advances in Technology

*Adv. Technol. 2024, 4(2), 104-120*

## Full Paper

# A Study of Clustering Approaches Applied to Customer Reviews in the Digital Era

M.N.S. Tissera*, P.P.G.D. Asanka, and R.A.C.P. Rajapakse

Department of Industrial Management, Faculty of Science, University of Kelaniya, Sri Lanka

Corresponding Author: nimeshshamika@gmail.com

Received: 24 May 2024; Revised: 25 July 2024; Accepted: 23 August 2024; Published: 19 January 2025

**Abstract**

The digital revolution has reshaped the landscape of business transactions, with online platforms generating vast amounts of text data through customer reviews. This paper explores the transformative potential of harnessing this data for customer segmentation, comparing traditional methods such as Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words (BoW) with state-of-the-art Large Language Models (LLMs) for sentence embeddings. The primary objective is to identify the most effective approach for customer segmentation based on textual data by conducting a comprehensive analysis using clustering approaches. The study investigates the impact of LLMs, specifically BERT, RoBERTa, XLNet, and MPNet, in contrast to TF-IDF and BoW. Through experimentation and evaluation metrics, including the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, the research sheds light on the nuanced effectiveness of each method. While LLMs, particularly RoBERTa, demonstrate superior clustering performance, the study acknowledges the subtle impact of spelling correction on these models. The findings provide valuable insights for businesses seeking to understand customer sentiments and preferences, enabling more targeted and personalized strategies in the dynamic digital age. This research contributes to the evolving field of customer analytics by offering a comparative analysis of clustering approaches, laying the foundation for future advancements in text-based customer segmentation.

**Keywords:** Clustering, Customer Segmentation, Language Model, Marketing, Text Analysis

## Introduction

The digital revolution has fundamentally transformed the landscape of business transactions. The advent of online platforms has led to a significant shift in how businesses interact with their customers. This shift has not only streamlined the transaction process but has also led to the generation of a vast amount of text data through customer reviews. If harnessed effectively, this data can offer valuable insights into customer behavior and preferences, enabling businesses to customize their services to meet customer needs more accurately. This phenomenon is a testament to the power of the digital age, where data has become a crucial asset for businesses [1].

The significance of text data in the digital age cannot be overstated. A plethora of studies have shown that customers generate a substantial amount of text data through online transactions. This data, often in the form of reviews, feedback, and comments, can serve as a rich source of information for businesses seeking to understand their customers better. However, the unstructured nature of this data presents a challenge to traditional data analysis techniques. This necessitates the development of more sophisticated methods that can effectively handle and analyze this data to extract meaningful insights [2].

The primary objective of this study is to conduct a comparative analysis of clustering approaches for customer segmentation based on textual data (customer reviews). We explore the impact of utilizing Large Language Models (LLMs) for sentence embedding creation in contrast to traditional methods. Bag-of-Words (BoW) and TF-IDF are some well-known and widely used traditional methods for text processing. They have been used for tasks involving but not limited to sentiment analysis, document classification, and clustering [3-6]. This motivated us to explore and compare these approaches with more sophisticated approaches such as the use of LLMs. Our investigation involves three approaches: (1) clustering using Term Frequency-Inverse Document Frequency (TF-IDF) vectors, (2) clustering using Bag-of-Words (BoW) representations, and (3) clustering based on LLM-generated embeddings.

When exploring TF-IDF and BoW, we concatenated all the reviews of a customer before generating vectors for their reviews. These vectors were then clustered to segment the customers. This approach allowed us to generate a single vector representation for each customer rather than having multiple vectors corresponding to each review. This helped us effectively compare the effectiveness of traditional text representation techniques with that of LLMs in the context of customer segmentation.

The research aims to identify the most effective approach for customer segmentation. This will provide businesses with a robust tool to understand and cater to their customers in the digital age. Furthermore, it will contribute to the broader field of customer analytics by providing a comparative analysis of various clustering approaches in the context of customer segmentation. This research is expected to pave the way for more sophisticated and effective customer segmentation techniques in the future.

## Background

### *Customer Segmentation Methods*

Customer segmentation has emerged as a critical strategy in the realm of customer-oriented marketing, especially in the burgeoning e-commerce sector [7]. This process involves the division of a company's customer base into distinct groups, each characterized by shared attributes or behaviors. The primary objective of customer segmentation is to gain a deeper understanding of the interests and motivations of individual customers, thereby enabling businesses to tailor their marketing efforts to meet the unique needs of each segment [7].

The advent of machine learning has brought about a significant enhancement in the effectiveness of customer segmentation. Various machine-learning techniques have been employed to facilitate this process. Among these, K-means clustering is widely used due to its simplicity, versatility, and scalability. K-means works by partitioning data into a predetermined number of clusters, where each data point is assigned to the nearest cluster center (centroid). The algorithm iteratively adjusts the centroids by minimizing the variance within clusters, thus enhancing the homogeneity of data points in the same cluster. Its effectiveness in handling large datasets makes K-means particularly relevant for applications like customer segmentation, where the goal is to categorize customers into well-defined, actionable segments based on their behaviors or characteristics [7].

Recent research has delved into the exploration of different machine-learning techniques for customer segmentation. For instance, a comprehensive review by Gomes and Meisen (2023) provides an in-depth overview of various segmentation methods and their current state-of-the-art. The authors conducted an extensive literature search, identifying 105 publications between 2000 and 2022 that deal with the analysis of customer behavior using segmentation methods. This body of work underscores the growing interest in leveraging machine learning for customer segmentation and the continuous evolution of these techniques.

Another noteworthy study by Luo et al., (2022) delved into the sentiment analysis of spa leisure consumption during different holidays and across different cities, with the aim of optimizing customer segmentation [8]. The authors proposed a novel general framework and related sentiment analysis methods, which were applied to a collection of datasets from customers' textual reviews of foot bath spa merchants in three cities in China. This study exemplifies the potential of sentiment analysis in enhancing customer segmentation, particularly in the context of service industries where customer reviews play a pivotal role.

Hence, in summary, the literature indicates that customer segmentation is a vital tool for businesses seeking to understand and cater to their customers' needs. The use of machine learning techniques, particularly clustering algorithms, has shown promise in enhancing the effectiveness of customer segmentation. However, further research is needed to explore the potential of these techniques in different contexts and with different types of data. This paper aims to contribute to this body of knowledge by providing a comparative analysis of different clustering approaches for customer segmentation using text data generated from online transactions.

***Text Data Analysis and Sentiment Analysis in Customer Feedback***

The increasing volume of textual data generated by online customer reviews has sparked a surge in research seeking to extract valuable insights for customer segmentation and understanding. This section delves into existing studies utilizing text data analysis and sentiment analysis in customer feedback, particularly focusing on how these techniques have been employed to unveil customer sentiments and preferences within reviews.

Sentiment analysis, a key technique used in this context, involves the use of Natural Language Processing (NLP) and text mining to identify and extract subjective information from text. This technique has been instrumental in analyzing comments and reviews concerning day-to-day activities, as highlighted by Wankhade et al., (2022). They provide a comprehensive overview of sentiment analysis methods, applications, and challenges, emphasizing its importance in the realm of customer feedback analysis [9]. Furthermore, Nandwani et al., (2021) discusses the various levels of sentiment analysis and emotion models [10]. They underscore the role of sentiment analysis in understanding customer reviews on various e-commerce sites. According to them, sentiment analysis assists marketers in understanding their customer's perspectives better, enabling them to make necessary changes to their products or services.

Opinion mining, also known as sentiment analysis, has been utilized to extract valuable insights from customer feedback. A study by Subhashini et al., (2021) provides a detailed survey of recent opinion-mining literature [11]. They discuss how to extract text features in opinions that may contain noise or uncertainties. They highlight the importance of opinion mining in e-commerce, where customer preference patterns can significantly impact companies' overall profits. Another study conducted a sentiment analysis on customer feedback data from Amazon product reviews [12]. They utilized opinion mining and text mining to understand customer sentiments toward specific products. The study underscores the importance of sentiment analysis in changing customer opinions about a product.

In conclusion, the analysis of text data, particularly through sentiment analysis and opinion mining, plays a crucial role in extracting customer insights. These techniques provide a deeper understanding of customer sentiments and preferences, enabling businesses to make informed decisions and improve their products and services. The advent of machine learning has further enhanced these techniques, allowing for the analysis of large volumes of data and the identification of patterns that may not be apparent through manual analysis. Machine learning algorithms can segment customers based on various factors such as purchasing behavior, demographics, and psychographics. By building upon existing research and addressing challenges head-on, this study seeks to contribute to the advancement of text-based customer segmentation, empowering businesses to harness the wealth of online reviews for a deeper understanding of their customers and more effective engagement strategies.

***Language Model-Based Approaches for Sentence Embeddings***

The emergence of powerful Large Language Models (LLMs) like BERT, RoBERTa, XLNet, and MPNet has revolutionized NLP, offering a quantum leap in our ability to capture the semantic richness and intricacies of human language [13, 14]. A key element of this revolution is the creation of high-quality sentence embeddings, dense vector representations that encode the meaning and context of a sentence [13, 14].

Traditionally, NLP relied on simple techniques like bag-of-words or TF-IDF to represent sentences, often resulting in shallow and context-agnostic representations [15]. LLMs, trained on massive text corpora, overcome these limitations by capturing the complex relationships between words and their nuances within a sentence [13, 14]. This allows them to generate contextualized word embeddings, where the

meaning of a word depends on its surrounding context [13, 14]. By aggregating these word embeddings, we can derive robust sentence embeddings that reflect the true semantic meaning of a sentence [13, 14]. Studies showcase the effectiveness of LLM-based sentence embeddings in various NLP tasks. For instance, Adoma et al., (2022) demonstrated that these LLMs significantly outperform traditional methods in recognizing emotions from texts [16]. Another research reported a comprehensive clustering and network analysis targeting sentence and sub-sentence embedding spaces [17]. The study found that one method generates the most clustering-friendly embeddings and that the embeddings of span sub-sentences have better clustering properties than the original sentences.

Furthermore, Adoma et al., (2022) analysed the efficacy of BERT, RoBERTa, DistilBERT, and XLNet pre-trained transformer models in recognizing emotions from texts [16]. The study found that using the same hyperparameters, the recorded model accuracies in decreasing order were RoBERTa, XLNet, BERT, and DistilBERT, respectively [16]. Moreover, Li et al., (2020) discusses how pre-trained contextual representations like BERT have achieved great success in NLP. However, the sentence embeddings from the pre-trained language models without fine-tuning have been found to capture the semantic meaning of sentences poorly [18]. The study argues that the semantic information in the BERT embeddings is not fully exploited [18].

**Methodology**

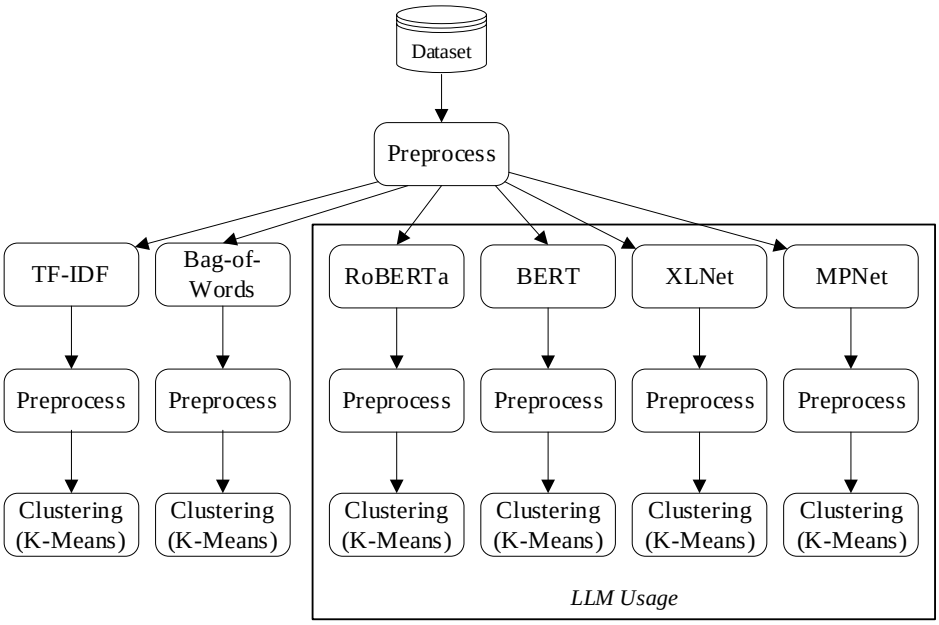


**Figure 1.** Overall methodology comparing traditional and LLM-based methods for clustering with K-means

The overarching goal was to assess the effectiveness of traditional methods, specifically TF-IDF and Bag-of-Words, in comparison to modern approaches utilizing LLMs for sentence embeddings. The experiment aimed to gain insights into the optimal technique for customer segmentation in the context of abundant textual data generated in the digital age. The primary hypothesis behind this study hypothesized that the

utilization of LLMs for generating sentence embeddings would lead to more meaningful and contextually rich representations of textual information, resulting in improved clustering performance compared to traditional methods such as TF-IDF and Bag-of-Words. The hypothesis was tested through a systematic analysis of clustering results and evaluation metrics. The overall experimentation methodology is shown in Figure 1.

The experiment was structured into three main arms, each corresponding to a distinct clustering approach: TF-IDF, BoW, and LLMs as shown in Figure 1. For each arm, the dataset underwent specific preprocessing steps, and clustering was performed using the K-means algorithm. The choice of K-means was motivated by its simplicity, efficiency, and widespread use in customer segmentation studies. The dataset used in this study was sourced from Olist, a Brazilian e-commerce platform that connects small and medium-sized businesses with large online marketplaces. Olist's dataset, which is publicly available [19], includes a comprehensive collection of customer reviews across various product categories, providing a rich source of textual data for analysis. The dataset had around 40,500 customer reviews. The reviews in the dataset were in both English and Portuguese. For our analysis, we used an English version of this dataset where the reviews that were in Portuguese were also translated into English leaving the English ones as they were. It is important to note that ground-truth clusters, or predefined clusters known in advance, were not available in the dataset under investigation.

The effectiveness of each clustering approach was assessed using a combination of quantitative and qualitative evaluation metrics. To measure the internal cohesion and separation of clusters, the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index were employed. These metrics are three of the most popular techniques for internal clustering evaluation [20]. The Silhouette Score evaluates how similar a data point is to its own cluster compared to other clusters, with scores closer to 1 indicating well-defined clusters and scores closer to -1 suggesting overlapping clusters. The Davies-Bouldin Index measures the average similarity ratio of each cluster with the one most similar to it, where lower values indicate better clustering quality. The Calinski-Harabasz Index, also known as the Variance Ratio Criterion, assesses the ratio of the sum of between-cluster dispersion and within-cluster dispersion, with higher scores signifying more distinct and well-separated clusters. In addition to numerical metrics, visualizations (cluster plots) were employed to provide a holistic understanding of the clustering results.

***Using TF-IDF***

Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistic used to reflect how important a word is to a document in a collection or corpus [5, 21]. It is often used in text mining and information retrieval systems to evaluate the importance of a term to a document in a corpus [5, 21]. The TF-IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general [5, 21].

In this research, our approach to the application of TF-IDF is illustrated in Figure 2.

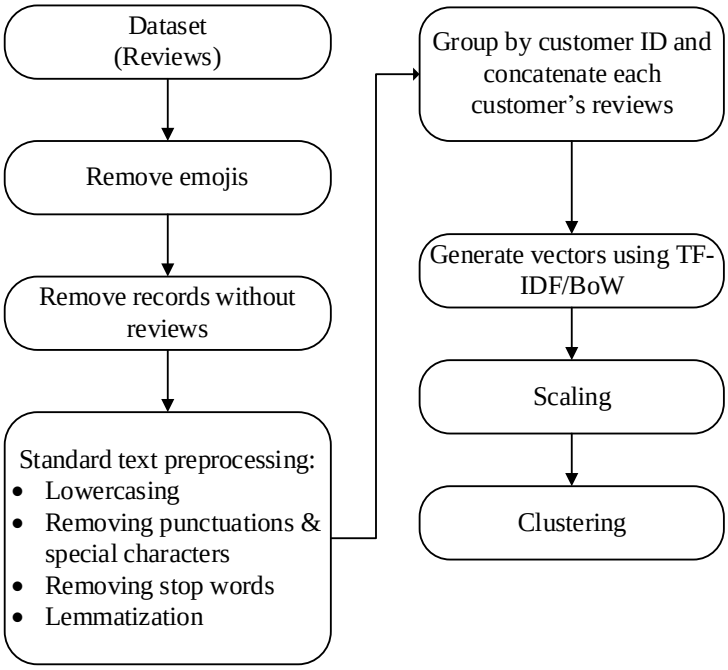


**Figure 2.** Methodology for performing TF-IDF and BoW

***Using Bag-of-Words***

The Bag-of-Words (BoW) model is a simplifying representation used in NLP and information retrieval. In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity [22]. The BoW model has been used in document classification where the occurrence (frequency) of each word is used as a feature for training a classifier [22].

The strategy followed for BoW was similar to that followed for TF-IDF as illustrated in Figure 2.

***Using Language Models***

Language Models like BERT, RoBERTa, XLNet, and MPNet have revolutionized NLP, offering a significant leap in our ability to capture the semantic richness and particulars of human language [23]. BERT captures nuanced meanings by processing text bidirectionally, understanding context from both directions simultaneously. RoBERTa builds on BERT by optimizing training procedures and focusing solely on masked language modeling, enhancing performance. XLNet introduces permutation-based training, combining bidirectional and autoregressive approaches to handle word dependencies more effectively. MPNet further refines these techniques by integrating both masked and permutation-based pre-training. These models are trained on massive text corpora and can generate contextualized word embeddings,

where the meaning of a word depends on its surrounding context [23]. These embeddings are simply vectors and hence standard vector interpretations should be applicable [24].

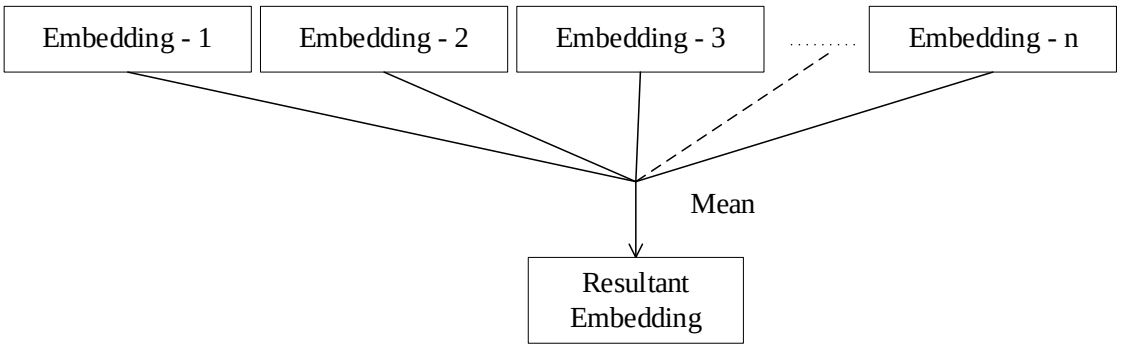


**Figure 3.** Resultant embedding for customer representation based on the mean of other embeddings

To come up with an overall sentence embedding (i.e., vector) that represents each customer, the fairest method would be to come up with an 'average' vector such that it represents a customer by taking into account all their reviews as shown in Figure 3. Hence, to arrive at this resultant vector, we calculated the mean value along each dimension. Figure 4 illustrates this intuition in a 2-D plane. Note that $c = (a + b)/2$ and $z = (x + y)/2$.

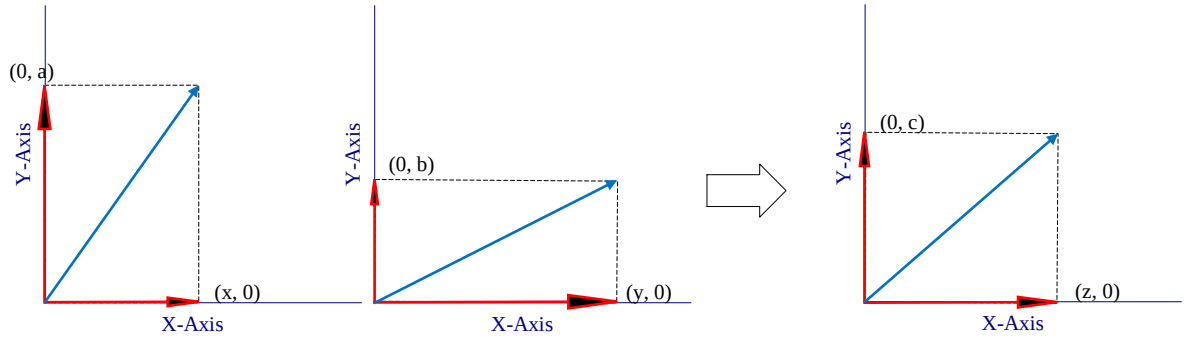


**Figure 4.** Using all embeddings of a customer to get resultant

## Experimentation and Results

The experiment focused on comparing the performance of customer segmentation after using TF-IDF, BoW, and various language models. Each approach underwent a systematic process, including data preprocessing, application of clustering algorithms, and evaluation.

### *TF-IDF*

The calculation of Term Frequency (TF) utilized a straightforward counting method, reflecting the frequency of each term within individual documents. The Inverse Document Frequency (IDF) employed the standard formulation, offering a measure of term significance across the entire document corpus. To refine the quality of TF-IDF embeddings, a custom vectorizer was designed with two critical filtering

mechanisms. Firstly, words appearing in over 50% of the documents were removed to eliminate ubiquitous terms lacking discriminative power. Secondly, words occurring in less than 1% of the documents were excluded to filter out rare terms introducing potential noise.

Furthermore, we concatenated all reviews for each customer before TF-IDF vectorization. This is because a customer may have multiple orders and each order would have at least one review. Hence a customer may have multiple reviews. By concatenating all reviews, we were able to associate all reviews provided by that customer while having only one record for that customer.

When performing TF-IDF we used simple count for TF and standard IDF as mentioned previously. The choice of the simple count method for TF was justified in this context, as it reflects the raw frequency of terms within individual documents, which is pertinent to customer reviews. The standard IDF formulation was deemed appropriate, as it penalizes terms that appear frequently across the entire document corpus, ensuring that common words receive lower weights in the TF-IDF embeddings.

We then used the K-means clustering algorithm to categorize customers into distinct segments. The value used for k was 3 as suggested by the elbow method.

For visualization purposes, we used Principal Component Analysis (PCA) to reduce the dimensionality. Figure 5 shows the visualization of the data points on a 3-D (to the left) and 2-D (to the right) plane.

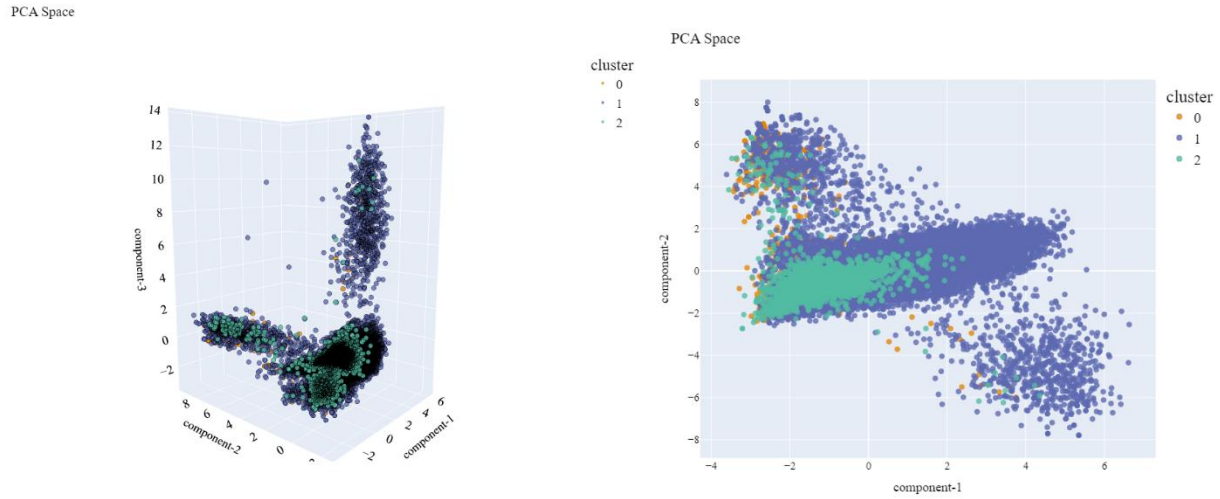


**Figure 5.** Clustering using vectors from TF-IDF (PCA)

As seen in Figure 5, there seems to be very little separation between clusters and hence differentiating one cluster from the rest is a challenging task, as the data points of one cluster overlap with those of another.

When using PCA, it is important to also consider the cumulative explained variance. It helps us understand how much of the total variance in a dataset is captured by a certain number of principal components. A value of 80% is generally considered good for descriptive purposes. However, in this case, the cumulative

explained variance of these three components of PCA was a low value (<50%). This low value could be due to the complex polynomial relationships between features. We used the t-distributed Stochastic Neighbor Embedding (t-SNE) method to capture these complex relationships. The visualizations are in Figure 6.

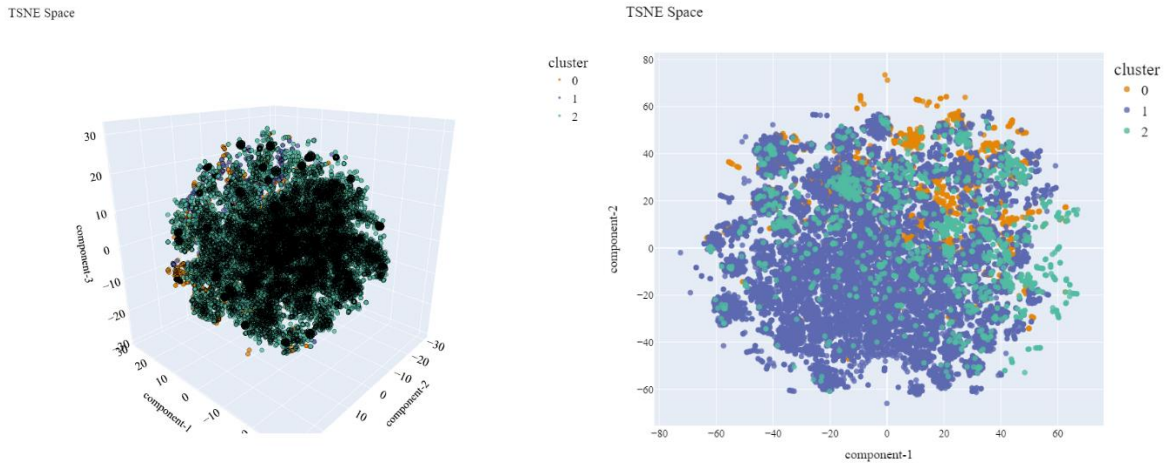


**Figure 6.** Clustering using vectors from TF-IDF (t-SNE)

Table 1 provides various clustering metrics. Based on these scores, we can conclude that the clustering algorithm used may not have produced well-defined clusters. The Calinski-Harabasz Index is somewhat high, indicating that there is a somewhat good separation between the clusters. However, the Davies-Bouldin Index is also relatively high, indicating that the clusters are not very similar to each other with low separation. The Silhouette score has a low value, indicating that the clusters are not well-separated.

**Table 1.** Clustering metrics for TF-IDF vectors

| Davies-Bouldin Index | Calinski-Harabasz Index | Silhouette Score |
|---|---|---|
| 4.10 | 1,438.85 | 0.05 |

***Bag-of-Words***

Similar to the preprocessing techniques employed in the TF-IDF experiment, our data underwent careful cleaning to ensure relevance and accuracy. Irrelevant information was removed, missing values were handled, and textual data were standardized. To enhance the quality of the BoW embeddings, we developed a customized count vectorizer that systematically excluded words that appeared in more than 50% of the documents and those occurring in less than 1% of the documents. This filtering mechanism aimed to eliminate words lacking meaningful content, including misspelled or improperly spaced terms such as "someth" and "ing" instead of "something."

Similar to the strategy used when generating vectors using TF-IDF, we concatenated all reviews for each customer before BoW vectorization. This was particularly crucial given that a single customer could

contribute multiple reviews. Aggregating these reviews aimed to provide a holistic representation of the customer's feedback.

The preprocessing steps were followed by the application of the custom Count Vectorizer, generating BoW embeddings for each customer based on the concatenated reviews. Subsequently, the K-means clustering algorithm was employed to categorize customers into distinct segments. The value used for k was 3. To arrive at this value, we utilized the elbow method.

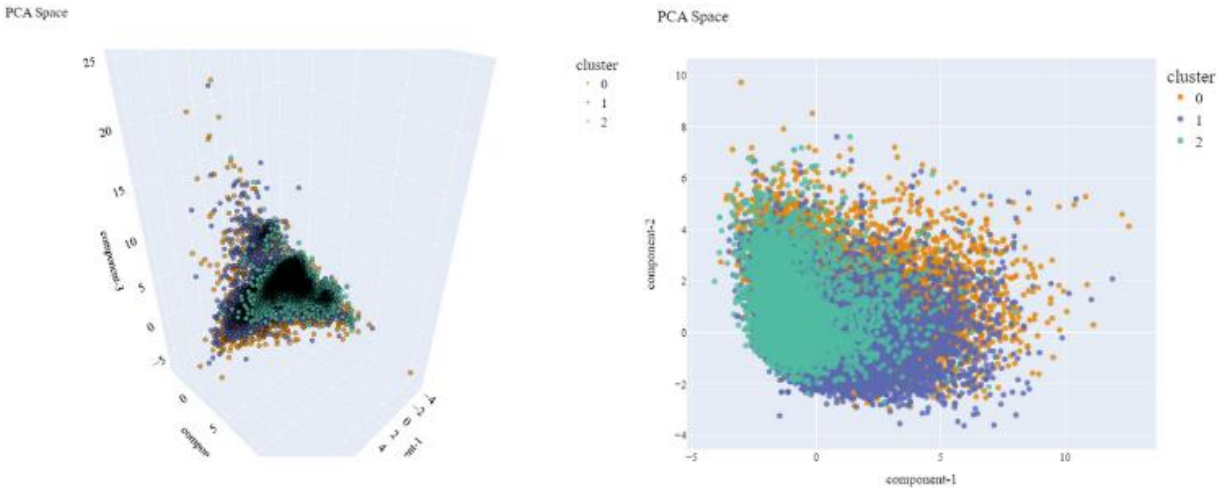


**Figure 7.** Clustering using vectors from BoW (PCA)

Figure 7 presents the visualizations for the BoW vectors. We used PCA to reduce dimensionality for visualization. However, the cumulative explained variance was not a satisfactory value. Hence, for better explanatory visualizations we used t-SNE. We arrived at the visualizations in Figure 8.

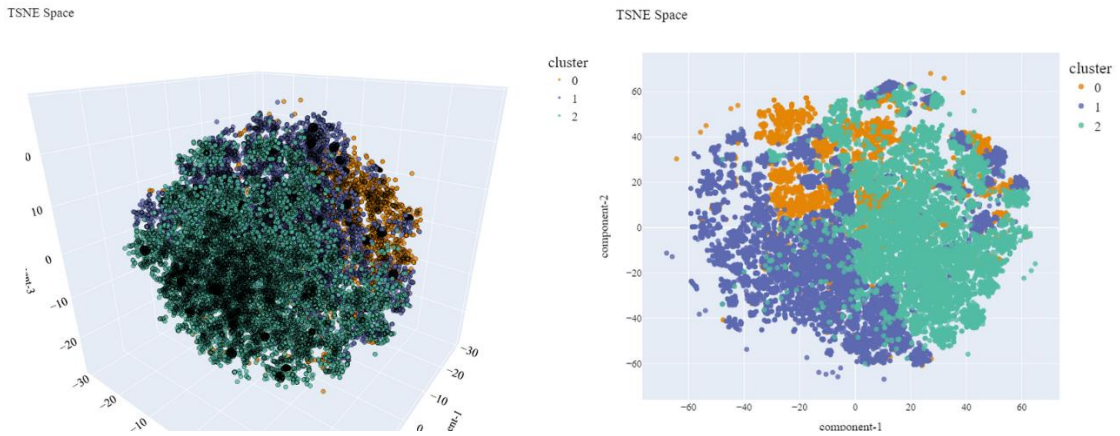


**Figure 8.** Clustering using vectors from BoW (t-SNE)

Based on the metrics we obtained in Table 2; clustering does not seem to be satisfactory. However, the metrics are better than those obtained for TF-IDF.

**Table 2.** Clustering metrics for BoW vectors

| Davies-Bouldin Index | Calinski-Harabasz Index | Silhouette Score |
|---|---|---|
| 3.10 | 2,440.00 | 0.08 |

***Large Language Models***

In this stage, we used LLMs to generate embeddings for customer reviews. The LLMs we used were RoBERTa, BERT, XLNet, and MPNet which were available to be downloaded via the 'Hugging Face' platform. After generating embeddings, we clustered them and compared their performance with that of TF-IDF and BoW based on Davies-Bouldin Index, Calinski-Harabasz Index, and Silhouette Score. We also compared the performance of clustering the embeddings generated from multiple LLMs as shown in Table 3.

**Table 3.** Clustering metrics for language model embeddings – before spelling correction

| LLM | Davies-Bouldin Index | Calinski-Harabasz Index | Silhouette Score |
|---|---|---|---|
| RoBERTa | 2.08 | 11,760.88 | 0.16 |
| BERT | 2.70 | 5,847.56 | 0.13 |
| XLNet | 3.03 | 3,767.20 | 0.10 |
| MPNet | 3.39 | 2,434.04 | 0.06 |

The clustering metrics in Table 3 provide insights into the performance of different LLMs, namely RoBERTa, BERT, XLNet, and MPNet. RoBERTa stands out with a Davies-Bouldin Index of 2.08, indicating well-defined clusters, and a high Calinski-Harabasz Index of 11,760.88, suggesting dense and well-separated clusters. The Silhouette Score of 0.16 further supports the notion of clear separation between clusters compared to the other language models. BERT follows closely with a slightly higher Davies-Bouldin Index of 2.70 and a lower Calinski-Harabasz Index and Silhouette Score, suggesting clusters that are somewhat less well-defined and less dense compared to RoBERTa. XLNet exhibits a higher Davies-Bouldin Index and Silhouette Score (3.03 and 0.10, respectively), indicating less well-defined clusters with

moderate separation. MPNet shows the lowest performance among the models, with a Davies-Bouldin Index of 3.39 and a Silhouette Score of 0.06, suggesting less well-defined and less separated clusters. In summary, the models vary in their ability to form distinct and dense clusters, with RoBERTa demonstrating superior clustering performance, followed by BERT, XLNet, and MPNet.

This stage prompted an investigation into the potential impact of spelling correction on clustering performance. We hypothesized that correcting spelling errors would enhance the language model's ability to comprehend the semantic content of the reviews, thereby leading to improved cluster formation. The results of this experiment are in Table 4.

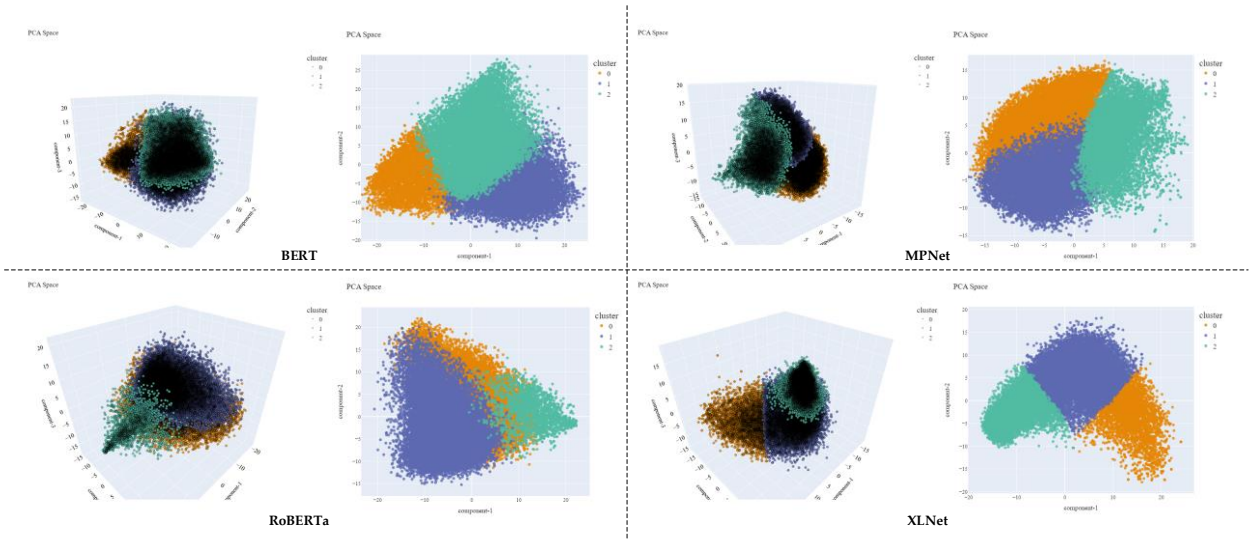


**Figure 9.** Visualizations of clusters – before spelling correction

**Table 4.** Clustering metrics for language model embeddings – after spelling correction

| LLM | Davies-Bouldin Index | Calinski-Harabasz Index | Silhouette Score |
|---|---|---|---|
| RoBERTa | 2.14 | 9,941.48 | 0.15 |
| BERT | 2.11 | 8,554.14 | 0.15 |
| XLNet | 3.13 | 3,430.69 | 0.10 |
| MPNet | 3.52 | 2,175.56 | 0.06 |

Upon comparing the clustering metrics before and after spelling correction for the LLMs—RoBERTa, BERT, XLNet, and MPNet—it becomes evident that the changes in metrics are generally subtle. For RoBERTa, there is a marginal increase in the Davies-Bouldin Index from 2.08 to 2.14, suggesting a slightly less well-

defined cluster structure after spelling correction (based on this metric alone). However, the Calinski-Harabasz Index remains high at 9941.48, and the Silhouette Score marginally declines from 0.16 to 0.15, indicating that the overall impact on clustering quality is not substantial. Similarly, BERT sees an increase in the Davies-Bouldin Index from 2.70 to 2.11, signifying marginally degraded cluster definition, while the Calinski-Harabasz Index increases to 8554.14 after spelling corrections. The Silhouette Score sees a slight increment to 0.15. Both XLNet and MPNet exhibit marginal increases in the Davies-Bouldin Index after spelling correction, implying slightly less well-defined cluster structures. The Calinski-Harabasz Indexes for both models decrease, indicating potential cluster density or separation reductions, while the Silhouette Scores remain relatively stable.

Overall, the changes in metrics, while noticeable, are not drastically significant, suggesting that the impact of spelling correction on the clustering performance of these models is nuanced. The subtlety of changes may be attributed to the inherent robustness of language models trained on diverse datasets, which allows them to handle variations in language, including misspellings, to some extent. Another reason might be attributed to the low percentage of misspelled words in the dataset, estimated at around 2%. With such a small proportion of misspelled words, the overall impact of spelling correction on the clustering metrics may be limited.

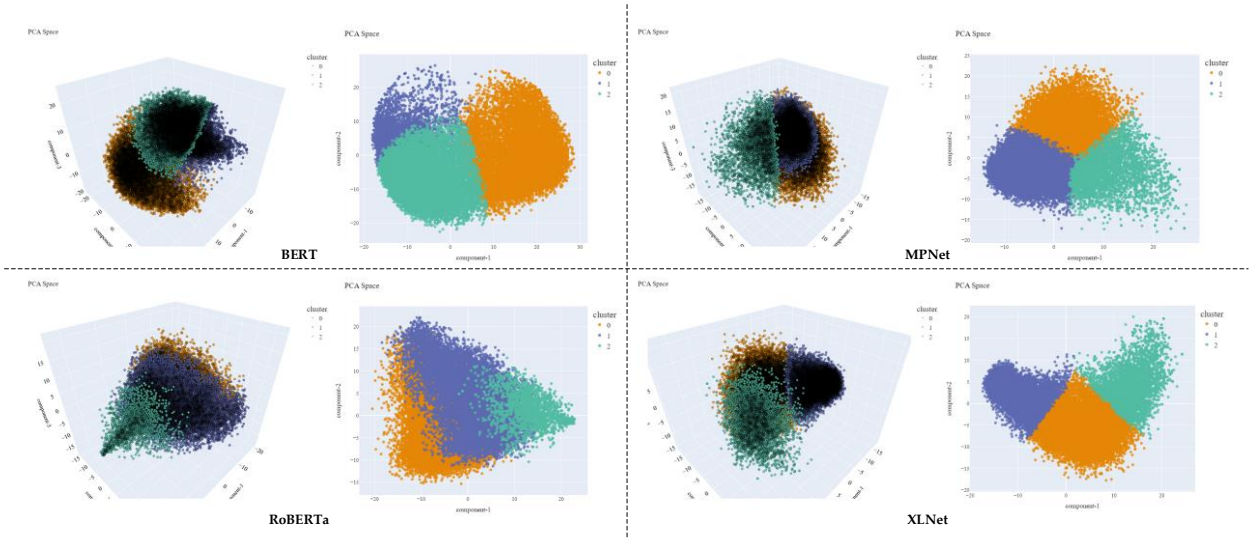


**Figure 10.** Visualizations of clusters – after spelling correction

Comparison between Figure 9 and Figure 10 justifies the conclusions made by the clustering metrics (i.e., there is no significant improvement in clustering even after spelling correction). We did not have to use t-SNE in this case since the cumulative explained variance of the 3 components when using PCA for visualizing clustered embeddings generated by the language models in each instance was satisfactory.

## Discussion

The comparative analysis of clustering approaches applied to customer reviews has yielded valuable insights into the effectiveness of traditional methods and language model-based approaches. Examining

the performance metrics and visualizations for TF-IDF, Bag-of-Words, and various language models, it is evident that the choice of representation method significantly influences the resulting customer segmentation. The TF-IDF approach, while commonly used in text mining, showed limited effectiveness in creating well-defined customer clusters, as indicated by the low Silhouette Score and high Davies-Bouldin Index. The Bag-of-Words model exhibited slightly improved results, emphasizing the importance of word occurrence frequencies in capturing some semantic nuances. Notably, language model-based embeddings, particularly those generated by RoBERTa, demonstrated superior clustering performance, with well-separated and dense clusters, as evidenced by comparatively higher Silhouette Scores and Calinski-Harabasz Indexes.

The discussion of language models extends beyond their performance to consider the practical implications of their application. While the experimentation with spelling correction yielded subtle improvements, it is crucial to acknowledge the nuanced impact of such corrections on clustering quality. Despite the overall effectiveness of language models in handling variations in language, the limited prevalence of misspelled words in the dataset may have contributed to the marginal changes observed post-correction.

The observed variations in clustering performance among different language models prompt further exploration into model-specific characteristics. For instance, RoBERTa consistently outperformed other models, suggesting that its training architecture and contextual embeddings play a crucial role in capturing the inherent structure of customer reviews. The discussion also underscores the importance of considering the trade-offs between computational resources, model complexity, and clustering performance, as these factors can influence the practical applicability of language models in real-world scenarios.

**Conclusion**

The primary objective of the study was to conduct a comparative analysis of clustering approaches for customer segmentation based on textual data, specifically customer reviews. By focusing on traditional methods such as TF-IDF and Bag-of-Words, as well as advanced language model-based approaches, the research aimed to determine the effectiveness of these techniques in creating meaningful and contextually rich customer segments. The primary hypothesis posited that the utilization of LLMs for generating sentence embeddings would result in more meaningful and contextually rich representations of textual information, leading to improved clustering performance compared to traditional methods. The results of this study support the hypothesis, as the language model RoBERTa demonstrated superior clustering performance, emphasizing the significance of leveraging state-of-the-art NLP models for customer segmentation in the digital age. The findings underscore the nuanced impact of representation methods on the formation of well-defined customer clusters. While TF-IDF and Bag-of-Words exhibited limitations in capturing semantic nuances, language models, especially RoBERTa, showcased enhanced capabilities in creating dense and distinct clusters. The subtle improvements observed post-spelling correction highlighted the robustness of language models in handling variations in language, albeit within the context of a dataset with a low prevalence of misspellings. As businesses increasingly rely on customer insights for tailored services, the implications of this research extend to practical applications in marketing and

customer relationship management. By understanding the strengths and limitations of various clustering approaches, businesses can make informed decisions on selecting the most suitable approach for customer segmentation, with this study affirming the advantages of LLMs in achieving more effective segmentation outcomes.

**Future Research Directions**

While the current study has provided valuable insights, there are avenues for further exploration. Future research could delve deeper into the interpretability of clusters generated by language models, examining the specific features that contribute to the effectiveness of RoBERTa in comparison to other models. Additionally, investigating the transferability of these findings to different industries or cultural contexts would contribute to the broader applicability of the proposed techniques. However, the current study lays a solid foundation for advancing the understanding of customer segmentation through sophisticated text analysis, paving the way for more targeted and personalized business strategies in the evolving landscape of digital transactions.

**Conflicts of Interest**

All the authors of this study state that there is no conflict of interest.

**References**

[1] Chen, Chiang, and Storey, Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 2012. 36(4): p. 1165. 10.2307/41703503.

[2] Zhao, Z., Liu, W., and Wang, K., Research on sentiment analysis method of opinion mining based on multi-model fusion transfer learning. Journal of Big Data, 2023. 10(1). 10.1186/s40537-023-00837-x.

[3] Rui, W., Xing, K., and Jia, Y., *BOWL: Bag of Word Clusters Text Representation Using Word Embeddings*, in *Knowledge Science, Engineering and Management*. 2016, Springer International Publishing: Cham. p. 3-14.

[4] Mounica, G., Bhavani, N.N., Rohit, P., and Suneetha, D., A novel method for fuzzy bag-of-words based on word clusters.

[5] Das, B. and Chakraborty, S., An improved text sentiment classification model using TF-IDF and Next Word Negation. arXiv [cs.CL], 2018. 10.48550/ARXIV.1806.06407.

[6] Akuma, S., Lubem, T., and Adom, I.T., Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets. International Journal of Information Technology, 2022. 14(7): p. 3629-3635. 10.1007/s41870-022-01096-4.

[7] Alves Gomes, M. and Meisen, T., A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. Information Systems and e-Business Management, 2023. 21(3): p. 527-570. 10.1007/s10257-023-00640-4.

[8] Luo, J., Qiu, S., Pan, X., Yang, K., and Tian, Y., Exploration of Spa Leisure Consumption Sentiment towards Different Holidays and Different Cities through Online Reviews: Implications for Customer Segmentation. Sustainability, 2022. 14(2): p. 664. 10.3390/su14020664.

[9] Wankhade, M., Rao, A.C.S., and Kulkarni, C., A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 2022. 55(7): p. 5731-5780. 10.1007/s10462-022-10144-1.

[10] Nandwani, P. and Verma, R., A review on sentiment analysis and emotion detection from text. Soc Netw Anal Min, 2021. 11(1): p. 81. 10.1007/s13278-021-00776-6.

[11] Subhashini, L.D.C.S., Li, Y., Zhang, J., Atukorale, A.S., and Wu, Y., Mining and classifying customer reviews: a survey. Artificial Intelligence Review, 2021. 54(8): p. 6343-6389. 10.1007/s10462-021-09955-5.

[12] Pankaj, Pandey, P., Muskan, and Soni, N., *Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews*, in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. 2019, IEEE. p. 320-322.

[13] Wei, C., Wang, Y.-C., Wang, B., and Kuo, C.C.J., An overview on language models: Recent developments and outlook. arXiv [cs.CL], 2023.

[14] Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A., A comprehensive overview of large Language Models. arXiv [cs.CL], 2023.

[15] Khurana, D., Koli, A., Khatter, K., and Singh, S., Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl, 2023. 82(3): p. 3713-3744. 10.1007/s11042-022-13428-4.

[16] Adoma, A.F., Henry, N.-M., and Chen, W., *Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition*, in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. 2020, IEEE. p. 117-121.

[17] An, Y., Kalinowski, A., and Greenberg, J., Clustering and network analysis for the embedding spaces of sentences and sub-sentences. arXiv [cs.CL], 2021.

[18] Li, B., Zhou, H., He, J., Wang, M., Yang, Y., and Li, L., *On the Sentence Embeddings from Pre-trained Language Models*, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, Association for Computational Linguistics. p. 9119-9130.

[19] Olist and Sionek, A., *Brazilian E-Commerce Public Dataset by Olist*. 2018, Kaggle.

[20] Liu, G., A new index for clustering evaluation based on density estimation. arXiv [cs.LG], 2022.

[21] Hasan, T. and Matin, A., *Extract Sentiment from Customer Reviews: A Better Approach of TF-IDF and BOW-Based Text Classification Using N-Gram Technique*, in *Proceedings of International Joint Conference on Advances in Computational Intelligence*. 2021, Springer Singapore: Singapore. p. 231-244.

[22] Bouachir, W., Torabi, A., Bilodeau, G.-A., and Blais, P., A bag of words approach for semantic segmentation of monitored scenes. arXiv [cs.CV], 2013.

[23] Mirjalili, S., Wu, J., Akhtar, N., Shaikh, M.B., Zafar, A., Irfan, M., muneer, a., Shah, A., Qureshi, R., tashi, q.a., and Hadi, M.U., Large Language Models: A comprehensive survey of its applications, challenges, limitations, and future prospects. 2023. 10.36227/techrxiv.23589741.v3.

[24] Wieting, J. and Kiela, D., No training required: Exploring random encoders for sentence classification. arXiv [cs.CL], 2019.