applied sciences

MDPI

Article

A Neural Topic Modeling Study Integrating SBERT and Data Augmentation

Huaqing Cheng [1,2], Shengquan Liu [1,2,*], Weiwei Sun [1,2] and Qi Sun [1,2]

[1] College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China; 107552001251@stu.xju.edu.cn (H.C.); sunqi@stu.xju.edu.cn (Q.S.)
[2] Xinjiang Multilingual Information Technology Laboratory, College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China
* Correspondence: liu@xju.edu.cn

Abstract: Topic models can extract consistent themes from large corpora for research purposes. In recent years, the combination of pretrained language models and neural topic models has gained attention among scholars. However, this approach has some drawbacks: in short texts, the quality of the topics obtained by the models is low and incoherent, which is caused by the reduced word frequency (insufficient word co-occurrence) in short texts compared to long texts. To address these issues, we propose a neural topic model based on SBERT and data augmentation. First, our proposed easy data augmentation (EDA) method with keyword combination helps overcome the sparsity problem in short texts. Then, the attention mechanism is used to focus on keywords related to the topic and reduce the impact of noise words. Next, the SBERT model is trained on a large and diverse dataset, which can generate high-quality semantic information vectors for short texts. Finally, we perform feature fusion on the augmented data that have been weighted by an attention mechanism with the high-quality semantic information obtained. Then, the fused features are input into a neural topic model to obtain high-quality topics. The experimental results on an English public dataset show that our model generates high-quality topics, with the average scores improving by 2.5% for topic coherence and 1.2% for topic diversity compared to the baseline model.

Keywords: SBERT; topic models; data augmentation; feature fusion

check for updates

Citation: Cheng, H.; Liu, S.; Sun, W.; Sun, Q. A Neural Topic Modeling Study Integrating SBERT and Data Augmentation. *Appl. Sci.* **2023**, *13*, 4595. https://doi.org/10.3390/app13074595

Academic Editor: Luis Javier García Villalba

Received: 10 March 2023
Revised: 1 April 2023
Accepted: 2 April 2023
Published: 5 April 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

A topic model [1] is an unsupervised model that statistically discovers potential topics from large corpora for application to downstream tasks in natural language processing, including text classification [2], sentiment analysis [3], etc. There are also Bayesian approaches represented by latent semantic analysis (LSA) [4], probabilistic latent semantic analysis (PLSA) [5], and hierarchical Dirichlet process (HDP) [6]. The textual content of the topic model is usually represented by a bag-of-words representation and the generation of the bag-of-words data is modeled using an underlying probabilistic structure. This approach allows the calculation of topic distributions and topic-word distributions for each article. Although traditional Bayesian topic models have been successful, these models often require complex computations and have a poor portability.

The neural topic model (NTM), which includes a neural component (the variational autoencoder [7]), produces high-quality topics and superior performance compared to the traditional latent Dirichlet allocation model [8–10]. The NTM utilizes a multilayer perceptron (MLP) to process the bag-of-words input and then performs a variational inference via a neural network to sample the potential document topic vector. The decoder network then utilizes an MLP to reconstruct the topic-word representation of the original document. The NTM outperforms traditional topic models in terms of topic quality and scalability and is well-suited for mining topics from massive amounts of text without considering the complexity of model computation and scalability. However, these models still have drawbacks,

such as using the bag-of-words model representation, which neglects the semantic relevance of the context between words. As the robust representation is determined by the semantics between contexts, neglecting this information can lead to poor results. The development of word embedding and pretrained language models has facilitated the advancement of topic modeling techniques. Specifically, static word embedding techniques such as word2vec [11] and Glove [12] have an edge over the bag-of-words model since word embedding captures syntactic and semantic rules by encoding the local context of word co-occurrence patterns. This overcomes the limitations of the bag-of-words model, which cannot reflect the differences and similarities between words and loses sequential information. For instance, the ETM [13] method is a document generation model that combines the traditional topic model LDA with word embeddings. Specifically, it models each word using a categorical distribution, where the natural parameter of the categorical distribution is the dot product between the word embedding and its given topic embedding. The ETM has a good predictive performance, but it cannot handle issues such as polysemy of words. Moreover, the development of pretrained language model technology has brought great advancements to NLP tasks, such as BERT [14], RoBERTa [15], DeBERTa [16], etc. ZeroShot^{TM} [17] replaces the bag-of-words input representation in topic models with pretrained contextualized word embeddings, allowing it to capture contextually relevant semantic information. However, this approach has some limitations. Firstly, due to the sparsity of features and the lack of word co-occurrence patterns, the model often fails to extract high-quality topics from short texts. Secondly, the model's performance in analyzing semantic information from short texts deteriorates, and it typically fails to extract feature information relevant to the topic in the text, resulting in low-quality topics that are often repetitive and lack coherence.

To address the above issues, we propose a neural topic model that integrates SBERT [18,19] and data augmentation. Specifically, we introduce a new data augmentation technique that incorporates keyword information. The data augmentation technique [20] uses simple random replacements, insertions, deletions, and other operations to enhance the robustness of text data. The keyword information is obtained through the TextRank algorithm [21], which efficiently and quickly extracts important words from a large amount of text or other materials. Therefore, our method combining EDA with keyword data augmentation is beneficial for overcoming the sparsity problem of text features. Next, we vectorize the augmented text data and input them into a BiLSTM-Att module to obtain the long-distance dependency information and overcome the influence of noisy words. Then, although the emergence of BERT has propelled the development of various NLP tasks, they still struggle to produce semantically meaningful embeddings for shorter language units such as sentences and phrases. The most common method for BERT sentence embedding is to take the average of the BERT output layer (referred to as BERT embedding) or use the output of the first token ([CLS] token). This common practice results in relatively poor sentence embeddings that are unsuitable for unsupervised clustering and other tasks [18]. Therefore, we use SBERT instead of BERT, which takes the entire sentence as the processing unit and uses Siamese networks and triplet networks to update weights, resulting in semantically meaningful sentence embeddings that are more suitable for tasks such as semantic similarity search and clustering. Furthermore, the SBERT model in this paper is trained on a large and diverse dataset and designed as a general-purpose model, achieving superior results compared to the sentence encoder in the baseline model ZeroShot^{TM} across various tasks. Moreover, our SBERT is capable of generating high-quality semantic embeddings for short texts. Finally, we merge the information that has been enhanced through data augmentation and processed through the attention mechanism with the high-quality semantic feature information. Then, we feed the resulting feature information into a neural topic model to learn high-quality topics. We use the ProdLDA model [22] as our neural topic model, which is a topic model based on automatic-encoding variational inference. It optimizes the structure of the topic model to be more suitable for tasks such as topic mining. Experimental results show that our model produces coherent and high-quality topics on the English public dataset; specifically, our contributions are as follows:

- We use an improved neural topic model combining SBERT and VAE for topic discovery tasks.
- We propose a new data augmentation method of fused keywords to overcome the sparsity problem of text and thus learn more accurate topics.
- We use the BiLSTM-Att mechanism to obtain context dependencies as well as to reduce the influence of noisy words (words irrelevant to the topic) on the model.
- Experimental results show that our approach leads to improved topic coherence and obtains diverse and interpretable topics.

2. Related Work

In recent years, the rise of neural networks and pretrained models has promoted the development of topic modeling. The emergence of neural topic models has helped people discover valuable and important topic information from massive quantity of noisy network data and promoted the development of downstream tasks in natural language processing such as text classification.

Miao et al. [23] introduced the concept of the neural variational document model (NVDM) based on the variational autoencoder. This early attempt to combine topic modeling and autoencoder utilized a multivariate Gaussian distribution instead of the original model's prior distribution. Later, Miao et al. were inspired by the NVDM and introduced the Gaussian model [24]. This model constructed topic distributions by providing parameterizable distributional transformations of topics within a variational inference framework. The ProdLDA topic model [22] is a variational-autoencoder-based model that uses a logistic Gaussian distribution to approximate the distribution and replaces the mixture model in the topic model with an expert's product. Card et al. [25] proposed a stochastic variational inference neural framework that enabled flexible combinations of metadata with various options and allowed the rapid exploration of alternative models. Nan et al. [26] proposed a model called W-LDA in the autoencoder framework. Unlike existing neural-network-based models, this model could directly execute the Dirichlet prior and outperformed other neural topic models in matching high-dimensional distributions. Additionally, the noise added by that model to the encoder module significantly improved topic consistency without compromising diversity. Lastly, the use of the maximum mean difference instead of ELBO in neural topics allowed the model to produce higher quality topics. Wang et al. [27] proposed a neural topic modeling approach called a two-way adversarial topic model. That approach employed a generator to learn the distribution of document topics to the distribution of document-words, and a discriminator was used to distinguish between true distribution pairs and pseudodistribution pairs. That technique helped the network (generator and encoder) to better learn article topics and topic-word distributions. Wu et al. [28] proposed the negative binomial neural topic model, which attempted to combine variational inference with a mixed counting model for a neural variational topic model. That model used a reparameterization of the gamma distribution and a Gaussian approximation of the Poisson distribution. It also developed a neural variational inference algorithm to infer the model parameters, while implicitly introducing non-negative constraints in the topic modeling to improve topic quality. Tian et al. [29] introduced a model called RRT-VAE, which utilized a rounded reparameterization technique to reparameterize the neural topic model of the Dirichlet distribution. That technique was an effective reparameterization method and extended the applicability of the model to other applications beyond the Dirichlet distribution. Gupta et al. [30] proposed iDocNADE, a neural autoregressive topic model designed to enhance the context of short texts. Experiments demonstrated that iDocNADE outperformed other topic models.

Topic models and their variants have been widely applied by scholars in various fields. The LB-MMT model [31] was a new probabilistic topic model used to explore latent topics in multimodal data from social media. It integrated labels, text, and visual information into a unified framework, addressing three key challenges: the top-down relationship from labels to text and images, the one-to-many relationship between labels and topics, and

representing visual and textual information in the model. Experimental results showed that LB-MMT outperformed all baseline models in a quantitative evaluation and had significant practical implications. Mishra et al. [32] discussed the development of a digital tourism recommendation system, which utilized online comments, blogs, and rating data for analysis. A machine-learning-based system was established to achieve three subgoals: predicting star ratings from comments, a feedback model, and a knowledge-based recommendation system. The system used both random forest classifiers and decision tree classifiers to predict star ratings and employed clustering and topic modeling to identify themes from the comments. Due to the majority of work being focused on identifying fake news, it has become difficult to apply these models for news classification in the real world. Therefore, Dauad et al. [33] utilized machine learning techniques for classifying online news articles, proposing a method based on a hyperparameter optimization of support vector machines and compared it with five other maximum likelihood classification techniques. The results demonstrated that the optimized support vector machine model exhibited better performance than the other models. Finally, the authors used real-world datasets to classify news articles. Past academic work on improving the YouTube learning environment has mainly collected learners' impressions through interviews and questionnaires. To overcome the above disadvantages, Alawadh et al. [34] proposed using the YouTube API to collect comments from three randomly selected popular YouTube channels and analyzed the comments using TextBlob and automated latent semantic analysis (LSA) methods to understand global and open discussion topics. Additionally, the authors proposed hypotheses and conclusions about YouTube EFL learning, as well as advice for English learners who hope to use both online and offline learning. Finally, that article highlighted the importance of reviews in order to provide practical information for learning English based on YouTube. The digital novel application Wattpad is one of the most popular apps on the Google Play Store. This research [35] aimed to conduct topic modeling by analyzing user reviews of the Wattpad app, with data obtained from the Google Play store using web scraping techniques. Unsupervised learning methods such as latent Dirichlet allocation (LDA) were used for topic modeling to effectively identify different themes relevant to Wattpad reviews. That study utilized the LDA method to generate words representing the major themes to provide a better understanding of the latest information about the Wattpad app. Liu et al. [36] primarily investigated the knowledge structure and evolutionary trends in the field of artificial intelligence. They employed a new latent feature topic model, called New-LDA, for topic recognition research and analyzed the knowledge structure of the field from the perspectives of topic recognition and coword analysis. The authors also introduced a time series model to establish a topic evolution network and conducted a comparative analysis of high-frequency words in three periods to identify the evolutionary patterns of the knowledge structure in AI.

In addition, several models based on word embedding and pretrained language models have emerged. Bianchi et al. [17] proposed a new topic model, which was a cross-lingual model that utilized the pretraining method. The model acquired knowledge from one language and applied it to predict topics for unseen documents written in different languages. Hoyle et al. [37] introduced a teacher–student model that utilized knowledge distillation for topic discovery tasks. That model was based on a variational autoencoder and could effectively combine various neural topic models, including the BAT + Scholar model. The BAT + Scholar model could seamlessly integrate the pretrained language model BERT with other neural topic models.

Topic models have been used with networks such as LSTM [38] networks in order to discover more coherent topics. Ding et al. [39] used an RNN to detect discontinued words and merged its output with document topic vectors to cluster the same semantic content. Jin et al. [40] proposed a matrix-decomposition model, LTMF, which integrated long short-term networks and topic modeling for comment understanding. That model showed a better topic clustering capability compared to traditional topic-model-based approaches.

Data augmentation is widely employed in text classification [41] and text clustering tasks [42], especially in situations where resources are limited. It has the potential to increase the size and enhance the quality of training data, as well as the robustness of deep neural networks. While data augmentation techniques have mainly been utilized in the image domain, their implementation in the text domain may have unforeseen consequences. This article explores various data enhancement techniques. Wei et al. [20] applied random swapping, random insertion, random deletion, and synonym replacement to improve text data, but that approach may eliminate useful information from the original text and adversely impact results. To overcome these limitations, Karimi et al. [43] suggested a straightforward data augmentation method that avoided the removal of information from the source dataset. That method involved the insertion of punctuation marks to alter their position within a sentence, thereby boosting the model's generalization ability and outperforming the baseline model.

In recent years, scholars have explored ways to improve the models based on a combination of word embedding and topic models. Our approach explores the fusion of data augmentation and the sentence encoder SBERT and neural topic models on this basis and obtains high-quality topics.

3. Materials and Methods

We introduce a neural topic model based on SBERT and data augmentation for a better discovery of high-quality topics. Our model is built around three main components: (1) SBERT, which is used to extract rich contextual features from text; (2) data augmentation and fine-grained feature information such as keywords, which are incorporated with attention mechanisms to reduce the influence of noisy words; and (3) the neural topic model ProdLDA. The overall framework of our model is shown in Figure 1.

3.1. SBERT

SBERT employs Siamese and triplet networks (one positive example + two negative examples, or one negative example + two positive examples) to update weights, so that semantically similar sentences have the smallest possible distance in vector space, while dissimilar ones have the largest possible distance. Additionally, SBERT uses similarity measures such as cosine similarity or the Manhattan/Euclidean distance to calculate sentence similarity. Compared to other pretrained language models (such as BERT), SBERT can quickly generate sentence embeddings and is better suited for unsupervised tasks such as semantic similarity search and clustering. In this article, we used the "all-mpnet-base-v2" module of SBERT instead of the module in the baseline model [17]. The model was trained on all available training data (more than 1 billion training pairs) and was designed as a general purpose model. It was intended to be used as a sentence encoder. Given an input text, it output a vector which captured the semantic information, and the model provided a better quality for both short and long texts. Then, we introduced the triplet network. Given an anchor point sentence a, the positive sentence p, and negative sentence n, the triplet loss adjusted the network so that the the distance between a and n was greater than the distance between a and p. Mathematically, the following loss function needed to be optimized, as shown in Equation (1).

$$\max(||s_a - s_p|| - ||s_a - s_n|| + \epsilon, 0) \tag{1}$$

where $||*||$ is a distance matrix, margin ϵ ensures that s_p is at least ϵ closer to s_a than s_n, and s_a, s_p, and s_n represent their sentence embeddings, respectively.

The classification objective function in SBERT is shown in Equation (2) where the sentence embeddings u and v are used in the elemental rank difference $|u - v|$, which is multiplied with the trainable weights w_t.

$$O = \text{softmax}(W_t(u, v, |u - v|)) \tag{2}$$

The model optimized the cross-entropy loss in SBERT to make it more convenient for natural language processing tasks, such as text classification, clustering, etc. The structure is shown in Figure 2.

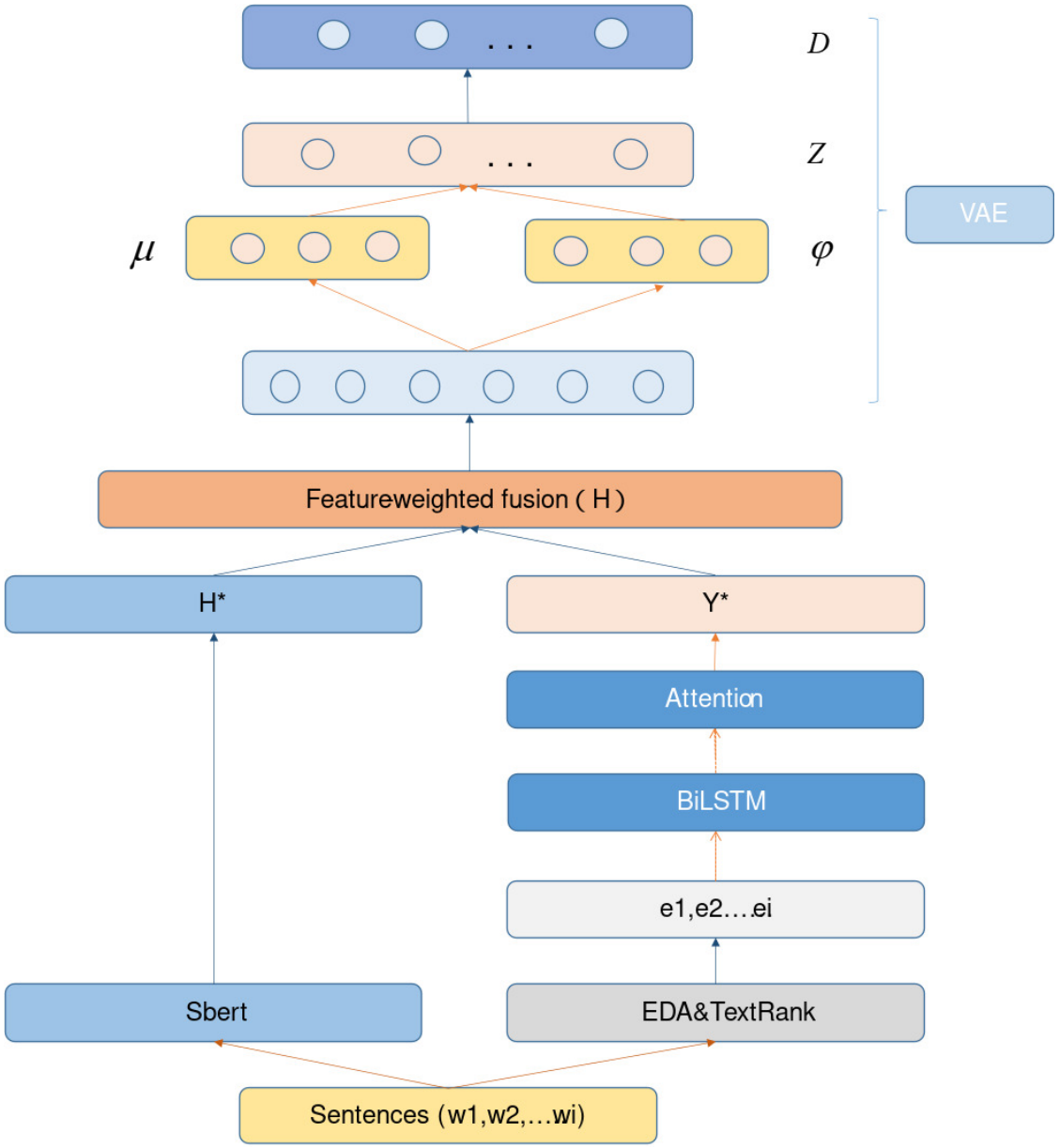


Figure 1. In the overall framework diagram of this paper, first, we use SBERT with a triplet network to obtain a rich representation of contextual sentence embeddings; then, we add input branches to the encoder to obtain the data-enhanced features. Specifically, we employ the EDA method combined with TextRank-algorithm-derived keywords to overcome the sparsity issue in the text. The text data after data augmentation are processed by static word embedding (word2vec), and then we filtered out the influence of noisy words by the BiLSTM-Att module. The processed data-augmented vector Y* is combined with the previously SBERT-processed sentence vector H* by a weighted feature fusion (where H = α H* + (1 − α)Y*, $H \in R^{B \times D}$, where B denotes the number of input samples, D denotes the latitude of the embedding vector, and α is the hyperparameter with a default value of 0.7, wi represents the ith word, and ei represents the embedding vector of the ith word). Finally, the resulting vector representation is used as the input to the encoder, and high-quality topics are extracted through the variational autoencoder (where μ is the mean, ϕ is the standard deviation, Z is the latent vector, and D is the reconstructed data obtained by the decoder).

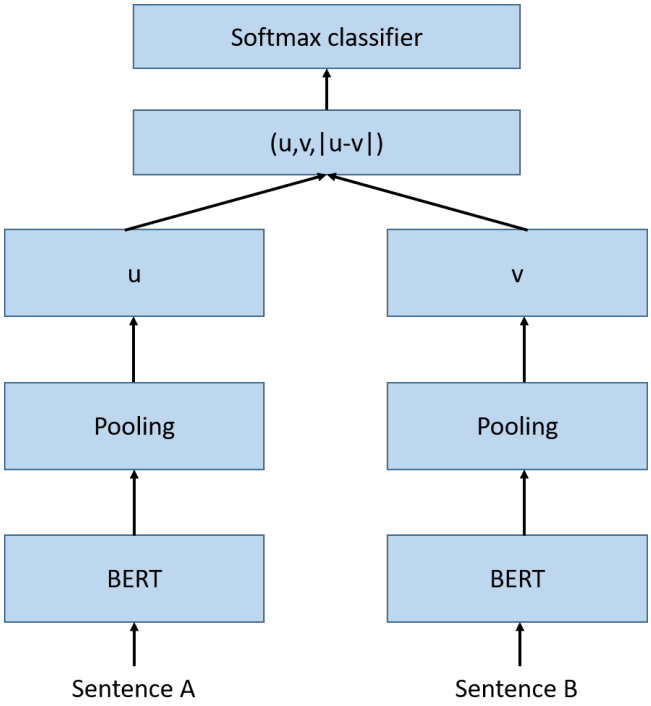


Figure 2. SBERT architecture with classification goals, for fine-tuning on the SNLI open dataset. The weights of the two BERT networks are the same.

The model also used a regression objective function to calculate the sentence similarity more accurately and conveniently, and the cosine-similarity between the two sentence embeddings u and v was computed (Figure 3).

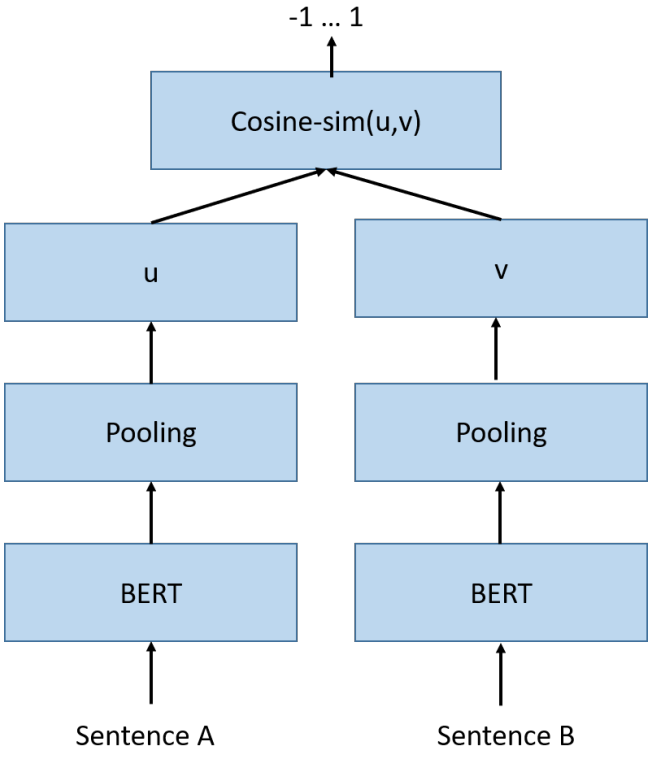


Figure 3. The architecture used by SBERT to calculate similarity scores; this architecture can also be used with the objective function of the regression.

3.2. Data Enhancement and Other Modules

Data augmentation techniques can enrich semantic information and resolve issues such as text data sparsity. One powerful yet simple method is to use random insertion, deletion, and substitution to enhance text quality. This approach enhances the training data quality and neural network robustness. However, its effectiveness in improving topic modeling tasks is limited. Thus, we incorporated the keyword feature information on top of this method to help the topic model extract high-quality topics. Specifically, we utilized the TextRank algorithm, a graph-based ranking algorithm for text. This algorithm partitions text into constituent units, constructs a node–link graph, uses word similarity as edge weights, iteratively calculates sentence (word) values, and extracts higher ranked phrases. TextRank is an improvement of the PageRank algorithm and is represented in Equation (3):

$$WS(V_i) = (1 - \mathrm{d}) + \mathrm{d} * \sum_{V_j \in In(V_i)} \frac{W_{Ji}}{\sum_{V_k \in Out(V_j)} W_{Jk}} WS(V_j). \tag{3}$$

where V_i denotes the ith vertex, W_{Ji} denotes the weights of the corresponding edges of the two vertices, $WS(V_i)$ denotes the coefficient (importance) of the ith vertex, and d is the damping coefficient, which generally has a default value of 0.85.

An LSTM network is an extended version of an RNN that effectively addresses the problem of gradient disappearance in RNNs. However, it can only acquire information in one direction. In contrast, a BiLSTM layer comprises two LSTM layers in opposite directions, enabling it to obtain semantic encoding that includes contextual information. Consequently, a BiLSTM layer can effectively represent the semantic features of text. The extracted information is then incorporated into the attention mechanism, which calculates the attention of words in the text. The weight assigned to each word corresponds to its level of importance in the task. Thus, the model can concentrate on groups of cue words that are relevant to the topic and effectively cluster the text together.

3.3. Neural Topic Model

ProdLDA is a neural topic model that uses a variational autoencoder. It replaces the mixture model in LDA with expert products and uses units such as the Adam optimizer and batch normalization in the encoder network to avoid component collapse issues. The encoder module transforms the feature representation of each document into a latent representation of the corresponding document topic distribution through a multilayer perceptron, while the decoder attempts to sample from these latent representations and map them back to the original document data. The marginal likelihood function of ProdLDA is shown in Equation (4):

$$p(w|\alpha, \beta) = \int_{\theta} \left(\prod_{n=1}^{N} p(w_n|\beta, \theta) \right) p(\theta|a) d\theta \tag{4}$$

where β is the probability distribution over the vocabulary corresponding to each topic, as well as the hidden variables θ and the reconstructed document w. Its improvement on the topic model construction allows it to generate higher-quality topics.

4. Experiments

Our task was similar to an unsupervised clustering task, in which we aimed to mine a large amount of text for potential topics for a quick understanding of the text's thematic content, and we used a neural topic model that incorporated data-enhanced knowledge and a sentence encoder to discover high-quality (relatively coherent and diverse) topics on English data.

4.1. DataSet

We used the M10 dataset in English [44] as our experimental dataset, which is a short-text dataset. The M10 dataset consists of 8355 documents with a vocabulary size of 1696.

We preprocessed the data by converting the text to lowercase, removing punctuation and eliminating stop words and other operations.

4.2. Settings

The model settings in this paper included the following parameters: learning _rate, dropout, batch_size, optimizer, embedding_dim, etc. For more visualization, the settings are shown in Table 1.

Table 1. Experimental parameter settings in this paper.

Name	Parameters
learning_rate	1×10^{-3}
Dropout	0.1
batch_size	64
Optimizer	Adam
embedding_dim	768

4.3. Baselines

We used the results of the following models run on the M10 dataset as a baseline.

LDA [1]: this method computes a topic distribution for each article (called the article-topic distribution) and approximates each topic to a probability distribution over the vocabulary (called the topic-word distribution); it is a classic method of topic modeling.

ProdLDA [22]: this neural topic model uses a logistic Gaussian distribution to approximate the Dirichlet distribution, optimizes its components, and uses the Adam optimizer, a batch normalization, and other units in the encoder network to avoid AEVB component crashes occurring; ProdLDA is a widely used classical neural topic model.

SBERT + LDA: an approach combining the sentence encoder SBERT and traditional topic models; we used the SBERT + LDA module from the method of Bianchi et al. [17], where the powerful contextualized embedding of pretrained language models helped the topic models to better mine topics.

ETM [13]: This model is a document-generation model that combines the traditional topic model LDA with word embeddings. The ETM is capable of identifying meaningful topics despite having a vast vocabulary that includes infrequent words and commonly used stop words.

ZeroShot^{TM} [17]: This model uses pretrained contextualized embeddings instead of the bag-of-words input representation in the topic model, which allows it to capture contextually relevant semantic information, and it is a fusion of a sentence encoder and a neural topic model (SBERT + ProdLDA). The method achieves excellent results on topic mining tasks.

4.4. Evaluation Indicator

Topic coherence: Topic coherence refers to the relatedness of words in a topic, i.e., whether the words in a topic share the same thematic characteristics. The purpose of topic coherence evaluation metrics is to assess the relevance and consistency of the topics extracted from a topic model. Topic coherence evaluation metrics commonly use measures such as NPMI and C_V to evaluate topic coherence. The evaluation metrics aim to assess whether the topic model can generate semantically coherent and easily understandable topics. For example, the group of words “apple, banana, orange, grape” is more coherent and aligned with human judgment than “apple, banana, packaging paper, price”.

Evaluating topic coherence is usually calculated using the normalized point mutual information (NPMI) [45,46]. The score for each topic is based on the word combinations of the first T words returned for that topic to calculate the co-occurrence, and then the topic coherence score is calculated based on the co-occurrence of the words. By default, the first 10 words of each topic are used to calculate the value of the NPMI. It is defined in Equation (5):

$$NPMI(t) = \sum_{j=2}^{N} \sum_{i=1}^{j-1} \frac{\log \frac{p(w_i,w_j)}{p(w_i)p(w_j)}}{-\log P(w_i,w_j)} \quad (5)$$

where $P(w_i)$ is the probability that the word w_i appears in a document, and $P(w_i, w_j)$ is the number of times the pair (w_i, w_j) appears in a document.

In addition, the C_V metric is another method used to evaluate the consistency of topic models, similar to the NPMI approach. C_V can measure both the internal coherence of a topic and the coherence between topics. Specifically, C_V utilizes the co-occurrence information of words to calculate the relevance within a topic and the separation between topics, thus assessing the quality of topics. We followed the approach used by Terragni et al. [44] to conveniently compute the coherence metrics (NPMI,C_V) for topics.

Topic diversity: The RBO [47] measure is based on a simple probabilistic user model. It compares the first ten words of two topics. We judged the diversity of topics by the reciprocal of the RBO measure [44,48]. For the same topic in a theme, it has a countdown of 0, while for completely different topics, it has a countdown of 1. The better the topic diversity, the higher the score, indicating a better model extraction.

We show some examples of topics mined by the model to better visually represent the topics proposed by the model in this paper and to view the coherence of the topics in a human-compatible way. When the semantics of the words of a set of topics are similar, we can consider them as semantically coherent and high-quality topics. When a set of topics is semantically coherent and the words are not identical, they can be considered as topics with a high diversity.

5. Results and Analysis

First, we evaluated the topic coherence metrics NPMI and C_V values, with the number of topics set to 50. As shown in Table 2, we can see that LDA performed the worst, with low scores in both NPMI and C_V metrics. LDA is a probabilistic topic model based on statistics. ProdLDA ranked second to last in NPMI and C_V values. It is a neural topic model that uses a variational autoencoder. The C_V value of SBERT + LDA was second only to ZeroShot^{TM} , but the NPMI value was lower and in the fourth place. The ETM model had a higher NPMI value than LDA and ProdLDA, but its score was relatively low in the C_V metric. ZeroShot^{TM} ranked second in the model list. Although it scored higher than the previous models in both metrics, it still had its drawbacks, which may be attributed to its lack of consideration for data sparsity issues. Its scores in both metrics were close to our model's scores. Our model performed the best in both metrics, thanks to our approach that overcame the data sparsity issue in text and enhanced the comprehensive information extraction ability. Thus, our model had a better topic quality.

Table 2. Comparison of the topic coherence (NPMI and C_V values) of different models.

Models	NPMI	C_V
Lda	−0.181	0.301
ProdLDA	−0.143	0.361
SBERT + LDA	−0.126	0.423
ETM	−0.083	0.313
ZeroShot^{TM}	0.011	0.441
Ours	**0.036**	**0.467**

Then, we evaluated the diversity metrics of different topic models using the reciprocal of the RBO measure as the score, with the number of topics set to 20. The results are shown in Table 3. From the table, we can see that ProdLDA performed second worst in terms of diversity. Diversity measured the frequency of different words appearing in each topic, and its performance was average. The LDA model had the worst diversity performance due to its high word repetition rate in topics. ZeroShot^{TM} had a similar diversity to that of

our model. Our model exhibited a good diversity in short-text datasets, with the highest diversity metric score.

Table 3. Comparison of topic diversity (RBO cepstral) scores of different models.

Models	K = 20
Lda	0.9061
ProdLDA	0.9635
ZeroShot^{TM}	0.9805
Ours	**0.9926**

Then, when the number of topics was set to 50, we compared the topic diversity scores of different models. The results are shown in Table 4. From the table, it can be seen that LDA performed relatively poorly, as it is a probability-based statistical model that is not as effective as the neural topic model in terms of performance. The topic diversity performance of the ProdLDA model was average, while the ZeroShot^{TM} model was second only to our model. Our model still demonstrated a good topic diversity, with a higher score in topic diversity than other baseline models.

Table 4. Comparison of topic diversity (RBO cepstral) scores of different models.

Models	K = 50
Lda	0.9131
ProdLDA	0.9513
ZeroShot^{TM}	0.9631
Ours	**0.9756**

Then, we performed ablation experiments, as shown in Table 5; we conducted further ablation experiments and demonstrated the variations in NPMI scores of our model. In this experiment, we employed the model ("distiluse-base-multilingual-cased-v2") with a dimension of 512 for SBERT(d), while SBERT(a) utilized the other module ("all-mpnet-base-v2") with a dimension of 768. For more details about the SBERT model, please refer to the paper [18]. As seen from the table, when a larger SBERT model was employed, the score increased by 0.7%. Furthermore, when we integrated the Keyword Easy Data Augmentation (KEDA) module and the BiLSTM-Att module on top of the SBERT model, the NPMI score increased by 1.9%. This improvement could be attributed to the adoption of the new data augmentation method and the use of the attention mechanism to filter out noisy words. The experimental results demonstrated the effectiveness of our proposed method.

Table 5. NPMI values under ablation experiments.

Models	NPMI
Ours (SBERT(a) + KEDA + BiLSTM-Att + ProdLDA)	**0.036**
SBERT(a) + ProdLDA	0.017
SBERT(d) + ProdLDA	0.010

Next, we show some examples of topics extracted by the ZeroShot^{TM} model in Table 6. First of all, we observe that in topic_id = 8, the topic should be about science and technology, while the other two topics are not very coherent and cannot be distinguished from the specific topic to be expressed. There is a need to improve the interpretability or coherence of the topics.

Table 6. Partial example of the model at K = 50 topic number.

Models	Topics	Topic_id
ZeroShot^{TM}	High, soft, metal, good, margin	Topic 5
	network, base, control, system, approach	Topic 8
	model, yield, soil, simulation, estimate	Topic 21

Finally, to make the comparison of topics more intuitive, we provide several examples of topics extracted by our model, as shown in Table 7. It can be observed that when Topic_id = 3 and Topic_id = 12, the instances extracted by the model should reflect topics such as technology and learning, and when Topic_id = 35, the model should extract topics such as the economy. From the table, we can see that the themes generated by our proposed model were more coherent and interpretable. However, there were still themes that did not work well, and this is where we need to improve in the future.

Table 7. Some examples of the model in this paper at K = 50 topics.

Models	Topics	Topic_id
Ours	Mobile, robot, robotic, human, map	Topic 3
	learn, language, recognition, evolution, search	Topic 12
	neural, network, base, approach, model	Topic 21
	market, financial, stock, return, investment	Topic 35

6. Conclusions

In this paper, we proposed a neural topic model that incorporated SBERT and new data enhancement methods, which incorporated data enhancement knowledge such as keywords to help overcome problems such as the sparsity of text, and the attention mechanism was then used to reduce the effect of topic-irrelevant words on the results. We used a new sentence vector encoder to represent meaningful rich sentence information. We merged the information enhanced by the data and processed by the attention mechanism with high-quality semantic feature information (using weighted feature fusion). Finally, the fused feature information was used as the input to the neural topic model encoder, which was processed by the neural topic model to mine high-quality topics, with an average score improvement of 2.5% for topic coherence and 1.2% for topic diversity compared to the baseline model. Thus, our model achieved excellent results in terms of coherence and diversity.

In the future, we plan to explore new pretrained language models, such as Phrase-Bert, which can extract fine-grained feature information, leading to more accurate and richer sentence representations. We will also explore adversarial training techniques to enhance the robustness of the topic model, as well as new neural topic models to improve the quality of topic extraction.

Author Contributions: Writing—original draft, H.C.; writing—review and editing, S.L., W.S. and Q.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Key Projects of Scientific Research Program of Xinjiang Universities Foundation of China under grant xjedu20191004.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Blei, D.M.; Ng, A.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
2. Chaudhary, Y.; Gupta, P.; Saxena, K.; Kulkarni, V.; Runkler, T.; Schütze, H. TopicBERT for energy efficient document classification. *arXiv* **2020**, arXiv:2010.16407.
3. Wu, D.; Yang, R.; Shen, C. Sentiment word co-occurrence and knowledge pair feature extraction based LDA short text clustering algorithm. *J. Intell. Inf. Syst.* **2021**, *56*, 1–23. [CrossRef]
4. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
5. Hofmann, T. Probabilistic latent semantic indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 15–19 August 1999; pp. 50–57.
6. Teh, Y.; Jordan, M.; Beal, M.; Blei, D. Sharing clusters among related groups: Hierarchical Dirichlet processes. *Adv. Neural Inf. Process. Syst.* **2004**, *17*, 1385–1392.
7. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
8. Das, R.; Zaheer, M.; Dyer, C. Gaussian LDA for topic models with word embeddings. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; Volume 1, pp. 795–804.
9. Wei, X.; Croft, W.B. LDA-based document models for ad-hoc retrieval. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, USA, 6–11 August 2006; pp. 178–185.
10. Mehrotra, R.; Sanner, S.; Buntine, W.; Xie, L. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 28 July–1 August 2013; pp. 889–892.
11. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 3111–3119.
12. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
13. Dieng, A.B.; Ruiz, F.J.; Blei, D.M. Topic modeling in embedding spaces. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 439–453. [CrossRef]
14. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
15. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
16. He, P.; Liu, X.; Gao, J.; Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv* **2020**, arXiv:2006.03654.
17. Bianchi, F.; Terragni, S.; Hovy, D.; Nozza, D.; Fersini, E. Cross-lingual contextualized topic models with zero-shot learning. *arXiv* **2020**, arXiv:2004.07737.
18. Reimers, N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv* **2019**, arXiv:1908.10084.
19. Reimers, N.; Gurevych, I. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv* **2020**, arXiv:2004.09813.
20. Wei, J.; Zou, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv* **2019**, arXiv:1901.11196.
21. Mihalcea, R.; Tarau, P. Textrank: Bringing order into text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 July 2004; pp. 404–411.
22. Srivastava, A.; Sutton, C. Autoencoding variational inference for topic models. *arXiv* **2017**, arXiv:1703.01488.
23. Miao, Y.; Yu, L.; Blunsom, P. Neural variational inference for text processing. In Proceedings of the International Conference on Machine Learning, PMLR, New York, NY, USA, 20–22 June 2016; pp. 1727–1736.
24. Miao, Y.; Grefenstette, E.; Blunsom, P. Discovering discrete latent topics with neural variational inference. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 2410–2419.
25. Card, D.; Tan, C.; Smith, N.A. Neural models for documents with metadata. *arXiv* **2017**, arXiv:1705.09296.
26. Nan, F.; Ding, R.; Nallapati, R.; Xiang, B. Topic modeling with wasserstein autoencoders. *arXiv* **2019**, arXiv:1907.12374.
27. Wang, R.; Hu, X.; Zhou, D.; He, Y.; Xiong, Y.; Ye, C.; Xu, H. Neural topic modeling with bidirectional adversarial training. *arXiv* **2020**, arXiv:2004.12331.
28. Wu, J.; Rao, Y.; Zhang, Z.; Xie, H.; Li, Q.; Wang, F.L.; Chen, Z. Neural mixed counting models for dispersed topic discovery. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6159–6169.
29. Tian, R.; Mao, Y.; Zhang, R. Learning VAE-LDA models with rounded reparameterization trick. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 1315–1325.
30. Gupta, P.; Chaudhary, Y.; Buettner, F.; Schütze, H. Document informed neural autoregressive topic models with distributional prior. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6505–6512.
31. Li, H.; Qian, Y.; Jiang, Y.; Liu, Y.; Zhou, F. A novel label-based multimodal topic model for social media analysis. *Decis. Support Syst.* **2023**, *164*, 113863. [CrossRef]

32. Mishra, R.K.; Jothi, J.A.A.; Urolagin, S.; Irani, K. Knowledge based topic retrieval for recommendations and tourism promotions. *Int. J. Inf. Manag. Data Insights* **2023**, *3*, 100145. [CrossRef]
33. Daud, S.; Ullah, M.; Rehman, A.; Saba, T.; Damaševičius, R.; Sattar, A. Topic Classification of Online News Articles Using Optimized Machine Learning Models. *Computers* **2023**, *12*, 16. [CrossRef]
34. Alawadh, H.M.; Alabrah, A.; Meraj, T.; Rauf, H.T. English Language Learning via YouTube: An NLP-Based Analysis of Users' Comments. *Computers* **2023**, *12*, 24. [CrossRef]
35. Awantina, R.; Wibowo, W. Computational Linguistics Using Latent Dirichlet Allocation for Topic Modeling on Wattpad Review. In Proceedings of the 4th International Conference on Science and Technology Applications, ICoSTA 2022, Medan, North Sumatera Province, Indonesia, 1–2 November 2022.
36. Liu, Y.; Chen, M. The Knowledge Structure and Development Trend in Artificial Intelligence Based on Latent Feature Topic Model. *IEEE Trans. Eng. Manag.* 2023; *early access*.
37. Hoyle, A.; Goel, P.; Resnik, P. Improving neural topic models using knowledge distillation. *arXiv* **2020**, arXiv:2010.02377.
38. Graves, A.; Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.
39. Dieng, A.B.; Wang, C.; Gao, J.; Paisley, J. Topicrnn: A recurrent neural network with long-range semantic dependency. *arXiv* **2016**, arXiv:1611.01702.
40. Jin, M.; Luo, X.; Zhu, H.; Zhuo, H.H. Combining deep learning and topic modeling for review understanding in context-aware recommendation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LO, USA, 1–6 June 2018; Volume 1, pp. 1605–1614.
41. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. *Information* **2019**, *10*, 150. [CrossRef]
42. Aggarwal, C.C.; Zhai, C. A survey of text clustering algorithms. In *Mining Text Data*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 77–128.
43. Karimi, A.; Rossi, L.; Prati, A. Aeda: An easier data augmentation technique for text classification. *arXiv* **2021**, arXiv:2108.13230.
44. Terragni, S.; Fersini, E.; Galuzzi, B.G.; Tropeano, P.; Candelieri, A. Octis: comparing and optimizing topic models is simple! In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Online, 19–23 April 2021; pp. 263–270.
45. Lau, J.H.; Newman, D.; Baldwin, T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, 26–30 April 2014; pp. 530–539.
46. Lau, J.H.; Baldwin, T. The sensitivity of topic coherence evaluation to topic cardinality. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 483–487.
47. Webber, W.; Moffat, A.; Zobel, J. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* **2010**, *28*, 20. [CrossRef]
48. Bianchi, F.; Terragni, S.; Hovy, D. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv* **2020**, arXiv:2004.03974.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.