sciendo Journal of Social and Economic Statistics

Vol. 14, No. 1, Summer 2025
***DOI**: 10.2478/jses-2025-0002*

OPEN ACCESS

Exploring the Efficacy and Bias of Sentiment Analysis Tools in Customer Reviews

Dumitru Alexandru MARA[1*], Lia Cornelia CULDA[2], Marian Pompiliu CRISTESCU[3], Raluca Andreea NERIȘANU[4], Radu-Anton MOLDOVAN[5]

[1, 2, 3, 4, 5] Lucian Blaga University of Sibiu, Romania

Abstract

Sentiment analysis is an important technique for understanding public sentiment in customer feedback. This study aims to compare the effectiveness and reliability of BERT and VADER in analysing customer reviews from an e-commerce platform. We analyzed mean sentiment scores and standard deviations for each tool, using Spearman's Rho to assess correlations between VADER and BERT results due to non-normal score distributions. Our findings reveal that VADER predominantly categorizes reviews as Neutral, with limited Positive and Negative classifications, while BERT shows a bias towards Negative sentiments with some Positive and no Neutral classifications. A weak positive correlation indicates limited agreement between the two tools. Additionally, neither tool showed a strong correlation with actual review ratings, highlighting challenges in accurately capturing nuanced sentiments. The study discusses potential biases inherent in each tool's methodology – VADER's lexicon-based approach may oversimplify sentiments, and BERT's deep learning architecture may introduce a negative bias. These findings suggest that combining VADER's efficiency with BERT's contextual sensitivity – for instance, through hybrid models – can mitigate individual tool limitations and offer a more robust sentiment analysis framework for e-commerce businesses. By integrating these insights, businesses can optimize customer feedback systems, refine marketing strategies, and improve product development processes

Keywords: customer relationship management, sentiment analysis, BERT, VADER, bias

JEL Classification: C63

Introduction

In the digital age, customer reviews serve as an useful feedback mechanism, offering valuable insights into consumer experiences, preferences, and satisfaction levels. Analyzing these reviews systematically allows businesses to gauge public sentiment, identify areas for improvement, and tailor their offerings to meet market demands effectively. Sentiment analysis, a subfield of natural language processing, has emerged as a powerful tool for interpreting the vast amounts of unstructured textual data generated by online reviews.

* Corresponding author, Dumitru Alexandru MARA – dumitrualexandru.mara@ulbsibiu.ro

By quantifying emotions and opinions expressed in customer feedback, sentiment analysis facilitates data-driven decision-making and enhances customer relationship management.

Despite the growing adoption of sentiment analysis in business intelligence, the efficacy and reliability of various analytical tools remain subjects of ongoing investigation. Among the most prominent tools are VADER (Valence Aware Dictionary and Sentiment Reasoner), a lexicon-based approach tailored for social media text, and BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art machine learning model renowned for its contextual understanding of language. While both tools are widely utilized for sentiment classification, their differing methodologies may introduce distinct biases and performance variances, particularly in diverse application contexts such as e-commerce. This study aims to critically evaluate and compare the effectiveness and inherent biases of VADER and BERT in analyzing customer reviews from the Vinted platform. Our main findings indicate that VADER predominantly categorizes reviews as Neutral, whereas BERT shows a bias towards Negative sentiments, with both tools exhibiting limited correlation with actual review ratings. By examining sentiment scores, distribution patterns, and correlations with actual review ratings, this research seeks to provide actionable insights for businesses in selecting the most appropriate sentiment analysis tool to align with their specific analytical needs and objectives.

This study addresses the following research questions: (1) How do VADER and BERT compare in capturing customer sentiment on an e-commerce platform? (2) What inherent biases does each tool exhibit? (3) How can a hybrid approach enhance sentiment analysis performance in real-world applications? Understanding these differences not only advances academic research but also provides actionable insights for businesses seeking to fine-tune their customer feedback strategies

1. Literature review

1.1. *The role of sentiment analysis in customer reviews feedback analysis*

In today's highly competitive market landscape, understanding customer sentiment is paramount for businesses aiming to enhance customer satisfaction, loyalty, and overall market performance. Customer reviews, predominantly available through online platforms, provide a rich source of unstructured textual data that reflects consumer opinions, preferences, and experiences (Liu et al., 2020). Sentiment analysis, also known as opinion mining, has emerged as a vital computational technique for extracting and quantifying the emotional tone embedded within these reviews (Cambria et al., 2013; Pang & Lee, 2008). By systematically analyzing large volumes of customer feedback, organizations can derive actionable insights that inform strategic decisions, optimize product offerings, and tailor marketing initiatives to better align with consumer expectations (Nassirtoussi et al., 2014).

The integration of sentiment analysis into customer feedback mechanisms enables businesses to transform qualitative data into quantitative metrics, facilitating the identification of prevailing sentiment trends and the detection of underlying factors influencing customer perceptions (Stancu et al., 2020). This capability is particularly crucial in the context of big data, where the sheer volume and velocity of customer reviews necessitate automated and scalable analytical solutions (Atanasova et al., 2010). Moreover, sentiment analysis contributes to the enhancement of decision support systems by providing nuanced insights that complement traditional business intelligence tools, thereby fostering a more comprehensive understanding of market dynamics and consumer behavior (Bâra & Lungu, 2012).

1.2. *Key tools for sentiment analysis*

1.2.1. *Lexicon-based approaches*

Lexicon-based sentiment analysis tools utilize predefined dictionaries that associate specific words with sentiment scores, enabling the classification of text into positive, negative, or neutral categories based on the aggregate sentiment of individual words (Loria, 2022). One notable lexicon-based tool is the Valence Aware Dictionary for Sentiment Reasoning (VADER), which is particularly adept at handling the informal and nuanced language often found in social media and customer reviews(Hutto & Gilbert, 2014). VADER assigns sentiment scores to words and considers factors such as punctuation, capitalization, and degree modifiers to accurately capture the sentiment intensity (Aljedaani et al., 2022). This makes VADER a robust tool for real-time sentiment analysis in dynamic online environments.

Similarly, TextBlob is another popular lexicon-based tool that leverages the Natural Language Toolkit (NLTK) and the Pattern library to perform sentiment classification (Loria, 2022). TextBlob assigns polarity scores ranging from -1 (indicating negative sentiment) to +1 (indicating positive sentiment), facilitating straightforward sentiment categorization. Despite their simplicity and ease of implementation, lexicon-based methods like VADER and TextBlob may encounter challenges in accurately interpreting context-dependent sentiments and complex linguistic constructs, potentially leading to oversimplified or biased sentiment assessments (Kotelnikova et al., 2022).

1.2.2. *Machine learning-based approaches*

In contrast to lexicon-based methods, machine learning-based sentiment analysis approaches leverage advanced models to capture the contextual and semantic intricacies of language. Bidirectional Encoder Representations from Transformers (BERT) exemplifies a state-of-the-art model that has significantly advanced the accuracy and depth of sentiment analysis (Devlin et al., 2019). BERT employs a transformer architecture that processes text bidirectionally, enabling it to understand the context of each word based on its surrounding words (Vaswani et al., 2017). This bidirectional understanding allows BERT to effectively handle nuanced sentiment expressions and contextual dependencies, surpassing the capabilities of traditional lexicon-based tools in capturing subtle sentiment cues (Sun et al., 2019).

BERT has been fine-tuned for various sentiment analysis tasks, demonstrating superior performance in identifying complex linguistic patterns and subtle sentiment shifts within customer reviews (Xu et al., 2020). For instance, in financial news sentiment analysis, BERT has shown a higher sensitivity to negative sentiments compared to TextBlob, highlighting its ability to discern contextual nuances that simpler models might overlook (Cristescu et al., 2024). However, the sophistication of BERT comes with increased computational requirements and complexity, which may pose challenges for real-time or resource-constrained applications (Li et al., 2019).

1.2.3. *Hybrid and advanced models*

Recent advancements have seen the development of hybrid models that combine the strengths of lexicon-based and machine learning-based approaches to enhance sentiment analysis accuracy and reliability. Models such as DomBERT integrate BERT's contextual understanding with domain-specific linguistic features to improve aspect-based sentiment analysis, thereby enabling more precise sentiment categorization tailored to specific business contexts (Li et al., 2019; Wu et al., 2019; Xu et al., 2020). These hybrid approaches aim to mitigate the limitations of individual models by leveraging their complementary strengths, resulting in more robust and nuanced sentiment assessments.

Hybrid models, which integrate the speed of lexicon-based methods with the nuanced contextual understanding of machine learning models, have shown promise in achieving balanced sentiment analysis. Such approaches can dynamically adjust to the complexity of customer language, thereby mitigating the oversimplification of lexicon-based tools and the potential bias in deep learning models.

1.2.4. *Efficacy and bias of sentiment analysis tools*

VADER, a lexicon-based model designed for sentiment analysis on social media, emphasizes simplicity and effectiveness for informal texts. Its rule-based structure allows it to process sentiments by considering linguistic factors like capitalization and punctuation, producing high F1 scores across various contexts, sometimes even surpassing human accuracy (Hutto & Gilbert, 2014). VADER is particularly effective for real-time analysis, though it often struggles with nuanced sentiments or context-specific terms that require a deeper linguistic understanding (HarishRao et al., 2017).

In contrast, BERT, a transformer-based model, uses bidirectional contextual understanding, allowing it to capture complex and subtle sentiments in text. Studies indicate that BERT consistently outperforms traditional lexicon-based tools in accuracy, recall, and precision due to its advanced processing of contextual dependencies. For instance, on datasets like Stack Overflow posts, BERT showed a significant improvement in F1 scores, underscoring its applicability in domain-specific sentiment analysis (Biswas et al., 2020). However, BERT's complexity also demands greater computational resources, making it less suited for scenarios requiring lightweight and immediate sentiment assessments (Marfani et al., 2022).

The simplicity of VADER's lexicon-based approach, while beneficial for general sentiment categorization, can introduce biases due to its reliance on predefined dictionaries. VADER's tendency to overemphasize common sentiment terms may lead to skewed sentiment ratings, particularly in datasets requiring nuanced understanding (He & Zheng, 2019).

BERT, though more contextually adaptive, also faces challenges with bias, especially when fine-tuned on datasets that reflect societal prejudices. For example, in detecting biases related to disabilities, BERT and similar models exhibited systematic bias, suggesting a need for further adjustments in model training to ensure fair sentiment classification (Venkit et al., 2023). BERT's higher sensitivity to context, while enhancing accuracy, can inadvertently amplify pre-existing biases in training data, thus impacting sentiment assessment in sensitive or domain-specific texts.

Both VADER and BERT have their strengths and limitations in sentiment analysis, with VADER excelling in simplicity and real-time capabilities, and BERT offering nuanced, context-aware analysis at the expense of computational efficiency. While BERT provides more accurate sentiment categorization, both tools demonstrate biases that must be addressed for balanced sentiment analysis in various applications.

1.2.5. *Addressing bias and enhancing efficacy in sentiment analysis*

Advancements in sentiment analysis have highlighted critical issues with model biases and the need for improving analytical efficacy. This section discusses contemporary strategies for mitigating bias and enhancing the effectiveness of sentiment analysis.

Bias in sentiment analysis models, often introduced during training on unbalanced datasets, can lead to skewed outcomes. Methods like runtime bias detection are emerging as viable solutions. For instance, BiasRV detects and flags biased predictions by creating "mutants" of the input text, altering attributes like gender and verifying if the model responds consistently. This approach improves fairness without requiring access to the model's internals, addressing gender biases effectively (Yang et al., 2021).

In addition, tools such as BiasHeal offer post-processing bias correction for third-party sentiment analysis systems by adjusting biased predictions based on external fairness metrics. This approach is especially useful for users who cannot modify or retrain proprietary sentiment models (Yang et al., 2021). Such solutions are vital for creating fairer, more robust models across various demographics.

Efficacy in sentiment analysis can be improved by integrating techniques that address domain-specific challenges. SA-Q is a tool that assesses intra-tool and inter-tool inconsistencies, guiding users toward the most reliable sentiment tool based on their dataset type. It also offers inconsistency resolution for improved sentiment prediction accuracy (Kouadri et al., 2022).

Moreover, advances in pre-processing, such as refining lexicon-based approaches with adaptive lexicons and neutral bias modules, can significantly boost accuracy. Experiments show that tailored lexicon frameworks reduce sentiment misclassification, especially in lexicon-based tools that traditionally struggle with context-dependent terms (Moussa et al., 2018).

While prior studies have demonstrated the strengths of these models in controlled settings, few have addressed the challenges posed by real-world, imbalanced datasets – highlighting a gap that the present study aims to fill. Also, despite extensive research on individual sentiment analysis methods, there remains a significant gap in direct comparative studies within real-world e-commerce settings. This study thus aims to bridge that gap by evaluating the efficacy and biases of VADER and BERT on customer reviews from Vinted.

VADER and BERT were selected due to their widespread adoption and their fundamentally different approaches – VADER's lexicon-based simplicity versus BERT's contextual deep learning – which facilitates a clear comparative analysis.

2. Methodology

This study employs a comparative experimental design to systematically evaluate the performance and bias of two sentiment analysis tools on a common dataset.

In this study, we aimed to evaluate and compare the effectiveness and reliability of two prominent sentiment analysis tools, VADER (Valence Aware Dictionary and sEntiment Reasoner) and BERT (Bidirectional Encoder Representations from Transformers), in analyzing customer reviews on the Vinted platform. The Vinted platform was selected due to its diverse seller base and rich, authentic customer review data, making it an ideal environment for evaluating sentiment analysis tools in an e-commerce context.

We initiated data collection by extracting customer reviews from multiple sellers on Vinted using the RapidAPI platform. The selection of sellers was based on their active engagement and diverse product offerings to ensure a comprehensive dataset. Specifically, we targeted reviews that provided genuine customer feedback by excluding entries labeled as "Auto-feedback" and those lacking a valid Product_ID. This filtering process was essential to maintain the relevance and quality of the data, thereby enhancing the reliability of our analysis.

Once the reviews were collected, they were systematically organized into a structured dataset comprising the following variables: FEEDBACKID (unique identifier for each feedback), VADER_Compound (compound sentiment score from VADER), VADER_Sentiment (sentiment category from VADER), BERT_Label (sentiment label from BERT), BERT_Score (confidence score from BERT), and BERT_Sentiment (sentiment category mapped from BERT). This structured format facilitated seamless integration of data into statistical analysis software for subsequent evaluation.

For sentiment analysis, we employed both VADER and BERT tools. The choice of VADER and BERT is justified by their contrasting approaches – VADER for rapid, rule-based sentiment scoring and BERT for context-sensitive analysis – which allows us to evaluate the trade-offs between efficiency and accuracy in practical applications. VADER, being a lexicon and rule-based tool, is particularly adept at analyzing sentiments expressed in social media contexts. It assigns a compound score ranging from -1 (most negative) to +1 (most positive), categorizing sentiments as Positive, Neutral, or Negative based on predefined thresholds. Conversely, BERT utilizes deep learning and transformer architectures to provide context-aware sentiment classifications, offering labels such as POSITIVE, NEGATIVE, and NEUTRAL along with associated confidence scores.

To accommodate BERT's input requirements, review texts were truncated to 512 characters and tokenized using the appropriate preprocessing functions, ensuring compatibility with the model's transformer architecture

The statistical analysis was conducted using SPSS software, following a systematic approach. Initially, we performed Tests of Normality using both the Kolmogorov-Smirnov and Shapiro-Wilk tests to assess whether the sentiment scores from VADER and BERT adhered to a normal distribution. Given that both tests indicated significant deviations from normality ($p < 0.05$), non-parametric methods were deemed appropriate for subsequent analyses.

The resulted dataset was formed of **294** customer reviews, each annotated with sentiment scores and classifications from both VADER and BERT.

Subsequently, we conducted Descriptive Statistics to summarize the central tendency and variability of the sentiment scores. To explore the relationship between the sentiment scores generated by VADER and BERT, we employed Spearman's Rho correlation analysis, which is suitable for non-normally distributed data. Additionally, we analyzed the Sentiment Distribution across both tools using frequency tables and bar charts to visualize the prevalence of each sentiment category.

Finally, we created various visualizations, including bar charts and histograms, to graphically represent the distribution and correlation of sentiment scores. These visual tools aid in the intuitive understanding of the data and support the interpretation of statistical findings.

The flowchart in Figure no. 1 presents our methodology for data collection and analysis.



Figure no. 1: Methodology used for data collection and analysis

Source: authors' own work

While this study primarily focuses on correlation analysis, future evaluations may incorporate additional metrics such as precision, recall, and F1-score to provide a more comprehensive assessment of tool performance.

3. Results

The sentiment analysis of customer reviews on the Vinted platform was systematically conducted using a Python-based script that integrated several open-source libraries and APIs. The process began with the importation of essential libraries, including requests for handling HTTP requests, pandas for data manipulation, nltk for natural language processing, and transformers from Hugging Face for leveraging advanced machine learning models (Devlin et al., 2019; Peng & Matsui, 2016).

Initially, the script utilized the nltk library to download the necessary VADER lexicon, which is pivotal for performing lexicon-based sentiment analysis (Hutto & Gilbert, 2014). This setup ensured that the VADER sentiment analyzer was equipped with a comprehensive set of lexical features to accurately gauge sentiment intensity in textual data.

```
import nltk
from nltk.sentiment import SentimentIntensityAnalyzer
# Step 1: Download necessary NLTK data
nltk.download('vader_lexicon')
```

Subsequently, the script defined a function, fetch_seller_reviews, designed to retrieve customer reviews from Vinted sellers via the RapidAPI platform. Utilizing the requests library, the function iteratively fetched reviews across multiple pages for each specified seller ID. This approach ensured a thorough collection of feedback, enhancing the dataset's comprehensiveness.

```
import requests
import time
# Step 2: Define the function to fetch seller reviews from RapidAPI
def fetch_seller_reviews(seller_id, country='us', max_pages=5):
    url = "https://vinted3.p.rapidapi.com/getSellerFeedback"
    headers = {
        "Accept": "application/json",
        "x-rapidapi-key": "YOUR_RAPIDAPI_KEY",
        "x-rapidapi-host": "vinted3.p.rapidapi.com",
        "x-rapidapi-ua": "RapidAPI-Playground"
    }
    reviews_list = []
    for page in range(1, max_pages + 1):
        querystring = {"country": country, "userId": seller_id, "page":
page}
        response = requests.get(url, headers=headers, params=querystring)
        if response.status_code == 200:
            data = response.json()
            if data:
                for review in data:
```

```
                    review['seller_id'] = seller_id
                reviews_list.extend(data)
                print(f"Fetched page {page} for seller {seller_id} with
{len(data)} reviews.")
            else:
                print(f"No data found on page {page} for seller
{seller_id}.")
                break
        else:
            print(f"Failed to fetch page {page} for seller {seller_id}.
Status Code: {response.status_code}")
            break
        time.sleep(1)  # To respect API rate limits
    return reviews_list
```

The core of the analysis was orchestrated within the main function. This function initiated the data collection by iterating through a predefined list of seller IDs, invoking the fetch_seller_reviews function for each seller to accumulate a comprehensive set of reviews. The collected data was then transformed into a pandas DataFrame, facilitating efficient data manipulation and analysis (Peng & Matsui, 2016).

```
import pandas as pd
# Step 3: Define the main function
def main():
    # Define list of seller IDs
    seller_ids = [
        182458315,
        219303007,
        204321845,
        155434839,
        196763679,
        233469634,
        233226505,
        150426467,
        198035324,
        168537252
    ]
    all_reviews = []
    # Fetch reviews for each seller
    for seller_id in seller_ids:
        print(f"Fetching reviews for seller ID: {seller_id}")
        reviews = fetch_seller_reviews(seller_id, country='us', max_pages=5)
# Adjust max_pages as needed
        if reviews:
            all_reviews.extend(reviews)
        else:
            print(f"No reviews fetched for seller ID: {seller_id}")
    if not all_reviews:
```

```
        print("No reviews fetched for any seller. Exiting.")
        return
    # Convert to DataFrame
    df_reviews = pd.DataFrame(all_reviews)
```

To ensure data quality, the script meticulously extracted relevant fields such as Product_Title, Product_ID, User_Username, and Review_Content from the nested JSON responses. Reviews containing "Auto-feedback" or lacking a valid Product_ID were systematically excluded, thereby refining the dataset for accurate sentiment analysis.

```
    # Extract relevant fields and handle missing data
    df_reviews['Product_Title'] = df_reviews['product'].apply(lambda x:
x['title'] if isinstance(x, dict) and 'title' in x else None)
    df_reviews['Product_ID'] = df_reviews['product'].apply(lambda x:
x['productId'] if isinstance(x, dict) and 'productId' in x else None)
    df_reviews['User_Username'] = df_reviews['user'].apply(lambda x:
x['username'] if isinstance(x, dict) and 'username' in x else None)
    df_reviews['User_ID'] = df_reviews['user'].apply(lambda x: x['userId']
if isinstance(x, dict) and 'userId' in x else None)
    df_reviews['User_URL'] = df_reviews['user'].apply(lambda x: x['url'] if
isinstance(x, dict) and 'url' in x else None)
    df_reviews['Review_Content'] = df_reviews['content']
    df_reviews['Review_Rating'] = df_reviews['rating']
    df_reviews['Created_At'] = df_reviews['createdAt']
    # Select and reorder relevant columns
    df_reviews = df_reviews[[
        'seller_id',
        'feedbackId',
        'Created_At',
        'Product_ID',
        'Product_Title',
        'Review_Content',
        'Review_Rating',
        'User_ID',
        'User_Username',
        'User_URL'
    ]]
    # Step 4: Filter out unwanted reviews
    # Exclude reviews containing "Auto-feedback"
    df_reviews =
df_reviews[~df_reviews['Review_Content'].str.contains("Auto-feedback",
na=False)]
    # Exclude reviews without Product_ID
    df_reviews = df_reviews[df_reviews['Product_ID'].notnull()]
    print(f"Total reviews after filtering: {len(df_reviews)}")
```

The sentiment analysis was bifurcated into two distinct methodologies: lexicon-based analysis using VADER and transformer-based analysis using BERT. For VADER, the script instantiated the SentimentIntensityAnalyzer to compute compound sentiment scores and categorize each review as Positive, Neutral, or Negative based on predefined thresholds (Hutto & Gilbert, 2014).

```
    # Step 5: Perform sentiment analysis using VADER
    print("Performing VADER sentiment analysis...")
    vader_analyzer = SentimentIntensityAnalyzer()
    def get_vader_sentiment(text):
        scores = vader_analyzer.polarity_scores(text)
        compound = scores['compound']
        if compound >= 0.05:
            sentiment = 'Positive'
        elif compound <= -0.05:
            sentiment = 'Negative'
        else:
            sentiment = 'Neutral'
        return pd.Series([compound, sentiment])
    df_reviews[['VADER_Compound', 'VADER_Sentiment']] =
df_reviews['Review_Content'].apply(get_vader_sentiment)
    print("VADER sentiment analysis completed.")
```

In parallel, BERT was utilized through the transformers library's sentiment-analysis pipeline, offering a more context-aware sentiment classification (Devlin et al., 2019). The script ensured that each review text was truncated to 512 characters to comply with BERT's input limitations. Each review was then analyzed, yielding a sentiment label and a corresponding confidence score.

```
    # Step 6: Perform sentiment analysis using BERT
    print("Performing BERT sentiment analysis...")
    bert_sentiment_pipeline = pipeline('sentiment-analysis')
    def get_bert_sentiment(text):
        # Truncate text to 512 characters to fit BERT's max token limit
        truncated_text = text[:512]
        try:
            result = bert_sentiment_pipeline(truncated_text)[0]
            label = result['label']
            score = result['score']
            # Map BERT labels to match VADER categories
            if label.lower() == 'positive':
                sentiment = 'Positive'
            elif label.lower() == 'negative':
                sentiment = 'Negative'
            else:
                sentiment = 'Neutral'
        except Exception as e:
            print(f"Error processing text: {e}")
            label = 'Neutral'
```

```
            score = 0.0
            sentiment = 'Neutral'
        return pd.Series([label, score, sentiment])
    df_reviews[['BERT_Label', 'BERT_Score', 'BERT_Sentiment']] =
df_reviews['Review_Content'].apply(get_bert_sentiment)
    print("BERT sentiment analysis completed.")
```

Upon completion of the sentiment analyses, the script consolidated the results into a single Excel file using pandas, ensuring that all relevant data, including sentiment scores and categories from both VADER and BERT, were systematically organized for subsequent statistical evaluation (Peng & Matsui, 2016).

```
    # Step 7: Save all results to a single Excel sheet
    excel_file = 'seller_reviews_sentiment.xlsx'
    df_reviews.to_excel(excel_file, sheet_name='Reviews', index=False)
    print(f"Sentiment analysis results saved to '{excel_file}' in 'Reviews'
sheet.")
```

The execution of the script resulted in a comprehensive dataset of **294** customer reviews, each annotated with sentiment scores and classifications from both VADER and BERT. The meticulous filtering ensured the exclusion of non-genuine feedback, thereby enhancing the dataset's reliability. The final Excel file, presented in Appendix 1, serves as a foundational resource for the ensuing statistical analyses, encapsulating both the raw review data and the derived sentiment insights.

Table no. 1: Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
VADER_Compound	294	1.3092	-0.3578	0.9514	0.0879	0.2357	0.0560
BERT_Score	294	0.4866	0.5132	0.9999	0.9229	0.1039	0.011
Valid N (listwise)	294						

Source: authors' own work

The **VADER_Compound** scores ranged from **-0.3578** to **0.9514**, with a mean of **0.0879** and a standard deviation of **0.2357**. In contrast, **BERT_Score** exhibited a narrower range from **0.5132** to **0.9999**, with a higher mean of **0.9229** and a lower standard deviation of **0.1039**. These statistics suggest that BERT_Score values are more consistently high, whereas VADER_Compound scores exhibit greater variability.

To ascertain the distribution characteristics of the sentiment scores, we conducted both the Kolmogorov-Smirnov and Shapiro-Wilk tests. The results are presented in Table no. 2.

Table no. 2: Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	df	Sig.
VADER_Compound	0.465	294	0.000	0.593	294	0.000
BERT_Score	0.345	294	0.000	0.662	294	0.000
a. Lilliefors Significance Correction						

Source: authors' own work

The p-values for both the Kolmogorov-Smirnov and Shapiro-Wilk tests were **0.000** for both **VADER_Compound** and **BERT_Score**, indicating that neither variable follows a normal distribution. Consequently, non-parametric statistical methods were employed for further analysis.

Given the non-normal distribution of both variables, **Spearman's Rho** was utilized to assess the correlation between **VADER_Compound** and **BERT_Score**. The results are detailed in Table no. 3.

Table no. 3: Spearman Correlation Between VADER and BERT Sentiment Scores

			VADER_Compound	BERT_Score
Spearman's rho	VADER_Compound	Correlation Coefficient	1.000	0.200**
		Sig. (2-tailed)	.	0.001
		N	294	294
	BERT_Score	Correlation Coefficient	0.200**	1.000
		Sig. (2-tailed)	0.001	.
		N	294	294
**. Correlation is significant at the 0.01 level (2-tailed).				

Source: authors' own work

The Spearman's Rho coefficient of **0.200** indicates a **weak positive** correlation between **VADER_Compound** and **BERT_Score**, which is statistically significant (**p = 0.001**). This suggests a slight tendency for higher VADER compound scores to be associated with higher BERT confidence scores, albeit the relationship is not strong.

The distribution of sentiment categories as classified by VADER and BERT is illustrated in Table no. 4 and Figure no. 2.

Table no. 4: Sentiment Distribution by VADER and BERT

Sentiment Category	VADER Count	VADER Percentage	BERT Count	BERT Percentage
Negative	11	3.7%	232	78.9%
Neutral	230	78.2%	62	21.1%
Positive	53	18.0%	0	0.0%
Total	**294**	**100.0%**	**294**	**100.0%**

Source: authors' own work

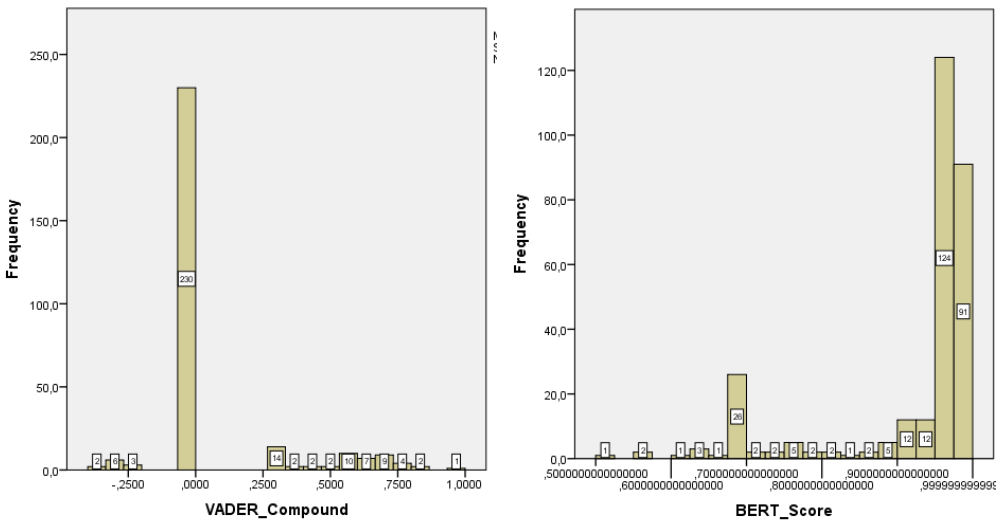


Figure no. 2: Sentiment Distribution by VADER and BERT

Source: authors' own work

The frequency tables reveal a stark contrast in sentiment classification between VADER and BERT. VADER predominantly categorized the majority of reviews as Neutral (78.2%), with smaller proportions labeled as Positive (18.0%) and Negative (3.7%). In contrast, BERT classified a substantial majority of reviews as Negative (78.9%) and a smaller segment as Positive (21.1%), with no reviews classified as Neutral. This discrepancy underscores the differing methodologies and sensitivity of the two tools in sentiment analysis.

Figure no. presents the confusion matrix comparing VADER's sentiment classifications to the true sentiment labels.

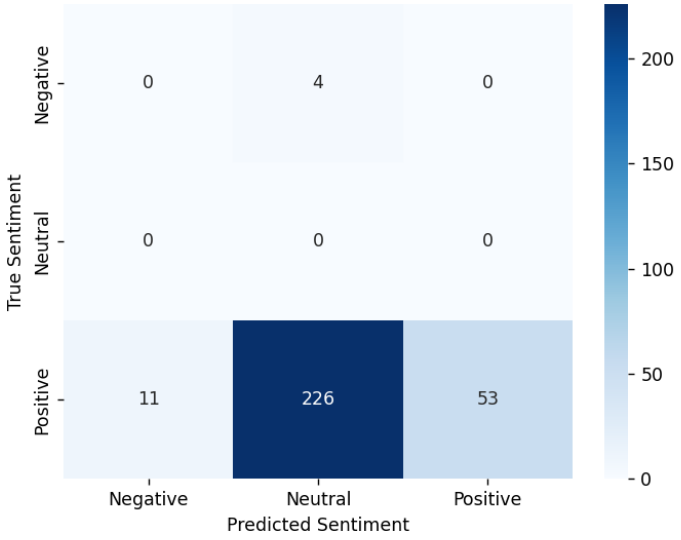


Figure no. 3: Confusion Matrix – Vader vs True Sentiment

Source: authors' own work

The sentiment analysis reveals that almost all reviews are truly **Positive** (290 out of 294). **VADER** achieved an **18% accuracy**, correctly identifying **18%** of Positive sentiments but failing to recognize any Negative sentiments. Figure no. displays the confusion matrix for BERT against the true sentiments

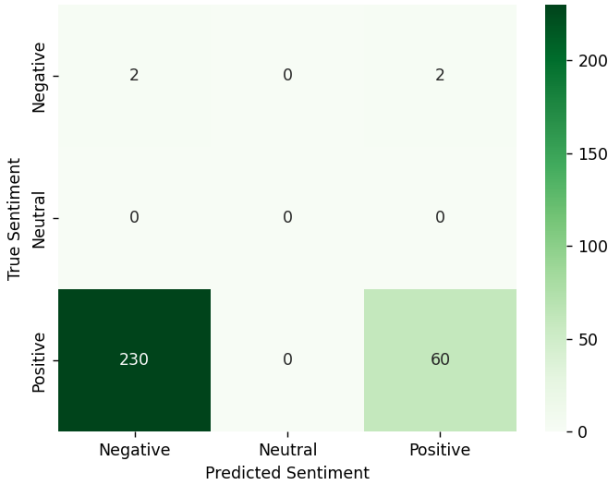


Figure no. 4: Confusion Matrix – BERT vs True Sentiment

Source: authors' own work

BERT slightly outperformed VADER with a **21% accuracy**, accurately identifying **21%** of Positive sentiments and **50%** of Negative sentiments. These results highlight that both VADER and BERT exhibit limited effectiveness in aligning with the highly imbalanced true sentiment distribution, indicating challenges in accurately capturing the nuanced sentiments within the customer reviews

The Cohen's Kappa between VADER and BERT sentiment classifications was **0.0954**, indicating only **slight agreement**. The low Cohen's Kappa value indicates only slight agreement between VADER and BERT, suggesting that the tools capture different dimensions of sentiment likely due to their distinct methodological biases.

Figure no. compares the sentiment classification results from VADER and BERT, revealing significant discrepancies in how the two tools label the same reviews.

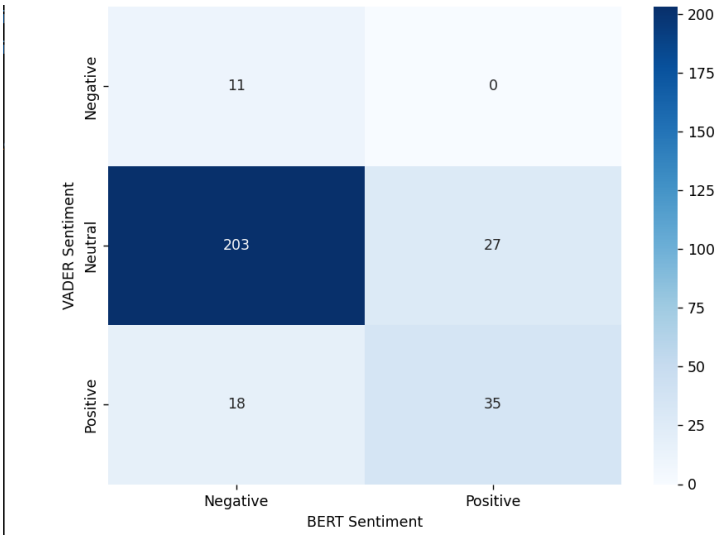


Figure no. 5: Sentiment Classification Agreement between VADER and BERT

Source: authors' own work

As shown in the contingency table, out of 294 reviews, **11** were consistently classified as Negative by both tools, **203** were labeled as Neutral by VADER but Negative by BERT, **27** as Neutral by VADER but Positive by BERT, **18** as Positive by VADER but Negative by BERT, and **35** as Positive by both. This demonstrates significant discrepancies and limited concordance between VADER and BERT in sentiment analysis.

Figure no. illustrates the variability of VADER compound sentiment scores across different sellers, showing that some sellers receive more inconsistent sentiment feedback than others.

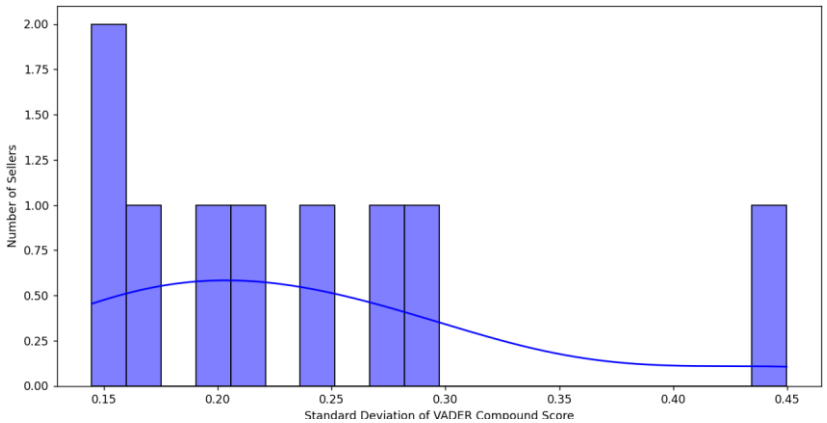


Figure no. 6: Distribution of VADER Compound Score Standard Deviation Across Sellers

Source: authors' own work

The volatility analysis reveals significant variability in sentiment scores across different sellers. For **VADER**, the most volatile seller is **233226505** with a standard deviation of **0.4497**, followed by **168537252** (0.2949), **196763679** (0.2758), **182458315** (0.2421), and **204321845** (0.2091). This indicates that the sentiments expressed in reviews for these sellers fluctuate considerably when assessed by VADER.

Figure no. presents the distribution of BERT sentiment score standard deviations across sellers, indicating that while most sellers have consistent textual patterns, a subset shows greater variability.

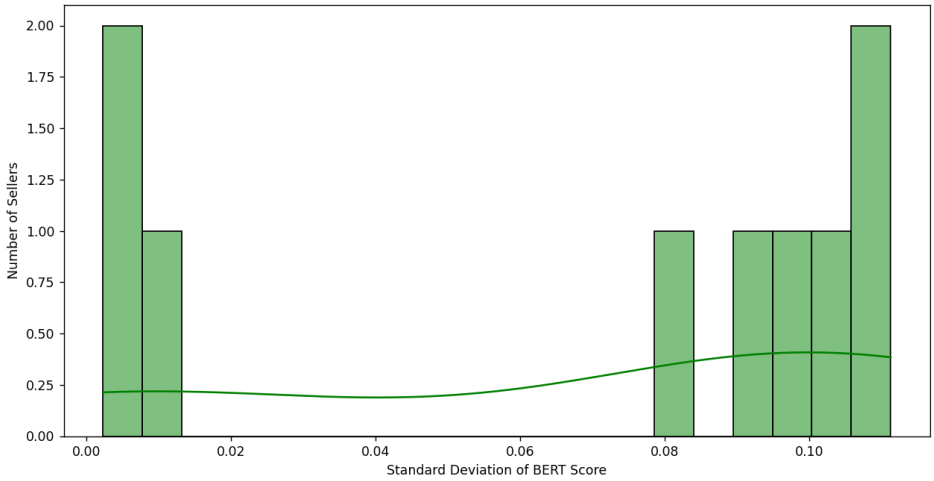


Figure no. 7: Distribution of BERT Score Standard Deviation Across Sellers

Source: authors' own work

Most sellers exhibit very low standard deviation values near 0, indicating consistent similarity in their textual data. However, a subset of sellers has higher standard deviations (e.g., above 0.1), reflecting greater variability in the BERT score, suggesting more diverse textual patterns or content variability for these sellers. The overall distribution is bimodal, with peaks at both low and high standard deviation ranges.

The variability in sentiment scores across sellers may reflect differences in customer engagement or product quality, offering businesses a potential indicator to identify areas for service improvement or targeted marketing strategies.

Figure no. visualizes the range of sentiment score fluctuations using a boxplot, highlighting sellers with more volatile customer feedback.

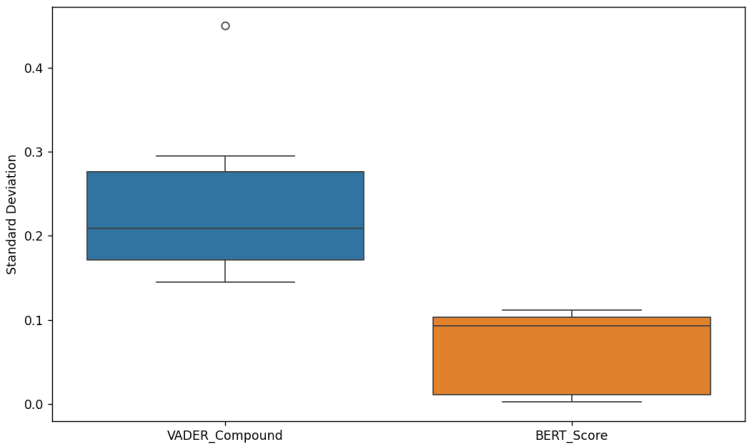


Figure no. 8: Boxplot of Sentiment Score Standard Deviations

Source: authors' own work

These findings suggest that certain sellers experience more inconsistent customer feedback, which could be attributed to factors like varying product quality, diverse customer expectations, or differing interaction experiences. Understanding this volatility is crucial for businesses to identify and address underlying issues that lead to fluctuating customer sentiments.

The correlation analysis began by assessing the normality of the sentiment score distributions for both VADER_Compound and BERT_Score using the Kolmogorov-Smirnov and Shapiro-Wilk tests. As it can be seen in Table no. 5, all tests yielded p-values of 0.0000, leading to the rejection of the null hypothesis

that the data follows a normal distribution. Consequently, Spearman's rank correlation was employed to evaluate the relationship between the sentiment scores and the actual Review_Rating.

Table no. 5: Normality Tests and Correlation Coefficients

Analysis	Variable Pair	Statistic	p-value
Normality Test	VADER_Compound (Kolmogorov-Smirnov)	0.4651	0.0000
Normality Test	VADER_Compound (Shapiro-Wilk)	0.5932	0.0000
Normality Test	BERT_Score (Kolmogorov-Smirnov)	0.3446	0.0000
Normality Test	BERT_Score (Shapiro-Wilk)	0.6625	0.0000
Correlation Coefficient	VADER_Compound vs Review_Rating	0.0099	0.8659
Correlation Coefficient	BERT_Score vs Review_Rating	0.1249	0.0323

Source: authors' own work

The Spearman correlation between VADER_Compound and Review_Rating was found to be 0.0099 with a p-value of 0.8659, indicating no significant correlation. This suggests that VADER's sentiment scores do not reliably reflect the actual review ratings provided by customers.

Figure no. 9 displays the relationship between sentiment scores and actual review ratings, showing a very weak Spearman correlation and suggesting that VADER and BERT scores do not strongly align with customers' numerical ratings.

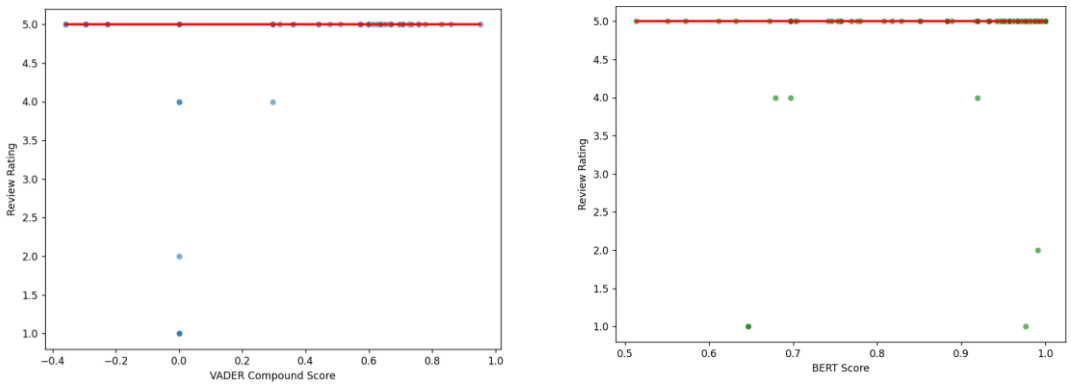


Figure no. 9: VADER Compound Score & BERT Score vs Review Rating (Spearman Correlation = 0.01)

Source: authors' own work

In contrast, BERT_Score exhibited a weak positive correlation of 0.1249 with a p-value of 0.0323, which is statistically significant. Although the correlation is still weak, it indicates that higher BERT sentiment scores tend to be associated with higher review ratings. This marginal relationship implies that BERT may have a slightly better efficacy in aligning sentiment analysis with actual customer ratings compared to VADER.

The weak correlations observed could be attributed to the highly imbalanced distribution of true sentiments and the inherent difficulty in capturing nuanced customer feedback through automated sentiment tools. Factors such as review length, context, and language subtleties may further contribute to this discrepancy.

However, the overall low correlation coefficients highlight the challenges both tools face in accurately capturing the nuanced sentiments embedded within the customer reviews, especially given the highly imbalanced true sentiment distribution (290 Positive vs. 4 Negative reviews). The extreme imbalance in true sentiments likely impacts the correlation results, making it difficult for sentiment analysis tools to accurately capture and reflect the true sentiment distribution. Additionally, the weak correlation observed

for BERT, while statistically significant, underscores the need for further refinement or alternative approaches to enhance sentiment classification performance in this context.

4. Discussions

The comparative analysis between VADER and BERT sentiment analysis tools reveals significant insights into their respective efficacies and inherent biases when applied to customer reviews on the Vinted platform, bolstered by findings from relevant literature.

In terms of efficacy and performance, the Descriptive Statistics indicated that VADER_Compound scores had a mean of 0.0879, suggesting a slight inclination towards positive sentiments on average, with a high standard deviation of 0.2357 reflecting considerable variability. Conversely, BERT_Score exhibited a higher mean of 0.9229 with a lower standard deviation of 0.1039, indicating that BERT consistently assigns high confidence scores, predominantly skewed towards negative sentiments. These findings align with studies that highlight BERT's superior accuracy and nuanced sentiment detection in customer feedback, often outperforming lexicon-based approaches like VADER (Sayeed et al., 2023). For example, in analyzing product reviews, BERT demonstrated enhanced F1 scores and precision over traditional models (Prasad et al., 2022).

Regarding bias in sentiment classification, the Sentiment Distribution further accentuated the divergence between the tools. VADER predominantly classified reviews as Neutral (78.2%), while BERT classified most as Negative (78.9%), with no Neutral classifications. This discrepancy reflects their distinct methodologies – VADER's lexicon-based approach can oversimplify context, while BERT's transformer-based architecture captures nuanced expressions but may amplify biases inherent in training data (Awajan et al., 2021). This aligns with findings that VADER may fail to account for subtle contextual factors, while BERT's over-sensitivity to negative expressions can skew sentiment assessments (Liu et al., 2020).

The Spearman's Rho correlation of 0.200 between VADER_Compound and BERT_Score indicates weak agreement between the tools, consistent with prior research showing limited concordance in sentiment assessments across lexicon-based and transformer-based models (Vidya & Janani, 2022). Moreover, neither tool exhibited strong alignment with actual review ratings, with weak positive correlations for BERT and no significant correlation for VADER. These results underscore challenges in accurately capturing nuanced sentiments, particularly in highly imbalanced datasets where positive reviews dominate (Marfani et al., 2022).

The findings emphasize the need for tailored tool selection based on specific use cases. For scenarios requiring speed and simplicity, VADER remains effective, especially when broad overviews are needed without intricate sentiment nuances (Mindoro et al., 2022). On the other hand, for nuanced analyses, BERT's contextual capabilities are more suitable, although its computational demands and sensitivity to training biases must be considered (Colón-Ruiz & Segura-Bedmar, 2020).

To overcome the limitations identified, several strategies can be employed. Combining VADER's efficiency with BERT's contextual analysis can provide balanced outputs, as hybrid approaches such as using VADER for initial classification and BERT for detailed analysis have shown promise in enhancing sentiment prediction accuracy (Liu et al., 2020). Additionally, addressing training data biases in BERT and refining VADER's lexicons with domain-specific terminology can improve both tools' efficacy (Awajan et al., 2021). Expanding datasets to include larger and more diverse samples can provide a more

comprehensive understanding of tool performance and reduce the impact of imbalanced sentiment distributions (Marfani et al., 2022).

Based on these findings, businesses are advised to consider a hybrid approach that leverages both VADER and BERT, thereby harnessing the speed of lexicon-based analysis alongside the contextual depth of transformer models. This integrated strategy can help overcome individual tool limitations and provide more balanced insights into customer sentiment.

In conclusion, while VADER offers simplicity and speed, BERT's contextual depth delivers more nuanced insights. A hybrid approach and continuous refinement of tools and datasets can address their limitations, enabling more accurate sentiment analysis in diverse applications.

5. Conclusions

This study contributes to sentiment analysis research by directly comparing VADER and BERT in an e-commerce setting, revealing their respective biases and operational limitations.

This study conducted a comparative analysis of two prominent sentiment analysis tools, VADER and BERT, to evaluate their efficacy and inherent biases in interpreting customer reviews on the Vinted platform. By systematically examining 294 filtered customer reviews, the research aimed to determine which tool provides a more accurate and reliable assessment of customer sentiments, thereby offering valuable insights for businesses in selecting appropriate sentiment analysis methodologies.

The findings revealed that VADER tends to categorize the majority of reviews as Neutral (78.2%), with limited identification of Positive (18.0%) and Negative (3.7%) sentiments. This high prevalence of neutral classifications suggests that VADER may underrepresent the true spectrum of customer sentiments, particularly in datasets with skewed sentiment distributions. On the other hand, BERT exhibited a contrasting pattern by predominantly classifying reviews as Negative (78.9%) and Positive (21.1%), with no instances of Neutral sentiment. While BERT showed a slightly higher accuracy in identifying Positive sentiments compared to VADER, it failed to capture Negative sentiments effectively, especially considering the dataset's actual sentiment distribution.

The correlation analysis further highlighted the challenges both tools face in accurately reflecting true customer sentiments. The Spearman's Rho correlation between VADER and BERT was weak yet statistically significant, indicating limited concordance between the two tools. Additionally, VADER's sentiment scores showed no significant correlation with actual review ratings, whereas BERT exhibited a weak positive correlation. These results underscore the difficulties both tools encounter in aligning sentiment analysis with the true sentiments expressed in customer reviews, particularly within highly imbalanced datasets.

Both VADER and BERT demonstrated distinct biases influenced by their underlying methodologies. VADER's reliance on predefined lexicons may lead to oversimplified sentiment classifications, potentially overlooking contextual nuances. Conversely, BERT's deep learning architecture, while adept at contextual understanding, showed a propensity to skew sentiments towards negativity, which may stem from biases present in the training data. The analysis of volatility across different sellers further revealed significant variability in sentiment scores, highlighting the influence of external factors such as product quality and customer interaction on sentiment expressions.

For businesses leveraging sentiment analysis, these divergent performances and bias profiles have practical implications. VADER is suitable for scenarios requiring rapid, real-time sentiment assessments with large

volumes of data due to its simplicity and efficiency. However, its tendency to default to neutral classifications may obscure critical positive or negative feedback. In contrast, BERT offers a more nuanced and context-aware sentiment analysis, making it advantageous for in-depth evaluations where understanding the subtleties of customer emotions is essential. Despite its computational demands and bias towards negative sentiments, BERT can be instrumental in identifying areas requiring attention to enhance customer satisfaction. We recommend that businesses adopt a hybrid sentiment analysis framework to capture both real-time and context-sensitive customer feedback, thereby improving decision-making and customer relationship management.

The study is limited by its sample size of 294 reviews, which, while adequate for preliminary analysis, may not fully capture the diverse range of customer sentiments on the Vinted platform. Future research should consider larger and more balanced datasets to better evaluate the tools' performances. Additionally, exploring hybrid approaches that integrate the strengths of both VADER and BERT could potentially yield more balanced and accurate sentiment assessments. Further investigation into bias mitigation techniques is also warranted to enhance the fairness and reliability of sentiment analysis tools. Incorporating adaptive lexicons, domain-specific fine-tuning, and post-processing bias correction methods could address the identified limitations and improve the overall efficacy of sentiment analysis in varied business contexts.

Future research should focus on: (a) incorporating additional evaluation metrics (precision, recall, F1-score) to complement correlation analysis; (b) using larger, more balanced datasets; and (c) exploring the impact of linguistic and cultural diversity on sentiment analysis performance.

In conclusion, both VADER and BERT possess distinct strengths and limitations in the sentiment analysis of customer reviews. VADER's efficiency and simplicity make it suitable for broad sentiment monitoring, while BERT's contextual depth offers more detailed insights at the cost of increased computational resources and potential biases. Businesses must carefully consider their specific needs, the nature of their customer feedback, and the inherent biases of each tool when selecting a sentiment analysis approach. By understanding these dynamics, organizations can better harness sentiment analysis to inform strategic decisions, enhance customer relationships, and drive market success.

It is also important to address ethical considerations in sentiment analysis, including potential biases in model training, issues of fairness, and the need for transparency in AI decision-making. Future work should develop methods to detect and mitigate these biases to ensure responsible use of sentiment analysis in business contexts

References

Aljedaani, W., Rustam, F., Mkaouer, M. W., Ghallab, A., Rupapara, V., Washington, P. B., Lee, E., & Ashraf, I. (2022). Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry. *Knowledge-Based Systems*, *255*, 109780. https://doi.org/10.1016/J.KNOSYS.2022.109780

Atanasova, T., Kasheva, M., Sulova, S., & Vasilev, J. (2010). Analysis of the possible application of Data Mining, Text Mining and Web Mining in business intelligent systems. *The 33rd International Convention MIPRO*, 1294–1297. https://doi.org/10.13140/2.1.1027.5840

Awajan, I., Mohamad, M., & Al-Quran, A. (2021). Sentiment Analysis Technique and Neutrosophic Set Theory for Mining and Ranking Big Data from Online Reviews. *IEEE Access*, *9*. https://doi.org/10.1109/ACCESS.2021.3067844

Bâra, A., & Lungu, I. (2012). Improving Decision Support Systems with Data Mining Techniques". In *Advances in Data Mining Knowledge Discovery and Applications*. https://doi.org/10.5772/47788

Biswas, E., Karabulut, M. E., Pollock, L., & Vijay-Shanker, K. (2020). Achieving Reliable Sentiment Analysis in the Software Engineering Domain using BERT. *Proceedings - 2020 IEEE International Conference on Software Maintenance and Evolution, ICSME 2020*. https://doi.org/10.1109/ICSME46990.2020.00025

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*(2). https://doi.org/10.1109/MIS.2013.30

Colón-Ruiz, C., & Segura-Bedmar, I. (2020). Comparing deep learning architectures for sentiment analysis on drug reviews. *Journal of Biomedical Informatics*, *110*, 103539. https://doi.org/10.1016/J.JBI.2020.103539

Cristescu, M. P., Mara, D. A., Nerișanu, R. A., & Culda, L. C. (2024). Leveraging Free Tools in Financial Sentiment Analysis. In C. Oprean-Stan (Ed.), *Application of Novel Research Methods. The Study of Current Economic Phenomena*. Peter Lang Verlag. https://doi.org/10.3726/B21691

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://github.com/tensorflow/tensor2tensor

HarishRao, M., ShashikumarD.RHarishRao, M., & D.R, S. (2017). Automatic Product Review Sentiment Analysis Using Vader and Feature Visulaization. *International Journal of Computer Science Engineering and Information Technology Research*, *7*(4), 53–66. https://doi.org/10.24247/IJCSEITRAUG20178

He, L., & Zheng, K. (2019). How do general-purpose sentiment analyzers perform when applied to health-related online social media data? *Studies in Health Technology and Informatics*, *264*. https://doi.org/10.3233/SHTI190418

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*. https://doi.org/10.1609/icwsm.v8i1.14550

Kotelnikova, A., Paschenko, D., Bochenina, K., & Kotelnikov, E. (2022). Lexicon-Based Methods vs. BERT for Text Sentiment Analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13217 LNCS*. https://doi.org/10.1007/978-3-031-16500-9_7

Kouadri, W. M., Benbernou, S., Ouziri, M., Palpanas, T., & Amor, I. Ben. (2022). SA-Q: Observing, Evaluating, and Enhancing the Quality of the Results of Sentiment Analysis Tools. *Proc. VLDB Endow.*, *15*, 3658–3661. https://doi.org/10.14778/3554821.3554868

Liu, Y., Lu, J., Yang, J., & Mao, F. (2020). Sentiment analysis for e-commerce product reviews by deep learning model of Bert-BiGRU-Softmax. *Mathematical Biosciences and Engineering*, *17*(6). https://doi.org/10.3934/MBE.2020398

Li, X., Bing, L., Zhang, W., & Lam, W. (2019). Exploiting bert for end-to-end aspect-based sentiment analysis_. *W-NUT@EMNLP 2019 - 5th Workshop on Noisy User-Generated Text, Proceedings*. https://doi.org/10.18653/v1/d19-5505

Loria, S. (2022). *TextBlob: Simplified Text Processing — TextBlob 0.19.0 documentation*. https://textblob.readthedocs.io/en/dev/

Marfani, H., Hina, S., & Tabassum, H. (2022). Analysis of Learners' Sentiments on MOOC Forums using Natural Language Processing Techniques. *3rd International Conference on Innovations in Computer Science and Software Engineering, ICONICS 2022*. https://doi.org/10.1109/ICONICS56716.2022.10100401

Mindoro, J. N., Malbog, M. A. F., Nipas, M. D. S., Susa, J. A. B., Acoba, A. G., & Gulmatico, J. S. (2022). Sentiment Analysis in Customer Experience in Philippine Courier Delivery Services using VADER

Algorithm Thru Chatbot Interviews. *ICPC2T 2022 - 2nd International Conference on Power, Control and Computing Technologies, Proceedings*. https://doi.org/10.1109/ICPC2T53885.2022.9777007

Moussa, M., Mohamed, E., & Haggag, M. (2018). A generic lexicon-based framework for sentiment analysis. *International Journal of Computers and Applications*, *42*, 463–473. https://doi.org/10.1080/1206212X.2018.1483813

Nassirtoussi, K. A., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*(16), 7653–7670. https://doi.org/10.1016/J.ESWA.2014.06.009

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Peng, R. D., & Matsui, E. (2016). *The Art of Data Science: A Guide for Anyone who Works with Data*. Lulu.com. https://books.google.ro/books?id=ZDH9DAEACAAJ

Prasad, M. P. R. K., Jahnavi, M. S., Reddy, M. R. J., Rama, K. V., & Narendra, K. (2022). Sentiment Analysis on Reviews of E-commerce Sites Using BERT. *IJFANS International Journal of Food and Nutritional Sciences*, *11*(12), 2024–2033. https://doi.org/10.48047/IJFANS/V11/I12/214

Sayeed, M. S., Mohan, V., & Muthu, K. S. (2023). BERT: A Review of Applications in Sentiment Analysis. In *HighTech and Innovation Journal* (Vol. 4, Issue 2). https://doi.org/10.28991/HIJ-2023-04-02-015

Stancu, A.-M. R., Pompiliu, C. M., & Stoyanova, M. (2020). Data Mining Algorithms for Knowledge Extraction. In S. L. Fotea, I. Ş. Fotea, & S. A. Văduva (Eds.), *Challenges and Opportunities to Develop Organizations Through Creativity, Technology and Ethics* (pp. 349–357). Springer International Publishing. https://doi.org/10.1007/978-3-030-43449-6_20

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11856 LNAI*, 194–206. https://doi.org/10.1007/978-3-030-32381-3_16

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*. https://arxiv.org/pdf/1706.03762v7

Venkit, P. N., Srinath, M., & Wilson, S. (2023). Automated Ableism: An Exploration of Explicit Disability Biases in Sentiment and Toxicity Analysis Models. *ArXiv*, *abs/2307.09209*. https://doi.org/10.48550/ARXIV.2307.09209

Vidya, K., & Janani, S. (2022). Analogy of Machine Learning Approaches and BERT for Sentiment Analysis. *Journal of Information Technology and Digital World*, *4*(1), 52–60. https://doi.org/10.36548/jitdw.2022.1.006

Wu, X., Lv, S., Zang, L., Han, J., & Hu, S. (2019). Conditional BERT Contextual Augmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11539 LNCS*. https://doi.org/10.1007/978-3-030-22747-0_7

Xu, H., Shu, L., Yu, P., & Liu, B. (2020). Understanding Pre-trained BERT for Aspect-based Sentiment Analysis. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 244–250). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.21

Yang, Z., Asyrofi, M. H., & Lo, D. (2021). BiasRV: uncovering biased sentiment predictions at runtime. *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. https://doi.org/10.1145/3468264.3473117

Appendix 1 Screenshot of the Resulted Excel File with Analyzed Customer Reviews

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	seller_i	feedbackI	Created_At	Product_I	Product_Title	Review_Content	Review_Rating	User_ID	User_Username	User_URL	VADER_Compoun	VADER_Sentimer	BERT_Labe	BERT_Scor	BERT_Sentimer
35	182458315	2347681572	2024-10-10T17:49:04+02:00	4008177572	Tricou Bershka S	Tricoul a fost impecabil!	5	168524762	dressing.me	https://www.vint	0	Neutral	NEGATIVE	0,996989429	Negative
36	233469634	2383031611	2024-10-24T23:53:24+02:00	5254139357	Pulover dama oversize	Totul super!	5	153503223	pixiemisi	https://www.vint	0	Neutral	POSITIVE	0,888657331	Positive
37	168537252	2376886368	2024-10-22T09:19:56-06:00	3852961067	Hanorac cu glugă / pulover culoare	totul perfect! recomand calduros	5	170079733	dumidan	https://www.vint	0,6114	Positive	POSITIVE	0,998470962	Positive
38	204321845	2361223494	2024-10-16T02:26:25-06:00	5204018611	Pantaloni cu talie dubla Zara noi cu	Totul perfect recomand	5	146850351	miamihaela2015	https://www.vint	0,5719	Positive	POSITIVE	0,969069064	Positive
39	233469634	2400079077	2024-11-01T21:01:23+01:00	5246024314	Rochie dama	Totul ok,recomand!	5	164371643	a.ale.miha	https://www.vint	0	Neutral	NEGATIVE	0,817581356	Negative
40	182458315	2380343404	2024-10-23T21:38:48+02:00	4354505468	Tricou / rochie	totul ok	5	190676524	adrienn7.5	https://www.vint	0,296	Positive	POSITIVE	0,991116166	Positive
41	204321845	2394715059	2024-10-30T15:49:11+02:00	5276086675	Pulover Only marimea S	Totul în regulă! Recomand	5	207660600	mariaandreea2526	https://www.vint	0	Neutral	NEGATIVE	0,988317072	Negative
42	182458315	2357554673	2024-10-14T18:16:14+02:00	3987892933	Pulover oversized	totul în regula. atât hainele, cât și încălțămi	5	209619502	alle.pop	https://www.vint	0	Neutral	NEGATIVE	0,990408182	Negative
43	150426467	2248594164	2024-08-31T11:13:02+03:00	4849192412	Bluziță fetiță,marca dopodopo,mari	Totul excelent!!!	5	187623016	kristunq	https://www.vint	0	Neutral	POSITIVE	0,999585688	Positive
44	196763679	2039865325	2024-05-09T10:48:34-06:00	4434201432	Fustă	totul este super, recomand!! :-)	5				0,7772	Positive	NEGATIVE	0,917011142	Negative
45	204321845	2407526140	2024-11-05T22:14:34+02:00	5303963802	Camasa Zara	Totul a fost ok	5	189353163	melamoldovaan	https://www.vint	0,296	Positive	NEGATIVE	0,95973897	Negative
46	168537252	2374306695	2024-10-21T08:33:25-06:00	4672788535	Oferta 2+1 gratis! Rochie albă /vara	totul a fost conform descrierii.colabore plă	5	193049976	ponciolita	https://www.vint	0	Neutral	NEGATIVE	0,990991592	Negative
47	204321845	2381753549	2024-10-24T08:11:37-06:00	5237409917	Pulover în V tricotat	Totul a decurs perfect. Recomand!!	5	162496606	b1ancaionita	https://www.vint	0,6467	Positive	POSITIVE	0,769353151	Positive
48	204321845	2350032098	2024-10-11T07:32:08-06:00	5103503176	Bluza Zara tricotata noua cu etichet	totul a decurs ok!	5	197225332	alexandruchitu	https://www.vint	0,3595	Positive	POSITIVE	0,571565628	Positive
49	204321845	2274491505	2024-09-11T12:32:40-06:00	4991427095	Maleta Zara cu guler Polo femei	totul a decurs minunat	5	168031161	bonerichy	https://www.vint	0	Neutral	NEGATIVE	0,973904192	Negative
50	196763679	2050665605	2024-05-14T14:14:04-06:00	4466918808	Rochie ZARA	Totul a decurs foarte ok. Produsul este exa	5	199339492	cati08.95	https://www.vint	0,296	Positive	NEGATIVE	0,987419426	Negative
51	168537252	2415071563	2024-11-09T03:16:42+02:00	3922777175	Pulover/ maletă superb/ scurtă / po	Thank you very much! I really appreciate th	5	237727933	danaemp01	https://www.vint	0,8286	Positive	POSITIVE	0,999866605	Positive

Source: authors' own work