Aspect-Based Sentiment Analysis for Open-Ended HR Survey Responses

Lois Rink[*]
Universiteit van Amsterdam
Amsterdam, The Netherlands
`Lrink@hotmail.com`

Job Meijdam
Randstad Groep Nederland
Diemen, The Netherlands
`job.meijdam@randstadgroep.nl`

David Graus
Randstad
Diemen, The Netherlands
`david.graus@randstad.com`

Abstract

Understanding preferences, opinions, and sentiment of the workforce is paramount for effective employee lifecycle management. Open-ended survey responses serve as a valuable source of information. This paper proposes a machine learning approach for aspect-based sentiment analysis (ABSA) of Dutch open-ended responses in employee satisfaction surveys. Our approach aims to overcome the inherent noise and variability in these responses, enabling a comprehensive analysis of sentiments that can support employee lifecycle management. Through response clustering we identify six key aspects (salary, schedule, contact, communication, personal attention, agreements), which we validate by domain experts. We compile a dataset of 1,458 Dutch survey responses, revealing label imbalance in aspects and sentiments. We propose few-shot approaches for ABSA based on Dutch BERT models, and compare them against bag-of-words and zero-shot baselines. Our work significantly contributes to the field of ABSA by demonstrating the first successful application of Dutch pre-trained language models to aspect-based sentiment analysis in the domain of human resources (HR).

1 Introduction

Understanding employees' preferences and opinions can be of paramount importance in the full employee life cycle, e.g., from recruitment and selection to employee retention, and performance and career management (Bogers et al., 2022). In employee satisfaction surveys, open-ended questions may elicit a wide range of aspects. However, conducting large-scale analysis of these responses is challenging because of their user-generated nature.

Aspect-based sentiment analysis (ABSA) is the task of identifying and extracting sentiments toward specific aspects from free text, allowing a more detailed analysis of opinions (Pontiki et al., 2016), which means they can be a valuable tool for identifying specific areas of (dis)satisfaction.

Despite the promise of ABSA, limited research has explored its application beyond English (Nazir et al., 2020). Moreover, to our knowledge, ABSA has not been studied in employee satisfaction surveys. This study aims to bridge this research gap by studying ABSA on Dutch open-ended employee satisfaction survey responses. The central research question that we answer in this paper is:

RQ1 *How effective is a BERT-based machine learning model in extracting aspect-sentiments from Dutch open-ended responses of employee satisfaction surveys?*

To answer this question, we first aim to answer the following sub-questions:

RQ1.1 *How does the performance of few-shot classification using Dutch BERT models compare to bag-of-words based baselines in ABSA?*

RQ1.2 *How does the performance of few-shot classification using Dutch BERT models compare to a zero-shot classification baseline in ABSA?*

RQ1.3 *To what extent can an improvement in performance be achieved in few-shot aspect-based sentiment classification, by training on a data set enlarged through data augmentation?*

Prior work mainly focuses on ABSA in English microblogs and user reviews. Microblogs, like X, cover diverse topics using informal language with an assumed shared context. Reviews tend to adopt more formal styles and primarily revolve around specific products or services (Kumar, 2019).

Open-ended survey responses share similarities with microblogs in terms of writing and context, but have topic coverage narrowed by topics that affect employee (dis)satisfaction.

---

[*] Work done while on internship at Randstad Groep Nederland

This study compares the performance of Dutch pre-trained language models, BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020), for ABSA in a few-shot classification experiment. These BERT-based models leverage contextual information, which is advantageous for short texts with limited contextual cues (Chang et al., 2020). We assess their performance against zero-shot classification BERTje and RobBERT models and traditional bag-of-words models (Wu, 2020).

The specific contributions of our research are as follows:

1. Annotation of a data set of 1,458 open-ended survey responses for ABSA in Randstad, with publicly available annotation procedures and guidelines for future studies.

2. Development of an ABSA model for Dutch open-ended employee satisfaction survey responses, enabling automated aspect-sentiment extraction for efficient and accurate analysis.

2 Related Work

This section provides an overview of previous studies on aspect and sentiment classification, ABSA, and the Dutch BERT models we employ in this paper: BERTje and RobBERT.

2.1 Aspect classification

Recent studies have found transformer models' effectiveness in topic and aspect classification on short text through their ability to capture long-range dependencies and context. Chang et al. (2020) demonstrated fine-tuning a deep transformer network for extreme multi-label aspect classification in English. In contrast, Dadgar et al. (2016) utilized a combination of TF-IDF vectors and an SVM classifier for news article aspect detection, without requiring extensive training. Hu et al. (2021) demonstrated few-shot learning using prototypical networks for aspect classification is valuable when labelled data is scarce. Alternatively, zero-shot classification, as discussed by Yin et al. (2019), allows topic classification without specific training, relying solely on labelled data for validation.

2.2 Sentiment classification

Sentiment classification is a long-standing research focus. Jiménez-Zafra et al. (2017) used an SVM classifier for patient satisfaction categorization in Dutch and Spanish healthcare reviews. Karl and

Scherp (2022) demonstrated that larger transformer models, like RoBERTa, excel in sentiment classification over classic BERT models due to their ability to generalize to unseen data. These transformer models can also perform well in zero-shot multilingual sentiment classification, as shown by Tesfagergish et al. (2022). Dogra et al. (2021) illustrated the effectiveness of BERT-based models in few-shot sentiment classification. Between Dutch BERT models it was found that RobBERT outperforms BERTje in sentiment classification, credited to its enhanced training framework and a larger training corpus (De Bruyne et al., 2021).

2.3 Aspect-based sentiment analysis

Lin and He (2009) introduced joint aspect-sentiment analysis, combining LDA for aspect extraction and a polarity lexicon for sentiment classification in English movie reviews. However, Jiménez-Zafra et al. (2017) found this lexicon-based approach unsuitable for Dutch and Spanish.

ABSA gained prominence after SemEval-2014 task 4 (Pontiki et al., 2014), where researchers tackled identifying explicit terms or categories representing aspects of a target entity, and their polarities, in the context of restaurant and laptop reviews. Over time, the task expanded to encompass full and multilingual reviews.

De Clercq and Hoste (2016) attempted Dutch ABSA using SVMs, augmenting their bag-of-words (BoW) approach with semantic role labels, with limited success.

Recent advancements include Hoang et al. (2019) demonstrating the potential of BERT models in English ABSA, and Liao et al. (2021) improving performance with a RoBERTa-based model. Few-shot BERT classification and augmented training with BERT embeddings were explored by Hosseini-Asl et al. (2022). De Geyndt et al. (2022) employed a RobBERT model to extract features for SVM in ABSA, reporting better results than a full transformer-based approach for their pipeline.

ABSA in open-text survey responses was addressed by Cammel et al. (2020) and van Buchem et al. (2022) using techniques like LDA, rule-based methods, and non-negative matrix factorization with multilingual BERT. These studies primarily focused on patient survey questions in the healthcare domain. Additionally, Cammel et al. (2020) restricted to detecting a single aspect-sentiment per response, limiting broad insights. Both studies considered open-ended responses to different survey

questions simultaneously, some of which elicited one-word responses which provided insufficient context for successful aspect-sentiment extraction.

2.4 Transformer models for Dutch

Transformers have excelled in Dutch aspect and sentiment classification, with two dedicated to Dutch: BERTje and RobBERT.

BERTje, with 12 layers, a hidden size of 768, and 12 attention heads, is pre-trained on a diverse corpus encompassing Wikipedia, news articles, books, and web pages, enabling it to capture Dutch linguistic patterns and context (de Vries et al., 2019). RobBERT, a Dutch variant of RoBERTa, shares a similar architecture with BERTje, but benefits from a more extensive pre-training dataset, including Wikipedia, news articles, web pages, Dutch parliament debates, and social media (Delobelle et al., 2020).

Existing methods for Dutch sentiment analysis are suboptimal, requiring further exploration, especially in novel domains like HR surveys.

3 Methodology

This section outlines the development and evaluation of the proposed ABSA model for Dutch open-ended survey responses on employee satisfaction.

3.1 Data set

The data set used in this study comprises Dutch open-ended responses to survey questions conducted by *anonymized*. We derived a sample of $1,500$ responses, recorded between January 2019 and December 2022, through stratified sampling across three different sub-brands of Randstad.

An example of such a response which illustrates the challenging nature of ABSA in employee satisfaction survey responses is: *"Ik ben tevreden over mijn salaris. Ik mis wel een stukje persoonlijke aandacht."* ("I am satisfied with my salary. However, I do miss some personal attention.") Here, we distinguish two aspect-sentiment pairs; an employee expresses a positive sentiment toward their salary, yet a negative sentiment toward personal attention. We explain the range of identified aspects in Section 3.3.

3.2 Data Preparation

We filtered and anonymized our set of responses to ensure data quality and privacy. We excluded responses with less than 10 tokens, as manual inspection revealed how shorter responses often lacked

adequate contextual information for accurate aspect classification. We also excluded responses exceeding 512 characters to address computational constraints (Liao et al., 2021). The final average response length is 35.7 tokens (approximately 182 characters), with 10 tokens (39 characters) at a minimum, and 97 tokens (511 characters) at most.

To ensure anonymity of both respondents and individuals mentioned in responses, all personal information was removed and replaced by dummy variables by using the Dutch Named Entity Recognition (NER) SpaCy model and regular expressions. After identifying, we replaced person names with "Naam", email addresses with "Emailadres", and addresses with "Adres". After this, manual review corrected an additional 28 names missed by SpaCy.

3.3 Aspect selection

We preprocessed responses by retaining only nouns, proper nouns, and verbs because they convey the most informative content (Boguraev et al., 1999). We then applied lemmatization. Finally, we applied TF-IDF vectorization to represent each response. We then applied k-means clustering over these TF-IDF vectors, determining the optimal number of clusters at $k = 6$, using the elbow method.

We inspected the responses in each cluster. Table 1 shows the most important terms per cluster, indicated by the highest TF-IDF score. After analyzing these clusters, we found they roughly represent the aspects of *contact, salary, schedule, personal attention, communication,* and *agreements.* The cluster names were verified by two domain experts. We define each aspect below, accompanied by an illustrative example.

- **Contact**: refers to the extent to which an employee can get in touch with the agency, for example, by phone or email. - *It took a long time to receive a response from the contact person. I had to make multiple phone calls and send emails before getting a reply.*

- **Schedule**: is about scheduling, work hours, and days off and whether the agency is flexible in changing these. - *I appreciate receiving my schedule well in advance because it allows me to adjust my plans accordingly.*

- **Agreements**: relates to the arrangements made between the employee and the agency and whether these are upheld or not. - *I am*

satisfied with the schedule agreement because it allows me to take my kids to school.

- **Salary**: is about payment and any bonuses or extras such as travel expenses. Consider remarks about correct payment of salary, and the frequency of salary payment. *- I am happy that I can choose my own frequency of payment.*

- **Personal attention**: is about the extent to which the agency pays personal attention to the employee. It can include receiving feedback and receiving personal guidance. *- I feel valued as an employee, and my ideas and suggestions are listened to attentively.*

- **Communication**: refers to the way information is exchanged between the agency and the employee, and whether communication about important matters is timely. *- I am satisfied with the way in which I am informed about the changes that are happening within the company.*

The *contact* and *communication* aspects may appear similar but are distinct: both revolve around interaction, but *contact* pertains to the accessibility and responsiveness between employees and the agency, emphasizing ease of access and availability. *Communication* is about information exchange, including factors like clarity, completeness, and timeliness.

3.4 Annotation Study

To collect labelled data, we ran an annotation study with nine native Dutch-speaking trainees from Yacht, which is one of the staffing agencies within Randstad. We had each response in the set of 1,500 responses annotated by three annotators, i.e., each annotator annotated 500 responses.

Annotators attended an in-person session to familiarize themselves with the task and guidelines. Detailed written guidelines, encompassing the task's objective, annotation procedure, answering options, and category definitions, were provided, along with examples for each category. To address annotator bias, a preliminary sample of 20 responses was annotated and discussed before each annotator worked on their assigned batch. We have published an English translation of our annotation guidelines online.[1]

---

[1] https://anonymous.4open.science/r/AnnotationGuidelinesABSA-BB08/

We employed the Prodigy annotation tool (Montani and Honnibal, 2018) with a custom recipe for annotation purposes. Annotators individually reviewed responses, selected relevant aspects, and a binary (positive or negative) sentiment for each aspect. They could also select 'no topics' if none of the six aspects were discussed. Responses were presented in random order to ensure unbiased judgment, and disagreements among annotators were resolved through majority voting, relying on a fourth annotator re-annotating in cases of no consensus.

Our primary focus was to identify clear-cut positive and negative sentiments for actionable insights on employee satisfaction. To address conflicting sentiments toward the same aspect, an 'ignore' option was introduced. Sentences marked with 'ignore' were excluded, enabling the model to focus on identifiable sentiment patterns. In addition, as sentiment was modeled as a binary variable, for neutral sentiments or multiple sentiments toward a single aspect, annotators used the 'ignore' option to ensure data consistency (Hartmann et al., 2023).

3.4.1 Inter-Annotator Agreement

Reliability in assessing inter-annotator agreement (IAA) is crucial. In this study, we employed Fleiss' kappa to measure agreement among multiple annotators, an extension of Cohen's kappa for more than two annotators (Fleiss, 1971).

With an average kappa score of 0.537, we achieved a moderate level of IAA, which reflects reliable annotations (Dumitrache et al., 2015) considering the inherent language ambiguity and inter-annotator disagreement.

The kappa statistic can be strict, especially in a multi-label setting, as it does not reward partial overlaps between annotations. Upon examining disagreement cases, we observed that when two out of three annotators agreed on the exact annotation, 53.92% of the responses exhibited a partial overlap between the majority-vote annotation and the third one. Additionally, 206 annotations (13%) required re-annotation by a fourth annotator due to three annotators providing different answers.

A quantitative analysis investigated disagreement patterns among aspects. Out of 106 responses with disagreement, a notable pattern emerged regarding the 'communication' aspect. In these cases, two annotators selected 'no topics,' while one chose 'communication:NEG.' This suggests a lack of clear demarcation in defining 'communication,' particularly in negative discussions. A similar pat-

Cluster	contact		salary		schedule		personal attention		communication		agreements	
	contact	contact	krijgen	to receive	week	week	mens	human	communicatie	communication	gesprek	conversation
	persoon	person	komen	to come	dag	day	contact	contact	verlopen	to go	horen	to hear
Top 5 terms	opnemen	to pick up	willlen	to want	uur	hour	krijgen	to receive	contact	contact	evaluatie	evaluation
	vraag	question	vragen	to ask	maand	month	nummer	number	komen	to come	bellen	to call
	contactpersoon	contact person	jaar	year	krijgen	to receive	maken	to make	super	super	sollicitatie	job interview

Table 1: Six identified aspects obtained through clustering responses, with the top five terms with highest TF-IDF scores.

tern was observed for 'personal attention,' possibly due to its over-representation in the dataset. Additionally, aspects with the highest agreement also generated substantial disagreement, likely because of their high frequency in the dataset. We found similar patterns for aspects with low occurrence.

In conclusion, the annotation study resulted in a dataset of 1,500 responses. Among these, 42 responses were categorized as 'ignore' due to conflicting or neutral sentiments and were excluded. After their removal, the final dataset comprised 1,458 responses, with 267 discussing aspects positively and 1,091 featuring negative discussions. See Table 2 for a detailed distribution of aspect and sentiment labels.

Label	POS_Count	NEG_Count	total
agreements	8	67	75
communication	33	212	245
contact	57	155	212
personal attention	141	370	511
schedule	5	134	139
salary	23	153	176
no topics	0	0	376
total	267	1091	1734

Table 2: Distribution of aspects and sentiments. POS_Count indicates the number of positive occurrences in the data set; NEG_Count indicates the negatives.

3.5 Data Augmentation

Table 2 reveals significant label imbalance in both aspects and sentiments, e.g., 'personal attention' is disproportionately represented, occurring nearly five times more than 'agreements'. In addition, the majority (approximately 78.77%) of aspects have a negative sentiment.

To address this imbalance and prevent bias in the machine learning model, data augmentation was implemented using NLPaug (Ma, 2019). This approach involves generating contextual word embeddings using a BERT model and replacing some of the tokens in a sentence (Sarhan et al., 2022). For an example of an original sentence and its augmented version: *"De lonen zouden wel een keer flink omhoog mogen"* ("the wages could well

do with a substantial increase") is rephrased into *"De salarissen zullen tot twintig keer dik omhoog moeten."* Here, some tokens from the initial sentence were replaced with contextually analogous tokens to yield the augmented sentence. However, this carries the risk of creating ungrammatical augmented sentences, as in the example.

Data augmentation used RobBERT embeddings to increase label combination variety until each distinct label combination occurred at least 30 times. RobBERT was chosen for its wider training data set, enabling broader coverage of Dutch texts. Because the objective was to extract multiple aspect-sentiments from open-ended survey responses, augmentation focused on responses with two or more aspects. With a 30% set augmentation probability and a maximum of 50 tokens, data augmentation involved replacing up to 50 tokens in a response. The augmented responses were added to the training data set, as shown in Table 3, which displays the distribution of labels and sentiments across both the augmented and non-augmented training data set. Although Table 3 demonstrates the additional training samples improved balance, a slight imbalance remains.

Label	POS_Count		NEG_Count	
	original	augmented	original	augmented
agreements	6	200	43	516
communication	21	516	140	745
contact	28	432	111	803
personal attention	101	550	239	716
schedule	3	98	85	420
salary	13	380	102	539
total	172	2176	720	3739

Table 3: Aspect and sentiment distribution of the original versus the augmented training data set.

3.6 Model implementation

For aspect-based sentiment analysis of open-ended survey responses, we propose a two-tiered approach. The first step employs multi-label classification to determine the correct aspects for each response. The second utilizes the aspects identified by the first system as features, along with the response, and assigns binary sentiment labels to each aspect within the response. The adoption of a

two-tiered approach serves a dual purpose. Firstly, it allows the model to concentrate exclusively on aspect identification in its initial step. This deliberate isolation permits the model to specialize in autonomously recognizing aspects before undertaking sentiment classification. Moreover, the two-tiered framework is strategically designed to mitigate challenges associated with data sparsity in the dataset. Given that certain aspects may possess limited training examples, a 12-class multi-label classification experiment could potentially yield inadequate representations for specific aspects. This limitation may compromise the model's capacity to generalize beyond the training data and perform optimally across the entire spectrum of aspect-sentiment classes.

3.6.1 Baselines

For aspect and sentiment classification we employ support vector machines (SVM), multilayer perceptron (MLP), and two Dutch BERT models (BERTje and RobBERT) in a zero-shot classification setting as baselines. We selected BERTje and RobBERT for their success in similar tasks, and the advantage of pre-training on a larger corpus of Dutch texts (Cammel et al., 2020; van Buchem et al., 2022; De Geyndt et al., 2022).

We apply hyperparameter tuning on a validation set, relying on a 70/15/15 train/test/validation split. For SVM, we found C=1000 and gamma=0.01 to be the optimal hyperparameters. For MLP, we found ReLU activation, Adam solver, and a hidden layer size of (256, 128) as optimal hyperparameters. For BERTje and RobBERT, we did parameter tuning on the training set, since no actual training was done. For aspect classification, each open-ended response was paired with all possible aspects, generating six inputs per response. Tokenisation was performed using the BERTje and RobBERT model tokenisers, following the guidelines provided by De Vries et al. (2019) and Delobelle et al. (2020). The network produced a probability vector of length six, indicating the likelihood of each aspect's presence in the response. Predicted class probabilities were initially notably below 0.5, perhaps due to the models' lack of training on the target domain data, so we tuned classification thresholds through a grid search, resulting in thresholds of 0.45 for BERTje and 0.37 for RobBERT. Maintaining a 0.5 threshold would have led to numerous false negatives.

For sentiment classification, we followed the same approach for SVM and MLP. For SVM we applied a linear kernel to fit the binary nature of the task. The optimal C parameter was found at 10. For MLP, we found ReLU activation, Adam solver, and a hidden layer size of (128, 64) to be optimal parameters. The preprocessed data was passed through the network, and the model output was a two-element vector representing positive and negative sentiment classes, with the sentiment having the highest value assigned as the predicted sentiment.

3.6.2 Aspect classification

In the aspect classification task, we fine-tune BERTje and RobBERT in a few-shot setting. To maintain consistent input length, batches of 16 tokenized samples were generated. The data set was randomly shuffled before training to mitigate order bias. The neural network consisted of a 12-layered BERT model with a dropout layer (dropout=.3) for regularization, and as output layer a linear layer with six dimensions representing six aspects. As both models converged around epoch 10 and to avoid overfitting, we stopped training at 10 epochs (Yu et al., 2019), using Adam optimizer with a learning rate of 0.005. Binary cross-entropy loss was calculated separately for each class with sigmoid activation and network weights were updated based on the total loss. This sigmoid activation approach allows for independent and interpretable probability estimates for the presence of each aspect, facilitating a comprehensive multi-label classification strategy.

3.6.3 Sentiment classification

For sentiment classification, BERTje and RobBERT were used in a few-shot classification setting. Input responses were padded, tokenized, and shuffled using a batch size of 4. Categorical aspect features were encoded using an embedding layer, and their embeddings were concatenated with the BERT embeddings of the responses to generate distinct sentiment predictions for each aspect.

The neural network for sentiment classification consisted of a 12-layer BERT model followed by a ReLU layer for learning complex patterns (Goodfellow et al., 2016), with a dropout layer (dropout=.3), followed by a 2 dimensional linear layer for the binary sentiment labels.

The model's performance was evaluated using cross-entropy loss after each iteration. A training function trained the model for 10 epochs with a

learning rate of 0.005 using the Adam optimizer. For RobBERT's an BERTje's training loss and accuracy, we found they stabilized after the fourth epoch. However, we extended training to avoid premature stopping and underfitting (Yu et al., 2019). The loss continued to decrease until epoch 10, indicating no overfitting through extended training.

3.7 Evaluation Metrics

To evaluate aspect and sentiment classification tasks, we use the macro F1 score, which balances precision and recall and treats each category equally, mitigating the impact of larger classes. We also examine precision and recall to detect potential overfitting and underfitting as suggested by Sokolova and Lapalme (2009). The significance of the results is assessed using the Wilcoxon signed-rank test for aspect classification and McNemar's test for sentiment classification.

4 Results

In this section, we present the results of the experiments. We compare performance of $BERTje_{fewshot}$ and $RobBERT_{fewshot}$ to our traditional baselines (SVM and MLP), and zero-shot BERT baselines ($BERTje_{zeroshot}$ and $RobBERT_{zeroshot}$). In addition, we apply data augmentation for both ($BERTje_{fewshotDA}$ and $RobBERT_{fewshotDA}$).

4.1 Aspect classification

First, we turn to Table 4, which shows the performance of $BERTje_{zeroshot}$ and $RobBERT_{zeroshot}$. We note that, with 0.8793 recall and 0.1682 precision for $RobBERT_{zeroshot}$, and 0.9129 recall and 0.1568 precision for $BERTje_{zeroshot}$, both approaches show overprediction, which persists across all aspect categories.

The zero-shot models struggled to establish a reliable decision boundary for classifying aspects, indicating ineffective transfer of pre-training knowledge to novel data. The Dutch BERT models lacked domain-specific knowledge, aligning with prior findings for microblog texts (Chen et al., 2021).

Next, we compare the F1 scores of few-shot methods to all others, in Table 5. We see how $BERTje_{fewshot}$ and $RobBERT_{fewshot}$ at 0.5219 and 0.5449 respectively, significantly outperform all baselines ($p < 0.0001$). Between them, $RobBERT_{fewshot}$ significantly outperforms $BERTje_{fewshot}$ ($p < 0.0001$).

	precision		recall		f1-score	
	BERTje	RobBERT	BERTje	RobBERT	BERTje	RobBERT
agreements	0.0488	0.0784	1	0.9153	0.0928	0.1381
communication	0.1648	0.1623	0.9875	0.8441	0.2883	0.2662
contact	0.1478	0.1728	0.8394	0.8871	0.2575	0.2818
personal attention	0.3452	0.03591	0.892	0.8974	0.5084	0.513
schedule	0.0972	0.1072	0.8871	0.9167	0.1702	0.1983
salary	0.1283	0.1211	0.8954	0.8218	0.2148	0.2148
macro avg	0.1568	0.1682	0.9129	0.8793	0.2571	0.2671

Table 4: zero-shot aspect classification scores.

Our BoW baselines outperform the transformer-based zero-shot baselines, which suggest that the individual words captured by BoW models have a strong correlation with the aspects, and the contextual knowledge from BERT models may not offer sufficient information for distinguishing between aspects in open-ended survey responses.

We applied data augmentation to address label imbalance. $BERTje_{fewshotDA}$ achieves a lower F1 score (0.4982) than $BERTje_{fewshot}$ (0.5219), which suggests potential overfitting, or that the model gained limited novel information from augmented examples. However, $RobBERT_{fewshotDA}$ outperforms $RobBERT_{fewshot}$ with a significant increase in F1 score from 0.5449 to 0.6074 ($p = 0.017$). The performance improvement was particularly prominent in the 'agreements' category, which was underrepresented before augmentation. Nevertheless, data augmentation leads to decreased performance in some aspects.

Both zero-shot models suffer from overprediction, evidenced by high recall and low precision. This could be caused by a lack of knowledge from the target domain. This finding is supported by the outcomes of the few-shot classification experiment, where the significantly improved performance shows the model's improved capability to differentiate between the different aspect-classes, resulting in a higher macro F1 score. This illustrates the importance and benefits of fine-tuning, even when only a small amount of labelled target domain data is available.

4.2 Sentiment classification

Turning to sentiment classification results in Table 6, we see how $BERTje_{fewshot}$ with an F1 score of 0.8736 does not significantly outperform $RobBERT_{fewshot}$ at 0.8871 ($p = 0.292$). Both significantly outperform all baselines (all $p < 0.0001$). Again, BoW models outperform zero-shot models that seem to struggle to transfer contextual knowledge to the novel domain.

There seems to be no beneficial impact of data

	SVM	MLP	BERTje _zeroshot_	RobBERT _zeroshot_	BERTje _fewshot_	BERTje _fewshotDA_	RobBERT _fewshot_	RobBERT _fewshotDA_
agreements	0.2963	0.2222	0.0928	0.1356	0.2849	0.2871	0.1553	**0.4765**
communication	0.2857	0.1924	0.2883	0.2716	0.437	0.4298	0.3938	**0.4691**
contact	0.3619	0.4143	0.2575	0.2841	0.56	**0.5984**	0.5758	0.5546
personal attention	0.6025	0.5871	0.5084	0.5148	0.6324	0.7064	**0.7593**	0.6587
schedule	0.4	0.3238	0.1702	0.2019	0.4892	0.3581	0.6129	**0.7489**
salary	0.5227	0.4498	0.2148	0.2142	0.7329	0.6157	**0.7581**	0.6487
macro average	0.4115	0.3633	0.2571	0.2671	0.5219	0.4982	0.5449	**0.6074**

Table 5: Aspect classification performance in terms of F1 score for; best performing methods are boldfaced.

	SVM	MLP	BERTje _zeroshot_	RobBERT _zershot_	BERTje _fewshot_	BERTje _fewshotDA_	RobBERT _fewshot_	RobBERT _fewshotDA_
negative	0.9124	0.9024	0.2619	0.2938	0.9472	0.9423	**0.9576**	0.9482
positive	0.6228	0.5135	0.3348	0.3171	0.8	0.7912	**0.8314**	0.8186
macro average	0.7676	0.708	0.2983	0.3061	0.8736	0.8651	**0.8871**	0.8846

Table 6: Sentiment classification performance in terms of F1 score; best performing methods are boldfaced.

augmentation on sentiment classification, with neither BERTje$_{fewshotDA}$ nor RobBERT$_{fewshotDA}$ being able to outperform BERTje$_{fewshot}$ (0.8736 vs. 0.8651) and RobBERT$_{fewshot}$ (0.8871 vs. 0.8846).

5 Discussion

This study explores ABSA in Dutch employee satisfaction surveys, using Dutch BERT-based machine learning models. Our findings are in line with findings in prior research (Chang et al., 2020; Karl and Scherp, 2022) that highlight BERT's effectiveness for ABSA in English. Consistent with Karl and Scherp (2022), RobBERT outperforms BERTje, indicating the superiority of larger transformer models for sentiment classification, also observed by De Bruyne et al. (2021).

Additionally, our study underscores the success of few-shot classification in addressing limited labelled data, consistent with Hu et al. (2021) for aspect classification and Dogra et al. (2021) for sentiment classification in English. However, our findings contradict successful application of BERT models for zero-shot classification by Yin et al. (2019). This discrepancy in performance can be attributed to our domain-specific data, in contrast to the diverse dataset used by Yin et al. (2019).

Furthermore, this study identified significant label imbalance in aspects and their associated sentiments, as detailed in Section 4. To address this, we explored data augmentation using NLPaug (Ma, 2019) following Sarhan et al. (2022). However, this technique improved the macro F1 score for RobBERT in aspect classification only.

5.1 Limitations

Our dataset, comprising 1,458 responses, is relatively small which may affect its reliability. Despite time constraints, three annotators assessed each response for inter-annotator agreement. However, using larger and more diverse datasets can improve findings in future studies.

Using k-means clustering for aspect identification presents inherent limitations for internal and external validity. A substantial number of responses (376) didn't align with identified aspects, raising concerns about their reliability and comprehensiveness. Clustering does not support responses' potential membership of multiple clusters. Fuzzy clustering, as suggested by Zhao and Mao (2017), allows responses to belong to multiple clusters with varying membership degrees, offering a more comprehensive solution, at the cost of hindering clear boundaries and optimal cluster determination through the elbow method. Finally, using a supervised classification approach on an internal dataset provides accurate clustering results but reduces findings' transferability to emerging aspects, impacting external validity. Unsupervised clustering, as demonstrated by Cammel et al. (2020), may provide flexibility and adaptability to evolving contexts and domains.

To ensure high performance on future openended responses, further fine-tuning or retraining of the model with recent and relevant data is necessary.

6 Conclusion

Analyzing workforce opinions and preferences through aspect-based sentiment analysis has various HR applications. In this paper, we demonstrate the effectiveness of Dutch BERT models, BERTje and RobBERT, in a few-shot ABSA experiment using Dutch open-ended responses from employee satisfaction surveys. We address three sub-questions to gain insights into the models' performance and potential enhancements for aspect-based sentiment analysis (ABSA).

Regarding the first two sub-questions (RQ1.1 and RQ1.2), few-shot transformer models outperform baseline BoW models, significantly improving aspect-sentiment classification measured by macro F1 score. This emphasizes the importance of labeled data for fine-tuning BERT models, enhancing performance compared to relying solely on pre-trained knowledge in zero-shot scenarios. It also highlights BoW models' superior performance in leveraging individual words compared to zero-shot models struggling with domain transfer.

Regarding the third sub-research question (RQ1.3), this study demonstrates how data augmentation can enhance RobBERT's aspect classification performance. However, data augmentation does not improve aspect or sentiment classification for BERTje or sentiment classification for RobBERT. These findings suggest that the effectiveness of data augmentation varies across models and tasks.

In summary, regarding the main research question (RQ1) on BERT-based models' effectiveness in extracting aspect-sentiments, our findings demonstrate their superiority over traditional bag-of-words models and zero-shot classification approaches. To enhance model robustness, future studies should acquire larger and more diverse ABSA datasets, exposing models to varied open-ended survey responses for improved generalization to novel data. Considering the challenges posed by the limitations of traditional clustering methods, future studies could explore the incorporation of other clustering methods such as fuzzy clustering (Zhao and Mao, 2017) or semi-supervised (neural) topic modeling approaches (Chiu et al., 2022; Xu et al., 2023). Moreover, future studies should investigate the disparity in data augmentation success between BERTje and RobBERT. Specifically, exploring whether generating augmented sentences using BERTje embeddings improves BERTje model performance, similar to the favorable outcome observed for RobBERT in this study. Understanding such disparities would contribute to a deeper understanding of the relationship between specific models, their embeddings, and the efficiency of data augmentation techniques in ABSA's broader context.

Acknowledgements

References

Toine Bogers, David Graus, Mesut Kaya, Francisco Gutiérrez, Sepideh Mesbah, and Chris Johnson. 2022. Second workshop on recommender systems for human resources (recsys in hr 2022). In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, page 671–674, New York, NY, USA. Association for Computing Machinery.

Branimir Boguraev, Rachel Bellamy, and Christopher Kennedy. 1999. Dynamic presentation of phrasally-based document abstractions. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers*, pages 10–pp. IEEE.

Simone A Cammel, Marit S De Vos, Daphne van Soest, Kristina M Hettne, Fred Boer, Ewout W Steyerberg, and Hileen Boosman. 2020. How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (nlp) approach. *BMC medical informatics and decision making*, 20(1):1–10.

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S Dhillon. 2020. Taming pre-trained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3163–3171.

Qi Chen, Wei Wang, Kaizhu Huang, and Frans Coenen. 2021. Zero-shot text classification via knowledge graph embedding for social media data. *IEEE Internet of Things Journal*, 9(12):9205–9213.

Jeffrey Chiu, Rajat Mittal, Neehal Tumma, Abhishek Sharma, and Finale Doshi-Velez. 2022. A joint learning approach for semi-supervised neural topic modeling. In *Proceedings of the Sixth Workshop on Structured Prediction for NLP*, pages 40–51, Dublin, Ireland. Association for Computational Linguistics.

Seyyed Mohammad Hossein Dadgar, Mohammad Shirzad Araghi, and Morteza Mastery Farahani. 2016. A novel text mining approach based on tf-idf and support vector machine for news classification.

In *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, pages 112–116. IEEE.

Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. 2021. Emotional robbert and insensitive bertje: combining transformers and affect lexica for dutch emotion detection. In *Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), held in conjunction with EACL 2021*, pages 257–263. Association for Computational Linguistics.

Orphée De Clercq and Véronique Hoste. 2016. Rude waiter but mouthwatering pastries! an exploratory study into dutch aspect-based sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2910–2917.

Ellen De Geyndt, Orphée De Clercq, Cynthia Van Hee, Els Lefever, Pranaydeep Singh, Olivier Parent, and Veronique Hoste. 2022. Sentemo: A multilingual adaptive platform for aspect-based sentiment and emotion analysis. In *12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, collocated with ACL 2022*, pages 51–61. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. arXiv:1912.09582.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

Varun Dogra, Aman Singh, Sahil Verma, Kavita, NZ Jhanjhi, and MN Talib. 2021. Analyzing distilbert for sentiment classification of banking financial news. In *Intelligent Computing and Innovation on Data Science: Proceedings of ICTIDS 2021*, pages 501–510. Springer.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2015. Achieving expert-level annotation quality with crowdtruth. In *Proc. of BDM2I Workshop, ISWC*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd nordic conference on computational linguistics*, pages 187–196.

Ehsan Hosseini-Asl, Wenhao Liu, and Caiming Xiong. 2022. A generative language model for few-shot aspect-based sentiment analysis. *arXiv preprint arXiv:2204.05356*.

Mengting Hu, Shiwan Zhao, Honglei Guo, Chao Xue, Hang Gao, Tiegang Gao, Renhong Cheng, and Zhong Su. 2021. Multi-label few-shot learning for aspect category detection. *arXiv preprint arXiv:2105.14174*.

Salud María Jiménez-Zafra, M Teresa Martín-Valdivia, Isa Maks, and Rubén Izquierdo. 2017. Analysis of patient satisfaction in dutch and spanish online reviews. *Procesamiento del Lenguaje Natural*, 58:101–108.

Fabian Karl and Ansgar Scherp. 2022. Transformers are short text classifiers: A study of inductive short text classifiers on benchmarks and real-world datasets. *arXiv preprint arXiv:2211.16878*.

H. Kumar. 2019. Sentiment analysis on imdb movie reviews using hybrid feature extraction method. *Int. J. Interact. Multim. Artif. Intell.*

Wenxiong Liao, Bi Zeng, Xiuwen Yin, and Pengfei Wei. 2021. An improved aspect-category sentiment analysis model for text sentiment analysis based on roberta. *Applied Intelligence*, 51:3522–3533.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence*, to appear.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2020. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, 13(2):845–863.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics.

Maria Pontiki, Haris Papageorgiou, Dimitrios Galanis, Ion Androutsopoulos, John Pavlopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. *SemEval 2014*, page 27.

25

Injy Sarhan, Pablo Mosteiro, and Marco Spruit. 2022. Uu-tax at semeval-2022 task 3: Improving the generalizability of language models for taxonomy classification through data augmentation. *arXiv preprint arXiv:2210.03378*.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Senait Gebremichael Tesfagergish, Jurgita Kapočiūtė-Dzikienė, and Robertas Damaševičius. 2022. Zero-shot emotion detection for semi-supervised sentiment analysis using sentence transformers and ensemble learning. *Applied Sciences*, 12(17):8662.

Marieke M van Buchem, Olaf M Neve, Ilse MJ Kant, Ewout W Steyerberg, Hileen Boosman, and Erik F Hensen. 2022. Analyzing patient experiences using natural language processing: development and validation of the artificial intelligence patient reported experience measure (ai-prem). *BMC Medical Informatics and Decision Making*, 22(1):1–11.

Jheng-Long Wu. 2020. Identifying emotion labels from psychiatric social texts using a bi-directional lstm-cnn model. *IEEE Access*.

Weijie Xu, Xiaoyu Jiang, Srinivasan Sengamedu Hanumantha Rao, Francis Iannacci, and Jinjin Zhao. 2023. vONTSS: vMF based semi-supervised neural topic modeling with optimal transport. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4433–4457, Toronto, Canada. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Fuxun Yu, Zhuwei Qin, Chenchen Liu, Liang Zhao, Yanzhi Wang, and Xiang Chen. 2019. Interpreting and evaluating neural network robustness. *arXiv preprint arXiv:1905.04270*.

Rui Zhao and Kezhi Mao. 2017. Fuzzy bag-of-words model for document representation. *IEEE transactions on fuzzy systems*, 26(2):794–804.