M. (2011). Recruitment: A review of research and emerging directions.

Fei, H., Li, B., Liu, Q., Bing, L., Li, F., & Chua, T.-S. (2023). Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*.

Feloni, R. (2017). Consumer goods giant unilever has been hiring employees using brain games and artificial intelligence-and it's a huge success.

Fulmer, I. S., Gerhart, B., & Scott, K. S. (2003). Are the 100 best better? an empirical investigation of the relationship between being a "great place to work" and firm performance. *Personnel Psychology*, *56*(4), 965–993.

Geva, M., Schuster, R., Berant, J., & Levy, O. (2020). Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PloS one*, *15*(5), e0232525.

Hirebee. (2023). Quality of hire metrics [Accessed: 2023-07-04]. https://hirebee.com/quality-of-hire-metrics

Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect-based sentiment analysis using bert. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 187–196.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. *Proceedings of the first workshop on social media analytics*, 80–88.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177.

Jamil, U. (2023). Llama explained: Kv-cache, rotary positional embedding, rms norm, grouped query attention, swiglu [Accessed: 2024-07-10].

Knox, S., & Freeman, C. (2006). Measuring and managing employer brand image in the service industry. *Journal of Marketing Management*, *22*(7-8), 695–716.

Korfiatis, N., Stamolampros, P., Kourouthanassis, P., & Sagiadinos, V. (2019). Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications*, *116*, 472–486.

Li, X., Bing, L., Lam, W., & Shi, B. (2018). Transformation networks for target-oriented sentiment classification. *arXiv preprint arXiv:1805.01086*.

Lievens, F., & Slaughter, J. E. (2016). Employer image and employer branding: What we know and what we need to know. *Annual Review of Organizational Psychology and Organizational Behavior*, *3*, 407–440.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415–463). Springer.

Luo, Y., Cai, H., Yang, L., Qin, Y., Xia, R., & Zhang, Y. (2022). Challenges for open-domain targeted sentiment analysis. *arXiv preprint arXiv:2204.06893*.

Ma, D., Li, S., Zhang, X., & Wang, H. (2017). Interactive attention networks for aspect-level sentiment classification. *arXiv preprint arXiv:1709.00893*.

Ma, Y., Peng, H., Khan, T., Cambria, E., & Hussain, A. (2018). Sentic lstm: A hybrid network for targeted aspect-based sentiment analysis. *Cognitive Computation*, *10*, 639–650.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, *1*(14), 281–297.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113.

Michaels, E., Handfield-Jones, H., & Axelrod, B. (2001). *The war for talent*. Harvard Business Press.

Miles, S. J., & Mangold, G. (2004). A conceptualization of the employee branding process. *Journal of Relationship Marketing*, *3*(2-3), 65–87.

Miles, S. J., & McCamey, R. (2018). The candidate experience: Is it damaging your employer brand? *Business Horizons*, *61*(5), 755–764.

Mitchell, M., Aguilar, J., Wilson, T., & Van Durme, B. (2013). Open domain targeted sentiment. *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1643–1654.

NLP, A. (2024). Gte-large-en-v1.5 [Accessed: 2024-07-09].

Onan, A., Korukoglu, S., & Bulut, H. (2016). Lda-based topic modelling in text sentiment classification: An empirical analysis. *International Journal of Computational Linguistics and Applications*, *7*(1), 101–119.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1-2), 1–135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

Phillips, J. M., & Gully, S. M. (2015). Multilevel and strategic recruiting: Where have we been, where can we go from here? *Journal of Management*, *41*(5), 1416–1445.

Ployhart, R. E., Schmitt, N., & Tippins, N. T. (2017). Solving the supreme problem: 100 years of selection and recruitment at the journal of applied psychology. *Journal of Applied Psychology*, *102*(3), 291.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., Clercq, O. D., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., & Eryigit, G. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 19–30.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). Semeval-2015 task 12: Aspect based sentiment analysis. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 486–495.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 27–35.

Poria, S., Cambria, E., Ku, L. W., Gui, C., & Gelbukh, A. (2014). A rule-based approach to aspect extraction from product reviews. *Proceedings of the second workshop on natural language processing for social media (SocialNLP)*, 28–37.

Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, *108*, 42–49.

Qiu, J., Liu, C., Li, Y., & Lin, Z. (2018). Leveraging sentiment analysis at the aspects level to predict ratings of reviews. *Information Sciences*, *451*, 295–309.

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

Recruitee. (2021). Time to hire and application drop-off rate [Accessed: 2023-07-04]. https://recruitee.com/time-to-hire-drop-off-rate

Reimers, N., & Gurevych, I. (2020). All-minilm-l6-v2 [Accessed: 2024-07-09].

Russell, S., & Brannan, M. J. (2016). "getting the right people on the bus": Recruitment, selection and integration for the branded organization. *European Management Journal*, *34*(2), 114–124.

Schwab, D., Rynes, S., & Aldag, R. (1987). Theories and research on job search and choice. In K. Rowland & G. Ferris (Eds.), *Research in personnel and human resources management*. JAI Press.

Selim, S. Z., & Ismail, M. A. (1984). K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on pattern analysis and machine intelligence*, (1), 81–87.

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and knn models for the text classification. *Augmented Human Research*, *5*(1), 12.

Shazeer, N. (2020). Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Skeeled. (2019). Cost per hire analysis [Accessed: 2023-07-04]. https://skeeled.com/cost-per-hire-analysis

Song, Y., Wang, J., Jiang, T., Liu, Z., & Rao, Y. (2019). Targeted sentiment classification with attentional encoder network. *Artificial Neural Networks and Machine Learning–ICANN 2019: Text and Time Series: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part IV*, 93–103.

Spence, M. (1978). Job market signaling. In *Uncertainty in economics* (pp. 281–306). Elsevier.

Sreenivasan, R. (2024, May). Meta llama 3 70b llm – zero shot aspect based sentiment analysis fast response groq collab demo [[Video]].

Stein, R. A., Jaques, P. A., & Valiati, J. F. (2019). An analysis of hierarchical text classification using word embeddings. *Information Sciences*, *471*, 216–232.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., & Liu, Y. (2024). Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, *568*, 127063.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, *63*(1), 163–173.

Trier, Ø. D., Jain, A. K., & Taxt, T. (1996). Feature extraction methods for character recognition-a survey. *Pattern recognition*, *29*(4), 641–662.

Turney, P. D. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*.

Van Esch, P., Black, J. S., & Ferolie, J. (2019). Marketing ai recruitment: The next phase in job application and selection. *Computers in Human Behavior*, *90*, 215–222.

Varia, S., Wang, S., Halder, K., Vacareanu, R., Ballesteros, M., Benajiba, Y., John, N. A., Anubhai, R., Muresan, S., & Roth, D. (2022). Instruction tuning for few-shot aspect-based sentiment analysis. *arXiv preprint arXiv:2210.06629*.

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

Vo, D. T., & Zhang, Y. (2015). Target-dependent twitter sentiment classification with rich automatic features. *Twenty-fourth international joint conference on artificial intelligence*.

Walker, H. J., Feild, H. S., Giles, W. F., & Bernerth, J. B. (2008). The interactive effects of job advertisement characteristics and applicant experience on reactions to recruitment messages. *Journal of Occupational and Organizational Psychology*, *81*(4), 619–638.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, *33*, 5776–5788.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824–24837.

Xue, W., & Li, T. (2018). Aspect based sentiment analysis with gated convolutional networks. *arXiv preprint arXiv:1805.07043*.

Yang, S., Jiang, X., Zhao, H., Zeng, W., Liu, H., & Jia, Y. (2024). Faima: Feature-aware in-context learning for multi-domain aspect-based sentiment analysis. *arXiv preprint arXiv:2403.01063*.

Ye, S., Hwang, H., Yang, S., Yun, H., Kim, Y., & Seo, M. (2024). Investigating the effectiveness of task-agnostic prefix prompt for instruction following. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(17), 19386–19394.

Zhang, B., & Sennrich, R. (2019). Root mean square layer normalization. *Advances in Neural Information Processing Systems*, *32*.

Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023). Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

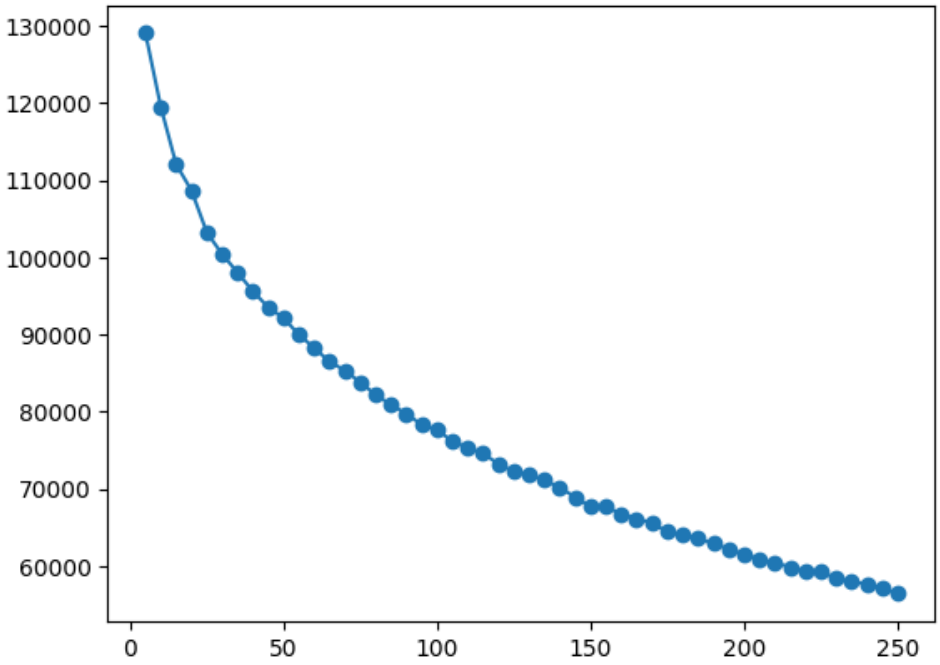
Zhao, A., & Yu, Y. (2021). Knowledge-enabled bert for aspect-based sentiment analysis. *Knowledge-Based Systems*, *227*, 107220.

Zhao, G., Luo, Y., Chen, Q., & Qian, X. (2023). Aspect-based sentiment analysis via multitask learning for online reviews. *Knowledge-Based Systems*, *264*, 110326.

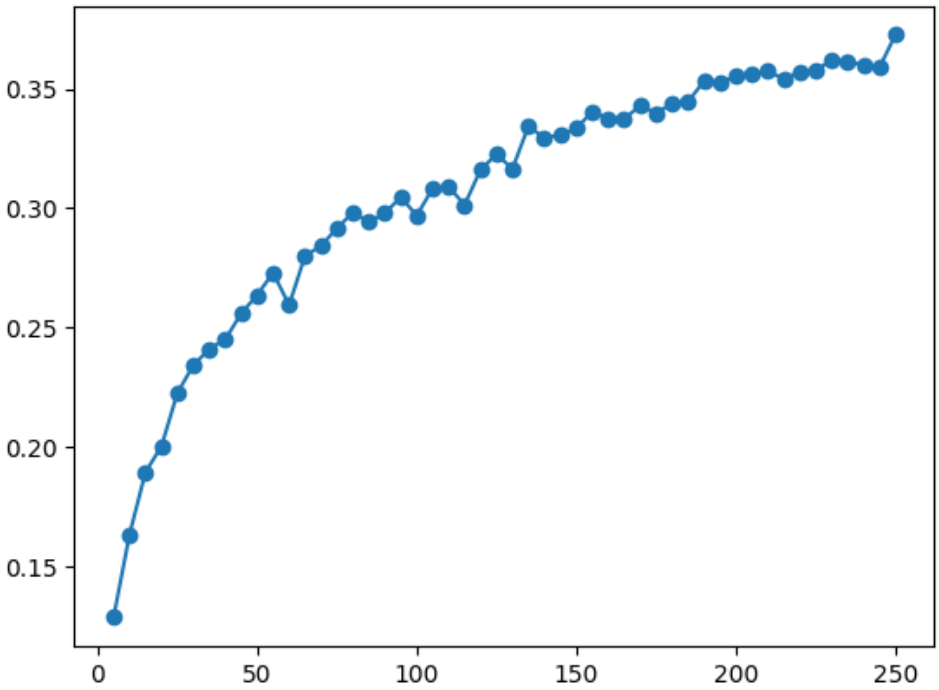
Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023). Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

9 Appendix

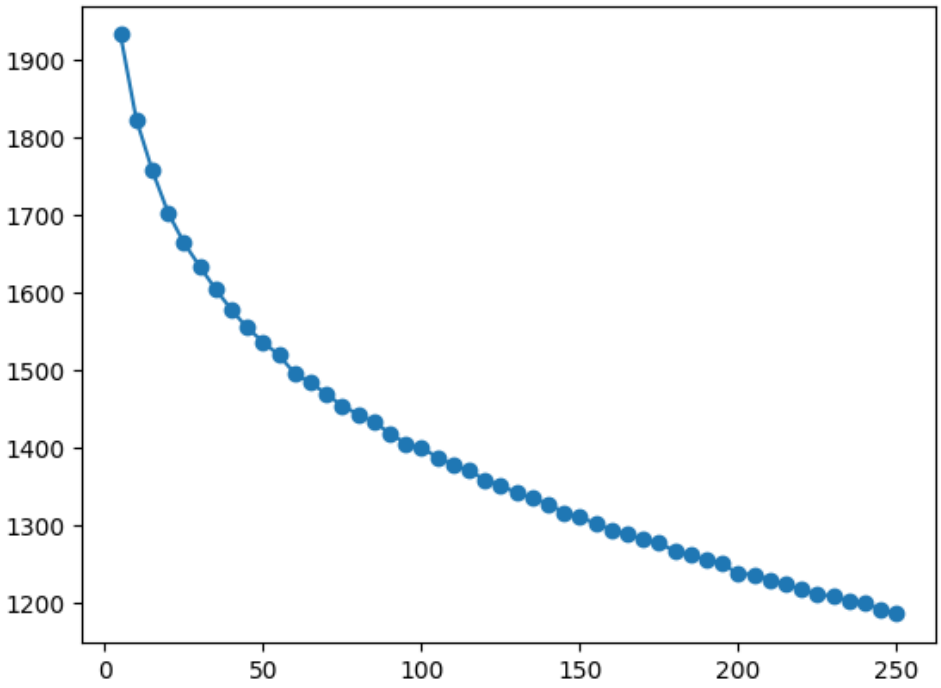
Appendix 1: K-means Interia Scores - Sentiment Scoring



Appendix 2: K-means Silhouette Score - Sentiment Scoring



Appendix 3: K-means Interia Scores - Binary Aspect Sentiment Encoding



Appendix 4: K-means Silhouette Score - Binary Aspect Sentiment Encoding