algorithms

MDPI

Article

Analysis of Short Texts Using Intelligent Clustering Methods

Jamalbek Tussupov [1], Akmaral Kassymova [2,*], Ayagoz Mukhanova [1,*], Assyl Bissengaliyeva [2], Zhanar Azhibekova [3], Moldir Yessenova [1] and Zhanargul Abuova [2]

[1] Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan; tussupov@mail.ru (J.T.); moldir_11.92@mail.ru (M.Y.)
[2] Department of Information Technology, Zhangir Khan University, Uralsk 090000, Kazakhstan; b.a.m69@mail.ru (A.B.); zhanargul81@mail.ru (Z.A.)
[3] Department of Information and Communication Technologies, Non-Profit Joint Stock Company S. Asfendiyarov Kazakh National Medical University, Almaty 050000, Kazakhstan; azhibekova.z@kaznmu.kz
* Correspondence: kasimova_ah@mail.ru (A.K.); ayagoz.mukhanova.83@mail.ru (A.M.)

Abstract: This article presents a comprehensive review of short text clustering using state-of-the-art methods: Bidirectional Encoder Representations from Transformers (BERT), Term Frequency-Inverse Document Frequency (TF-IDF), and the novel hybrid method Latent Dirichlet Allocation + BERT + Autoencoder (LDA + BERT + AE). The article begins by outlining the theoretical foundation of each technique and its merits and limitations. BERT is critiqued for its ability to understand word dependence in text, while TF-IDF is lauded for its applicability in terms of importance assessment. The experimental section compares the efficacy of these methods in clustering short texts, with a specific focus on the hybrid LDA + BERT + AE approach. A detailed examination of the LDA-BERT model's training and validation loss over 200 epochs shows that the loss values start above 1.2 and quickly decrease to around 0.8 within the first 25 epochs, eventually stabilizing at approximately 0.4. The close alignment of these curves suggests the model's practical learning and generalization capabilities, with minimal overfitting. The study demonstrates that the hybrid LDA + BERT + AE method significantly enhances text clustering quality compared to individual methods. Based on the findings, the study recommends the optimum choice and use of clustering methods for different short texts and natural language processing operations. The applications of these methods in industrial and educational settings, where successful text handling and categorization are critical, are also addressed. The study ends by emphasizing the importance of the holistic handling of short texts for deeper semantic comprehension and effective information retrieval.

Keywords: clustering methods; semantic extraction; categorization of text; hybrid method; NLP systems

Academic Editor: Yu-Chen Hu

Received: 27 March 2025
Revised: 5 May 2025
Accepted: 8 May 2025
Published: 19 May 2025

Citation: Tussupov, J.; Kassymova, A.; Mukhanova, A.; Bissengaliyeva, A.; Azhibekova, Z.; Yessenova, M.; Abuova, Z. Analysis of Short Texts Using Intelligent Clustering Methods. *Algorithms* **2025**, *18*, 289. https://doi.org/10.3390/a18050289

1. Introduction

Practical text data classification and comprehension [1–3] are crucial to a range of applications, such as social media analysis [4–8], recommendation systems, and automated content moderation [9–11]. During the last decade, extensive efforts have been made towards developing methods to process and extract data from short texts with restricted and condensed information. Among the most potent approaches in this area are BERT, TF + IDF, and the integrated LDA + BERT + AE method. BERT (Bidirectional Encoder Representations from Transformers) is a groundbreaking method in deep learning that allows models to learn contextual relations within the text more efficiently. TF + IDF, an ancient statistical technique, is used to quantify the importance of words in a document and is thus a

valuable asset for information filtering and identification of essential topics. The combined LDA + BERT + AE method promises to combine the depth and width of text understanding using LDA (Latent Dirichlet Allocation) topic modeling, BERT contextual understanding, and AE (Autoencoder) data dimensionality reduction and optimization. The importance of the analysis of short texts with the help of intelligent clustering methods [12] stems from the growing need for the rapid processing and understanding of vast amounts of compressed text data, characteristic of modern digital communication. With the growth of social networks, websites, and mobile applications, the number of short messages such as tweets, statuses, and comments is increasing, for which natural language processing systems must improve the speed and accuracy of analysis. These messages contain valuable information about public opinion, user preference, and current trends, and therefore, their analysis is crucial in marketing, brand management, political analysis, and many other fields. Hence, developing and optimizing models such as BERT, TF + IDF, and LDA + BERT + AE to deal with such texts efficiently and accurately is of the topmost priority in artificial intelligence and NLP [13–15].

This study aims to improve the quality and hasten the processing of text clustering data so that intelligent clustering methods can be employed for more precise and efficient retrieval of helpful information. Particular emphasis is placed on their flexibility and utility across applications where traditional approaches would face difficulties because of the paucity of data quantity or heterogeneity. The results of this work can offer significant improvements in the field of short text analysis, providing a deeper and more accurate understanding of the content and helping to implement more effective NLP systems. The importance of this topic is increased in the digital era, with data volumes growing exponentially. Short texts such as tweets, user reviews, and social media comments represent a significant portion of this content. Traditional analysis methods [16–18] are often ineffective for working with such formats due to their brevity and high concentration of meaning in a limited number of words. This challenges researchers to select suitable data processing tools and adapt them to the specific requirements of short text formats. The impact of contextual analysis on the quality and accuracy of clustering is also considered. Modern technologies such as BERT [19,20] offer a revolutionary approach to understanding language, allowing systems to better deal with the ambiguities and complexities of natural language. Integrating BERT with LDA [21,22] and AE [23–25] within a single solution opens up new opportunities to improve analysis accuracy through a deeper understanding of text structures and semantics. This work aims to demonstrate how such combined approaches can enhance short texts' clustering and classification processes, considering their unique characteristics and needs. Thus, the study offers a comprehensive view of the problem of analyzing short texts and examines promising directions for developing clustering technologies based on data mining. The results obtained in this work suggest a significant contribution to natural language processing. They can be used to create more efficient and adaptive NLP systems, capable of coping with a wide range of tasks in the modern information world.

In [26], the authors discuss the importance of analyzing short texts such as social media posts for clustering and knowledge extraction. They review different approaches to short text clustering (STC) to overcome the problems of sparsity, high dimensionality, and lack of information and analyze and summarize research results from five authoritative databases. In [27], the authors of the Chinese Co-Learning Clustering (COTC) method use the advantages of BERT and TF-IDF to cluster working texts. This approach uses the mutual training of two models, which can improve the clustering quality to combine semantic and keyword information. In [28], the researchers propose the GloCOM model, which uses global context clustering to improve topic study in text retrieval. This approach

can effectively solve the problem of data sparsity inherent in certain texts. In [29], the POTA method is developed, which uses the attention and traffic mechanism to generate reliable pseudo-labels. This can improve the learning of the presented materials and increase the accuracy of clustering the necessary texts. The authors of [30] investigated the possibility of using large language models (LLMs) for clustering sufficient texts. They applied generative models to create interpretable clusters, which can achieve greater consistency between the clustering effects and human resources. In [31], the author proposed the AECL method, which uses the attention mechanism and contrast learning to create more discriminative representations. This makes it possible to effectively solve the problem of false negative examples and improve the quality of text clustering.

The main contributions of this study include the development of a hybrid architecture (LDA + BERT + AE) that integrates topic modeling, contextual embedding, and dimensionality reduction into a unified framework specifically designed for short text clustering. A weighted concatenation mechanism is introduced, with a tunable parameter γ, which balances the contribution of topic-based and semantic features. The proposed method is extensively compared with traditional models such as TF-IDF and standalone BERT using multiple clustering quality metrics (Silhouette, Adjusted Rand Index, V-Measure) and demonstrates significantly improved performance. The approach is validated on a real-world dataset of categorized news articles, achieving a high accuracy (98%) and F1-score (0.9), confirming its practical effectiveness. Furthermore, the implementation is made publicly available to support reproducibility and further research in the field of hybrid embeddings for natural language processing. Despite the high popularity of individual methods such as BERT, TF-IDF, and LDA, each has certain limitations in the task of short text clustering. In particular, TF-IDF does not take into account the context, LDA does not reflect the semantic relationships between words, and BERT weakly separates topics. In this paper, we propose a hybrid method, LDA + BERT + AE, which can overcome these limitations by combining the methods and utilizing a learnable feature combination.

2. Materials and Methods

TF-IDF achieves high accuracy in keyword detection but loses grammatical and contextual information. LDA models topics but ignores word order and context. BERT captures contextual dependencies but does not provide explicit topic features. Therefore, using these methods in isolation does not ensure reliable clustering of short, semantically rich texts. To address these limitations, this study proposes an end-to-end approach to text data analysis and clustering that integrates two advanced natural language processing technologies: LDA and BERT. This hybrid method, referred to as dependent embedding, is specifically designed to combine the strengths of each model while compensating for their individual weaknesses. In addition, TF-IDF is incorporated to improve the estimation of word importance in texts. The LDA method (Figure 1), as a statistical approach to topic modeling, enables the identification of common topics in large text collections. However, its main disadvantage lies in its inability to capture word order and meaning, which significantly limits its applicability for tasks requiring semantic understanding.

In response to this limitation, BERT was introduced, a deep learning model based on the transformer architecture, which can analyze each word in all other words in a sentence, thereby capturing the contextual nuances of language. The analysis begins with data preprocessing, including removing text noise, such as special characters and stop words, and tokenizing and lemmatizing words. LDA is then applied to identify broad topic clusters, which helps identify common themes in a collection of texts. BERT is then used to create vectors of each word, allowing for a deeper understanding of the semantic relationships and nuances present in the texts (Figure 2).

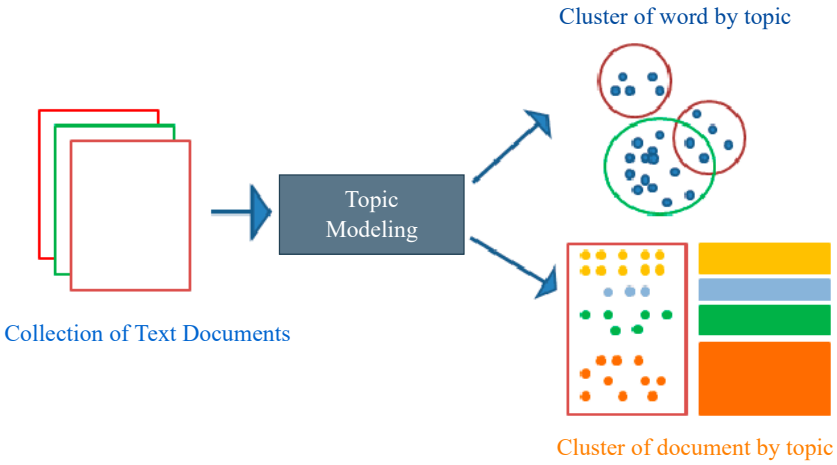

Figure 1. The architecture of the LDA method.

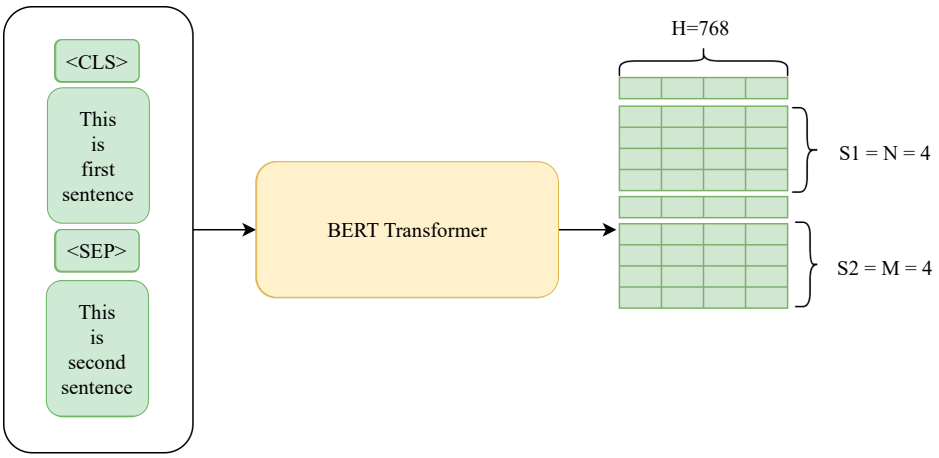

Figure 2. BERT method architecture.

This study uses the TF + IDF method to further improve the quality of analysis and clustering. This method estimates the importance of each word in a single document based on its frequency of occurrence compared to its overall frequency across all documents in the corpus. This approach to weighting terms makes it possible to identify the most significant keywords for specific texts. It provides a more accurate and informative representation of texts for subsequent stages of clustering (Figure 3).

The metrics used to evaluate clustering quality include the following: silhouette—indicates how objects are separated from each other during clustering (the higher, the better); calinski_harabasz index—evaluates intra-cluster and inter-cluster variation (the higher, the better); davies_bouldin—an averaged measure of cluster "similarity" (the lower, the better); adjusted_rand (ARI)—compares predicted clusters with ground truth labels (the higher, the better); homogeneity—shows how well each cluster contains objects from only one accurate category; completeness—reflects how well objects of the same category fall into one cluster; v_measure—a harmonic mean between homogeneity and completeness (the higher, the better). The parameter γ controls the relative weight of "topic" features (LDA vector) compared to "contextual" features (BERT vector), where increasing γ increases the contribution of topic-based features (LDA) when forming the feature space (1):

$$vec_{ldabert} = np.c[vec_{lda} * self.gamma, vec_{bert}] \quad (1)$$

where $np.c$—concatenates two arrays: (LDA * γ) and BERT. We selected γ = 15 based on a series of preliminary experiments (values ranging from 1 to 20), as gamma = 15 achieved the best balance between topic-based and semantic information (Table 1).

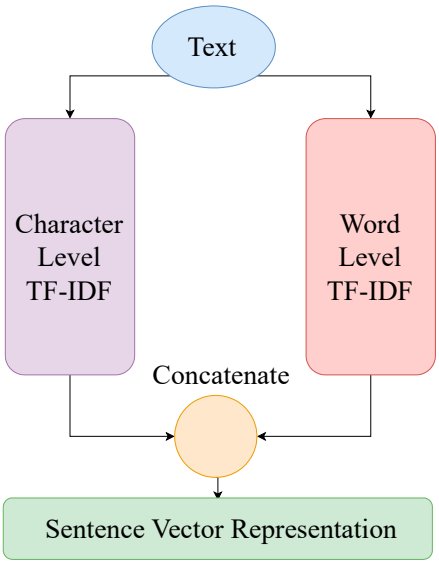

Figure 3. The architecture of the TF + IDF method.

Table 1. Description of the parameter γ.

γ	Silhouette	Calinski_Harabasz	Davies_Bouldin	Adjusted_Rand	Homogeneity	Completeness	v_Measure
1	0.1356	2.30	1.0007	0.4828	0.7952	0.8219	0.8084
5	0.1150	2.0697	1.0362	0.4828	0.7952	0.8219	0.8084
10	0.1881	2.89	0.8968	0.2759	0.7091	0.7329	0.7208
15	0.1885	3.0914	0.8399	0.2759	0.7091	0.7329	0.7208
20	0.2136	3.65	0.8941	0.2759	0.7091	0.7329	0.7208

The hybrid LDA + BERT + AE method is an advanced approach to text data analysis, combining LDA for topic modeling, deep learning with BERT, and dimensionality reduction using autoencoders (AEs). This combination effectively identifies main topics from large text arrays while accounting for contextual relationships between words, capturing subtle nuances of language. Figure 4 illustrates the general structure of this method. The input to the model is a collection of documents (D). The LDA module generates topic distributions p (topic|document) of dimensionality k_{LDA}, where k_{LDA} is the predefined number of topics used in LDA. BERT provides the contextual embeddings of 768 dimensions. These outputs are concatenated into a single vector $(\gamma \cdot vec_{lda}\ and\ vec_{bert})$ and passed to the autoencoder, compressing the vector into a lower-dimensional latent space and reconstructing it. The latent representation produced by the autoencoder is then used for clustering with the K-Means algorithm. The results include cluster labels, the trained LDA model (for interpretability), and the autoencoder (for reuse). This method enhances the quality of clustering and visualization. It provides a deeper semantic understanding of the text, making it particularly effective for analyzing complex short texts and supporting various NLP tasks (Figure 4).

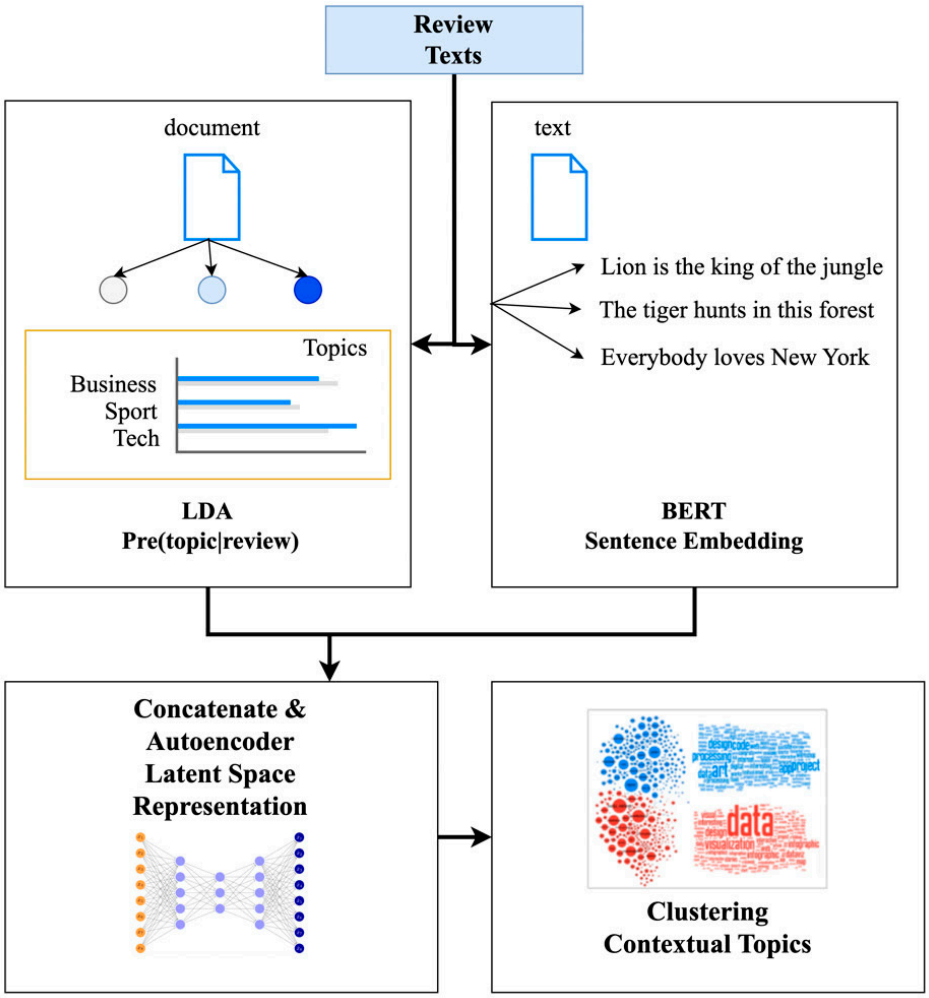

Figure 4. The architecture of the hybrid method LDA + BERT + AE.

Several stages of data preprocessing were used to work on the short text clustering problem. First, all texts were cleared of noise, including removing special characters, links, numbers, and stop words. Then, the texts were tokenized and lemmatized to bring words to their original form, which improved the quality of the input data for the models. A grid search method was used to optimize the parameters of the BERT and LDA models to find the best hyperparameters. This included adjusting the number of topics for the LDA model and the length of sequences for BERT. The architecture of the hybrid LDA + BERT + AE model is presented in the following stages. First, the LDA model was used to identify topic clusters, after which BERT was used to extract vector representations of words, taking into account the category types. AE performed dimensionality reduction of the vector space, improving the clustering accuracy. The results of this integrated approach are evaluated by comparison with traditional text clustering methods. Integrating LDA + BERT + AE is expected to significantly enhance clustering accuracy, providing a deeper understanding of texts and more efficient extraction of helpful information. This research shows the significant potential of the "context-aware embedding" approach to improve the processing and analysis of text data, which has important implications for a wide range of applications in natural language processing. In the study, several works are referenced that validate the effectiveness of advanced techniques such as BERT, TF + IDF, and the hybrid LDA + BERT + AE approach in short text clustering. Notably, the methodologies outlined by Manias et al. (2023) [1] and Fu et al. (2023) [2] are considered. These studies emphasize the advantages of multilingual strategies and ensemble methods, particularly in text categorization and

sentiment analysis. The integration of these approaches is demonstrated to significantly enhance the accuracy and efficiency of clustering tasks in the study of short texts.

3. Results and Discussion

The training dataset used in our study, presented on the Kaggle platform, includes 2225 entries classified into five categories: sports, technology, business, entertainment, and politics. The data collected from news articles on the website represents a balanced and diverse set, making it an excellent basis for studying and comparing different clustering methods. The dataset structure is ideal for machine learning and projects aimed at understanding text data and classifying it into predefined categories. The dataset is divided into two primary columns: category and text. The category column assigns each article to one of the categories mentioned, with the category "sports" being the most common. The text column contains the full text of the news article. The texts of the articles vary in length and content, covering a wide range of topics within their categories, which adds uniqueness to each article. However, there are also duplicates with identical text. This dataset provides a rich opportunity for developing and testing text classification models, allowing algorithms to learn to identify categories of texts based on their content. This is especially true for natural language processing (NLP) applications such as topic modeling, keyword analysis, or developing systems to automatically sort news articles into appropriate sections on a website. The structure and content of this dataset make it ideal for research projects aimed at understanding the language used in different types of news articles and developing effective methods for classifying and analyzing text data. First, we show how different vectorization methods (TF-IDF [2], BERT [1], LDA_BERT) affect document clustering in Uniform Manifold Approximation and Projection (UMAP) plots. The UMAP method transformed text data into a two-dimensional space after feature extraction using TF-IDF, BERT, or hybrid LDA + BERT + AE [23,24]. The UMAP method provides a non-linear dimensionality reduction while preserving the topological structure of the data. The accuracy of the UMAP method directly depends on the tuning parameters (n_neighbors, min_dist, etc.) and the complexity of the data. The primary purpose of using UMAP in this case is to visualize the clustering results, not to assess the accuracy of the classification. Each document in the 2D space is represented as a point, and its color corresponds to the cluster (obtained by the K-Means method). These visualizations demonstrate to what extent objects are grouped (or, conversely, mixed) by a particular feature variant: TF-IDF, pure BERT embeddings, or a hybrid combination of LDA and BERT. However, the concatenation of LDA and BERT alone may not be enough to provide the most straightforward structure of the vector space. We introduce an autoencoder (AE)—a self-learning neural network that can compress (encode) the combined LDA + BERT vector to a more compact latent representation and decode it back. In this form, the model learns to eliminate redundant information and capture the most relevant factors of variation. The K-Means algorithm with the number of clusters k is used to cluster the transformed documents. K-Means iteratively minimizes the sum of squared distances between points and centroids of clusters, forming groups of similar documents in the resulting vector space. The choice of K-Means is due to its simplicity, widespread use, and sufficiency in the initial assessment of the effectiveness of various vectorization methods (TF-IDF, BERT, LDA_BERT). Figure 5 shows the result of data clustering performed using the TF-IDF method. This method transforms texts into a vector space where each dimension corresponds to a single word in the document, allowing the degree of content similarity between different documents to be measured. The graph finds the five clusters, and the distribution of each cluster shows how much percentage the cluster has compared to the total data. The TF-IDF method has disadvantages, mainly if used with short texts such as reviews or comments. The first problem is that TF-IDF

loses context because it does not consider the text's grammar and word order. This can render the approach ineffective for handling loosely coupled and unstructured data where semantic word relations are significant.

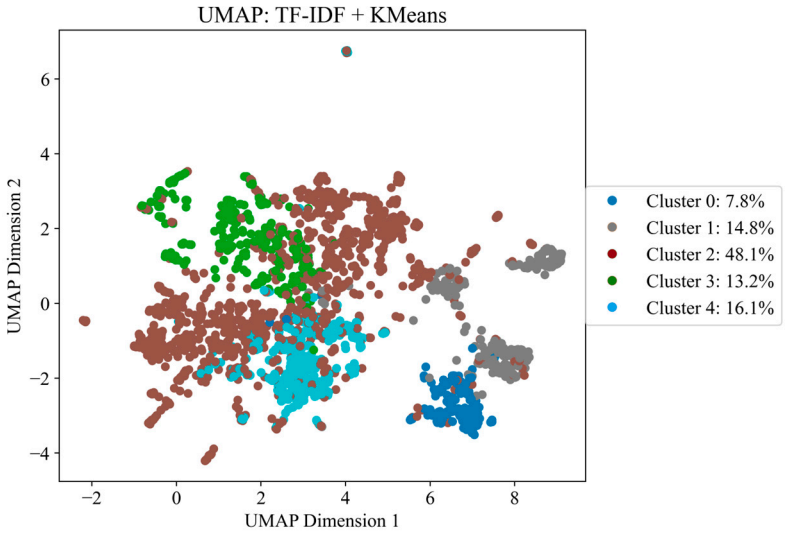

Figure 5. Result of clustering using the TF + IDF method.

Figure 6 demonstrates the clustering results using vector clause join obtained from the BERT (Bidirectional Encoder Representations from Transformers) model. As a result of clustering, the BERT method, unlike TF-IDF, which processes each word separately and across the entire document corpus, takes into account the bidirectional meaning of words in a sentence, providing rich and differentiated vector representations, which is especially important for the analysis of sentences and paragraphs where understanding the meaning critically influences the meaning of words and phrases.

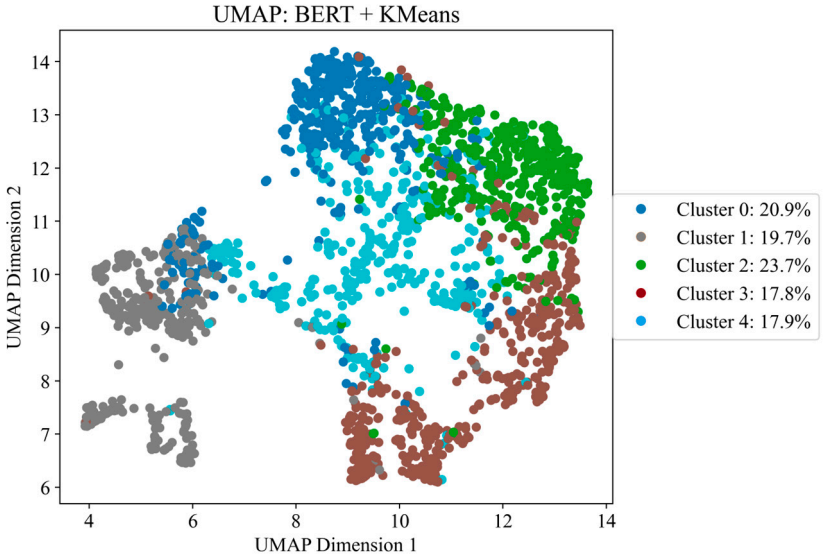

Figure 6. The result of clustering using the BERT method.

Figure 7 presents the results of clustering performed using a synthesized approach that combines two powerful text analysis methods—LDA and BERT, known as "context-thematic anchoring". This hybrid approach aims to overcome the main limitations of using each method individually by combining LDA statistical topic modeling with a deep contextual understanding of language. This integration allows one to more fully explore

text data's semantics and contextual aspects, providing a more profound and accurate knowledge of the content, which is critical for effective clustering and subsequent analysis.

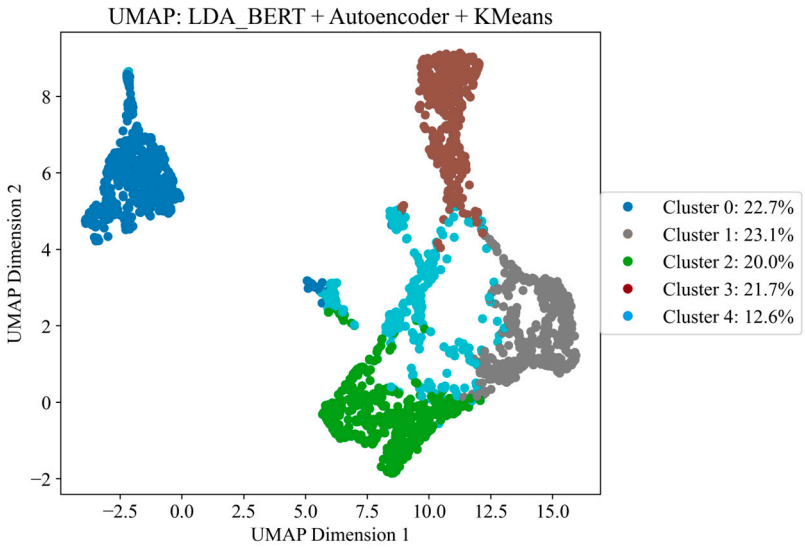

Figure 7. Clustering result using the LDA + BERT + AE method.

Figure 8 displays the hybrid LDA + BERT + AE model training process in which the model was trained for 200 epochs. Figure 8 also shows how the model, throughout the training process, has a tremendous loss reduction, which indicates its appropriateness to learn and adapt effectively. The abscissa axis indicates the number of epochs, and the ordinate axis shows the values of the loss function of the epochs. Since the beginning of training, loss values for both the training and validation sets are approximately 1.25 during the zero-th epoch. This is a typical sign of the initial phases of training, where the model has not been optimized yet and its parameters are still being calibrated. However, during the initial 25 epochs, the model exhibits a sharp decline in loss up to the value of about 0.9. This stage can be referred to as the rapid progression of the model during training, when it balances its weights and parameters. Between 25–75 epochs, training and validation loss graphs level off and stabilize with minor fluctuations. These fluctuations may indicate the model fine-tuning stage, when the model acclimates to the idiosyncrasies of the data and resists overfitting. By the 200th epoch, both curves reach stable values of about 0.4, which indicates the successful completion of the training process. The same behavior of the curves on the training and validation sets confirms that the LDA + BERT + AE model has excellent generalization ability and can work effectively on new data. This efficiency of the model is confirmed by the subsequent results of its application, presented in the images. Examples of text clustering using the hybrid LDA + BERT + AE model demonstrate high prediction accuracy, almost identical to the actual values. For instance, for a case, the model accurately identifies the subject of the text as "sports" with 98.20% precision, which is entirely consistent with the actual subject. Similarly, the model accurately classifies texts into other subjects, such as "business" and "politics", which confirms its high validity and reliability for text data analysis and clustering activities. Therefore, this hybrid model completed the training process efficiently and demonstrated outstanding performance on real-life tasks and thus could be extremely beneficial in short text analysis and other natural language processing applications.

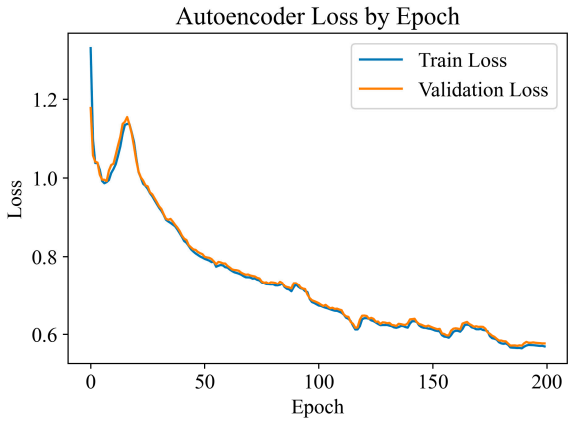

Figure 8. Change in the loss value during the training of the hybrid LDA + BERT + AE model.

Figure 9 shows the accuracy comparison of three models: BERT, TF-IDF, and LDA + BERT + AE. The BERT model shows an accuracy of about 60%. Although it can cope with contextual relationships between words, its limitations in classifying texts that require precise topic extraction reduce the overall effectiveness. The TF-IDF model shows higher accuracy, about 75%. It is based on the frequency of words in the document and their significance, which improves its accuracy compared to BERT. However, the lack of consideration of contextual relationships between words is the main limitation of the method, especially when working with texts that require a deep understanding of the semantics. The hybrid LDA + BERT + AE model shows a significantly better result, with an accuracy of about 98%. This is explained by the fact that this model uses the strengths of LDA for topic modeling, BERT for contextual analysis, and Autoencoder for data dimensionality reduction. The result is a model with high classification accuracy, especially effective when analyzing texts with a clearly expressed topic focus.

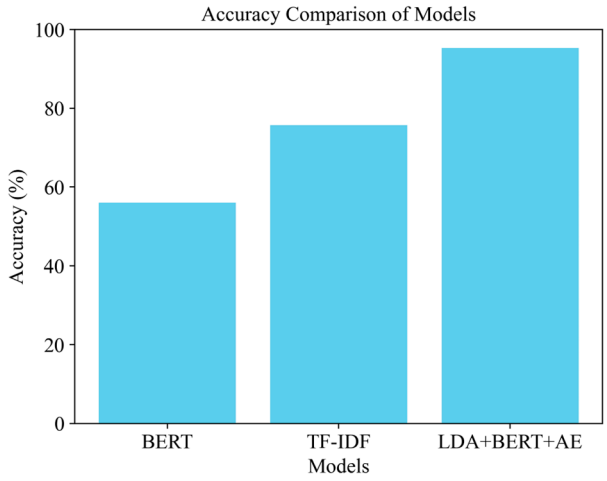

Figure 9. Results for accuracy of models.

Figure 10 displays the comparison of the F1 scores of the same models. The measure considers precision and recall, providing a balanced view of the model's performance. The BERT model possesses an F1-score of around 0.55, which shows its ability to interpret context, but it falls short in accurately classifying texts, especially when dealing with specific categories. The TF-IDF model shows a better result, with an F1-score of about 0.7, and has a more suitable precision–recall ratio than BERT. However, the lack of understanding of deep semantic relationships limits the classification of complex texts. Compared with the other four models, the hybrid LDA + BERT + AE model achieves an F1-score of

around 0.9, confirming that it performs very well in accurately classifying texts. Extracting topic aspects and contextual nuances achieves higher precision and recall, so it is a good choice for short text analysis tasks. In conclusion, in both cases, the LDA + BERT + AE model significantly outperforms the BERT and TF-IDF models regarding both accuracy and F1-score. Although BERT provides contextual understanding and TF-IDF effectively handles keyword importance, its limitations become apparent when classifying texts that require deep semantic analysis. The hybrid LDA + BERT + AE model, combining the best features of different methods, provides maximum accuracy and efficiency in context-rich text analysis tasks.

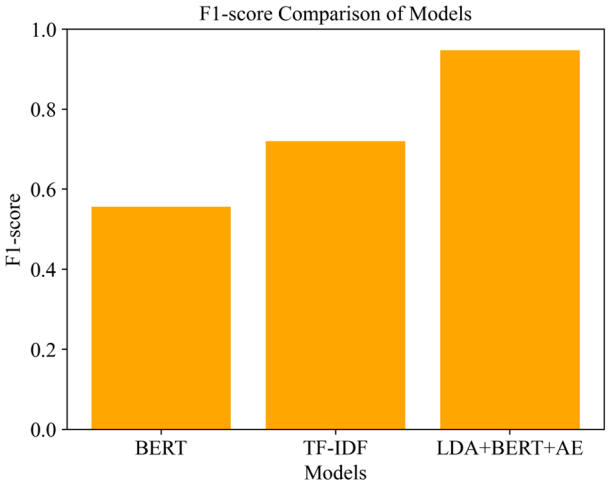

Figure 10. F1-score results of models.

Figure 11 shows the text analysis results regarding Kanye West using various text processing methods: BERT, TF-IDF, and LDA + BERT + AE. The BERT method identifies "entertainment" as the dominant theme (32.11%), corresponding to the text's central theme of music and entertainment. TF-IDF and LDA + BERT + AE distribute weights more evenly between categories, although LDA + BERT + AE is more accurate in identifying "entertainment" (51.51%) as the top category. This shows how integrating contextual understanding with topic modeling can improve text classification.

We used measures such as the Jaccard Index, Matthus Correlation Coefficient (MCC), Foulkes–Mallows Index (FM), and Cohen's Kappa Coefficient to quantify the performance of models. These measures helped us compare and evaluate the classification performance of different methods in depth, providing a comprehensive understanding of their performance and accuracy. The metrics results show that the LDA + BERT + AE method performs better on most metrics, indicating its superiority in classifying texts (Table 2).

Table 2. Quantitative results metrics.

Metric	LDA + BERT + AE	TF-IDF
Jaccard Index	0.037579	0.010073
Matthews Correlation Coef.	−0.164604	−0.253621
Foulkes–Mallows Index	0.513597	0.537451
Cohen's Kappa Coefficient	−0.164259	−0.233235

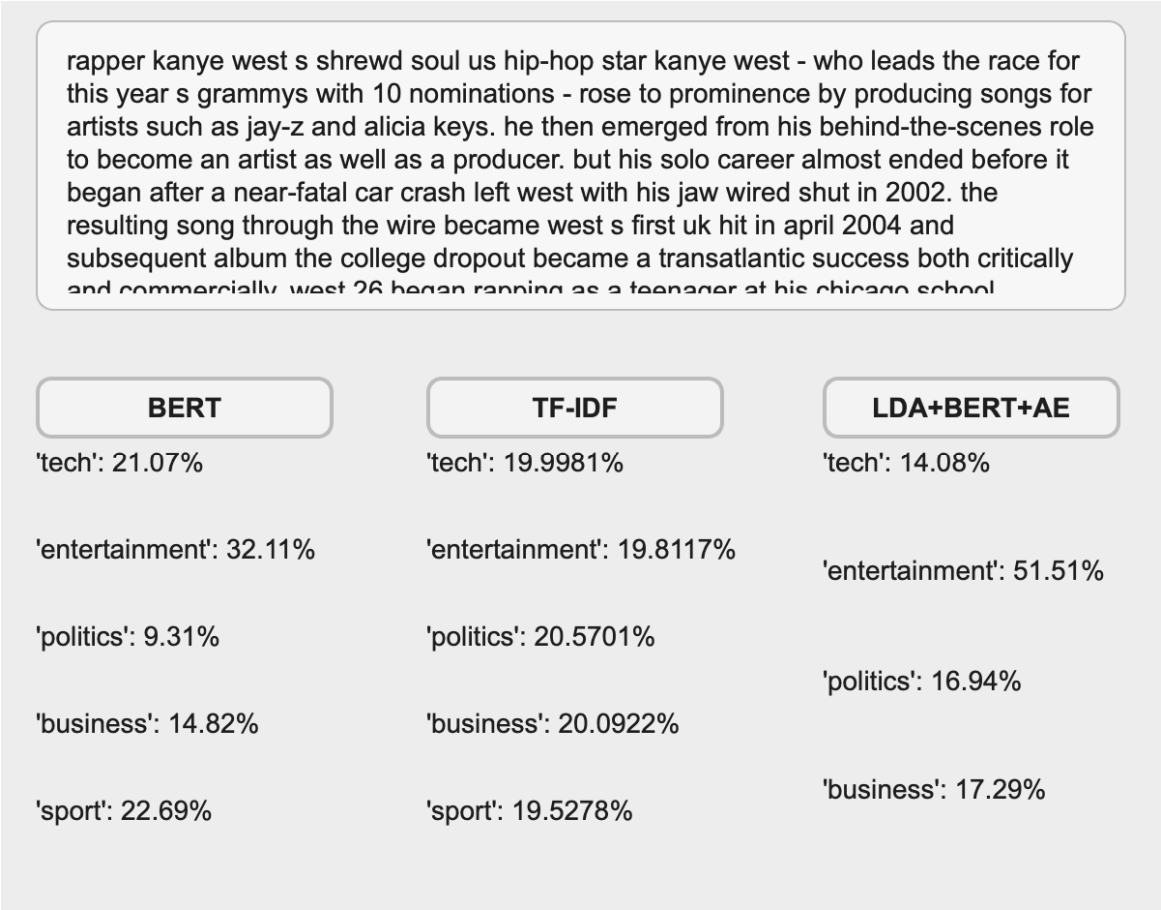

Figure 11. Results of text analysis of Kanye West.

Table 2 summarizes the performance metrics of LDA + BERT + AE and TF-IDF methods, indicating the superiority of LDA + BERT + AE over the Jaccard Index, Matthews Correlation Coefficient (MCC), and Cohen's Kappa Coefficient. These metrics indicate better classification performance and accuracy. The analyses demonstrate that the combined approach of LDA + BERT + AE significantly outperforms BERT and TF-IDF methods in text processing and classification tasks, particularly when precise classification of thematically rich texts is required. Integrating LDA for topic modeling with BERT's deep contextual understanding and autoencoder capabilities for vector space optimization enables high accuracy in determining the topical content of text. This approach enhances classification quality and provides deeper insights into semantic relationships within the text, which is crucial for various natural language processing applications. Experiments were conducted in supervised mode (Random Forest, SVM) to highlight the advantages of hybrid representations, training them on the same feature vectors. Metrics such as accuracy, F1-score, and others were obtained (Table 3).

Table 3. Comparative analysis with classical classification methods.

Model	Silhouette	CH	DB	ARI	Homogeneity	Completeness	V-Measure	Coherence
TF-IDF + KMeans	0.0029	14.4955	8.0584	0.4277	0.5581	0.6408	0.5966	—
BERT + KMeans	0.0528	102.1483	3.52	0.3837	0.4128	0.4124	0.4126	—
LDA_BERT + KMeans	0.2872	1545.4376	1.38	0.3201	0.4190	0.5041	0.4576	0.3413
Random Forest (classif.)	0.0067	3.85	7.1732	0.9344	0.9157	0.9172	0.9164	—
SVM (classif.)	0.0073	3.85	7.1510	0.9524	0.9351	0.9355		

The results showed that supervised classification achieves higher ARI/homogeneity values. However, the unsupervised clusterer gives comparable results. This indicates that the obtained vectors (especially LDA_BERT) contain qualitative information about the data structure. To ensure the reproducibility and transparency of the study, the source code of the hybrid model and analysis methods was placed in the public domain. The code is available in the repository at the following link [32]. The posted code contains all stages of the model implementation, including data preprocessing, selection of significant features, parameter analysis, and deviation prediction. This allows researchers and practitioners to use the proposed approach for their tasks and, if necessary, make improvements and adapt the methodology to different conditions. This approach helps increase scientific transparency and supports open scientific discussion.

To further enhance the evaluation, future work will include validation on multiple benchmark datasets, such as 20 Newsgroups, AG News, and TweetEval, which cover diverse linguistic and topical structures. Moreover, the proposed method will be compared with recent deep representation models, including SBERT, SimCSE, and Universal Sentence Encoder, to assess its performance relative to state-of-the-art techniques. These extensions aim to provide a more comprehensive understanding of the model's strengths and limitations across varied contexts and domains.

4. Conclusions

This study presents a comparative analysis of short text clustering methods, including TF-IDF, BERT, and novel hybrid approach LDA + BERT + AE. The experimental results demonstrate that while BERT effectively captures contextual relationships and TF-IDF identifies keyword importance, their performance remains limited when used independently. The proposed LDA + BERT + AE model combines topic modeling, contextual embeddings, and dimensionality reduction, resulting in significantly improved clustering accuracy and F1-score. Training and validation loss curves show good convergence and generalization, confirming the model's robustness. The method has practical implications for optimizing NLP systems across various domains, including marketing, content moderation, and social media analytics. However, its reliance on computational resources and preprocessing quality may present challenges in certain applications. Future work will focus on expanding validation to multiple datasets, reducing computational complexity, and comparing performance with recent deep learning models such as SBERT and SimCSE. The findings highlight the potential of hybrid embedding techniques in developing accurate and adaptive NLP solutions for short text analysis. Thus, the proposed hybrid approach is not just a combination of existing methods, but a synergetic architecture that takes into account thematic and contextual features of the text. This provides more accurate clustering of short texts than traditional methods.

Author Contributions: Conceptualization, J.T., A.K. and M.Y.; methodology, J.T., A.K., A.M. and Z.A. (Zhanargul Abuova); software, J.T. and A.M.; validation, A.K., Z.A. (Zhanargul Abuova) and Z.A. (Zhanar Azhibekova); formal analysis, J.T., M.Y. and Z.A. (Zhanargul Abuova); investigation, A.K., A.B. and Z.A. (Zhanar Azhibekova); resources, A.K. and A.B.; data curation, A.M. and M.Y.; writing—original draft preparation, J.T., A.K. and M.Y.; writing—review and editing, Z.A. (Zhanar Azhibekova), Z.A. (Zhanargul Abuova) and A.B.; visualization, A.M. and M.Y.; supervision, A.K. and Z.A. (Zhanar Azhibekova); project administration, A.K. and Z.A. (Zhanar Azhibekova). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19677451).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are publicly available at https://github.com/JamalbekTussupov01/Text-clustering/tree/main (accessed on 27 March 2025).

Conflicts of Interest: Author Mrs. Zhanar Azhibekova is employed by the company "Non-profit Joint Stock Company S. Asfendiyarov Kazakh National Medical University".The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AE	Autoencoder
ARI	Adjusted Rand Index
BERT	Bidirectional Encoder Representations from Transformers
CH	Calinski–Harabasz Index
DB	Davies–Bouldin Index
FM	Fowlkes–Mallows Index
KNN	K-Nearest Neighbors
Lasso	Least Absolute Shrinkage and Selection Operator
LDA	Latent Dirichlet Allocation
LSTM	Long Short-Term Memory
MAPE	Mean Absolute Percentage Error
MCC	Matthews Correlation Coefficient
NLP	Natural Language Processing
RMS	Root Mean Square
RMSE	Root Mean Square Error
SGs	Smart Grids
STC	Short Text Clustering
TF-IDF	Term Frequency–Inverse Document Frequency
TRTD	Topic Representative Term Discovery
UMAP	Uniform Manifold Approximation and Projection
V-Measure	Harmonic Mean of Homogeneity and Completeness
XGBoost	Extreme Gradient Boosting

References

1. Manias, G.; Mavrogiorgou, A.; Kiourtis, A.; Symvoulidis, C.; Kyriazis, D. Multilingual Text Categorization and Sentiment Analysis: A Comparative Analysis of Multilingual Approaches for Classifying Twitter Data. *Neural Comput. Appl.* **2023**, *35*, 21415–21431. [CrossRef] [PubMed]
2. Fu, G.; Li, B.; Yang, Y.; Li, C. Re-Ranking and TOPSIS-Based Ensemble Feature Selection with Multi-Stage Aggregation for Text Categorization. *Pattern Recognit. Lett.* **2023**, *168*, 47–56. [CrossRef]
3. Edara, D.C.; Vanukuri, L.P.; Sistla, V.; Kolli, V.K.K. Sentiment Analysis and Text Categorization of Cancer Medical Records with LSTM. *J. Ambient Intell. Humaniz. Comput.* **2023**, *14*, 5309–5325. [CrossRef]
4. Balaji, T.K.; Annavarapu, C.S.R.; Bablani, A. Machine Learning Algorithms for Social Media Analysis: A Survey. *Comput. Sci. Rev.* **2021**, *40*, 100395.
5. Abbas, A.F.; Jusoh, A.; Mas'od, A.; Alsharif, A.H.; Ali, J. Bibliometrix Analysis of Information Sharing in Social Media. *Cogent Bus. Manag.* **2022**, *9*, 2016556. [CrossRef]
6. McKitrick, M.K.; Schuurman, N.; Crooks, V.A. Collecting, Analyzing, and Visualizing Location-Based Social Media Data: Review of Methods in GIS-Social Media Analysis. *GeoJournal* **2023**, *88*, 1035–1057. [CrossRef]
7. Becken, S.; Friedl, H.; Stantic, B.; Connolly, R.M.; Chen, J. Climate Crisis and Flying: Social Media Analysis Traces the Rise of "Flightshame". *J. Sustain. Tour.* **2021**, *29*, 1450–1469. [CrossRef]
8. Chakraborty, K.; Bhattacharyya, S.; Bag, R. A Survey of Sentiment Analysis from Social Media Data. *IEEE Trans. Comput. Soc. Syst.* **2020**, *7*, 450–464. [CrossRef]

9. Horta Ribeiro, M.; Cheng, J.; West, R. Automated Content Moderation Increases Adherence to Community Guidelines. In Proceedings of the ACM Web Conference 2023, New York, NY, USA, 30 April–4 May 2023; pp. 2666–2676.
10. He, Q.; Hong, Y.; Raghu, T.S. The Effects of Machine-Powered Platform Governance: An Empirical Study of Content Moderation. *SSRN Electron. J.* **2021**. Available online: http://hdl.handle.net/10125/80064 (accessed on 5 May 2025). [CrossRef]
11. Fasel, M.; Weerts, S. Can Facebook's Community Standards Keep Up with Legal Certainty? Content Moderation Governance under the Pressure of the Digital Services Act. *Policy Internet* **2024**, *16*, 588–606. [CrossRef]
12. Saranya, S.; Usha, G. A Machine Learning-Based Technique with Intelligent WordNet Lemmatize for Twitter Sentiment Analysis. *Intell. Autom. Soft Comput.* **2023**, *36*, 339–352. [CrossRef]
13. Hupkes, D.; Giulianelli, M.; Dankers, V.; Artetxe, M.; Elazar, Y.; Pimentel, T.; Jin, Z. A Taxonomy and Review of Generalization Research in NLP. *Nat. Mach. Intell.* **2023**, *5*, 1161–1174. [CrossRef]
14. Chung, S.; Moon, S.; Kim, J.; Kim, J.; Lim, S.; Chi, S. Comparing Natural Language Processing (NLP) Applications in Construction and Computer Science Using Preferred Reporting Items for Systematic Reviews (PRISMA). *Autom. Constr.* **2023**, *154*, 105020. [CrossRef]
15. Xin, Q.; He, Y.; Pan, Y.; Wang, Y.; Du, S. The Implementation of an AI-Driven Advertising Push System Based on a NLP Algorithm. *Int. J. Comput. Sci. Inf. Technol.* **2023**, *1*, 30–37. [CrossRef]
16. Işıkdemir, Y.E. NLP Transformers: Analysis of LLMs and Traditional Approaches for Enhanced Text Summarization. *Eskişehir Osman. Univ. J. Eng. Archit. Fac.* **2024**, *32*, 1140–1151. [CrossRef]
17. Nelson, L.K.; Burk, D.; Knudsen, M.; McCall, L. The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods. *Sociol. Methods Res.* **2021**, *50*, 202–237. [CrossRef]
18. Zhang, X.; Ju, T.; Liang, H.; Fu, Y.; Zhang, Q. LLMs Instruct LLMs: An Extraction and Editing Method. *arXiv* **2024**, arXiv:2403.15736.
19. Zhou, C.; Li, Q.; Li, C.; Yu, J.; Liu, Y.; Wang, G.; Sun, L. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. *arXiv* **2023**, arXiv:2302.09419. [CrossRef]
20. Lamsiyah, S.; Mahdaouy, A.E.; Ouatik, S.E.A.; Espinasse, B. Unsupervised Extractive Multi-Document Summarization Method Based on Transfer Learning from BERT Multi-Task Fine-Tuning. *J. Inf. Sci.* **2023**, *49*, 164–182. [CrossRef]
21. Yu, D.; Xiang, B. Discovering Topics and Trends in the Field of Artificial Intelligence: Using LDA Topic Modeling. *Expert Syst. Appl.* **2023**, *225*, 120114. [CrossRef]
22. Lohith, C.; Chandramouli, H.; Balasingam, U.; Arun Kumar, S. Aspect Oriented Sentiment Analysis on Customer Reviews on Restaurant Using the LDA and BERT Method. *SN Comput. Sci.* **2023**, *4*, 399. [CrossRef]
23. Li, P.; Pei, Y.; Li, J. A Comprehensive Survey on Design and Application of Autoencoder in Deep Learning. *Appl. Soft Comput.* **2023**, *138*, 110176. [CrossRef]
24. Chen, S.; Guo, W. Auto-Encoders in Deep Learning—A Review with New Perspectives. *Mathematics* **2023**, *11*, 1777. [CrossRef]
25. Bengesi, S.; El-Sayed, H.; Sarker, M.K.; Houkpati, Y.; Irungu, J.; Oladunni, T. Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. *IEEE Access* **2024**, *12*, 69812–69837. [CrossRef]
26. Ahmed, M.H.; Tiun, S.; Omar, N.; Sani, N.S. Short Text Clustering Algorithms, Application and Challenges: A Survey. *Appl. Sci.* **2023**, *13*, 342. [CrossRef]
27. Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; Abuhaija, B.; Heming, J. K-Means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data. *Inf. Sci.* **2023**, *622*, 178–210. [CrossRef]
28. Kyaw, K.S.; Tepsongkroh, P.; Thongkamkaew, C.; Sasha, F. Business Intelligent Framework Using Sentiment Analysis for Smart Digital Marketing in the E-Commerce Era. *Asia Soc. Issues* **2023**, *16*, e252965. [CrossRef]
29. Murshed, B.A.H.; Mallappa, S.; Abawajy, J.; Saif, M.A.N.; Al-Ariki, H.D.E.; Abdulwahab, H.M. Short Text Topic Modelling Approaches in the Context of Big Data: Taxonomy, Survey, and Analysis. *Artif. Intell. Rev.* **2023**, *56*, 5133–5260. [CrossRef]
30. Habbak, H.; Mahmoud, M.; Metwally, K.; Fouda, M.M.; Ibrahem, M.I. Load Forecasting Techniques and Their Applications in Smart Grids. *Energies* **2023**, *16*, 1480. [CrossRef]
31. Yang, S.; Huang, G.; Cai, B. Discovering Topic Representative Terms for Short Text Clustering. *IEEE Access* **2019**, *7*, 92037–92047. [CrossRef]
32. Tussupov, J. Text Clustering with BERT and LDA. 2023. Available online: https://github.com/JamalbekTussupov01/Text-clustering/tree/main (accessed on 5 May 2025).