Integrating Text Classification in Topic Discovery with Semantic Embedding Models

Ana Laura Lezama-Sánchez [†] , Mireya Tovar Vidal [*,†] , José A. Reyes-Ortiz [*,†]

Posted Date: 12 May 2023

doi: 10.20944/preprints202305.0908.v1

Keywords: Deep Learning; Topic Discovery; Latent Dirichlet Allocation; Latent Semantic Analysis; Probabilistic Latent Semantic Analysis

Article

Integrating Text Classification in Topic Discovery with Semantic Embedding Models

Ana Laura Lezama-Sánchez [1,†], Mireya Tovar Vidal [1,*,†] and José A. Reyes-Ortiz [2,*,†]

[1] Faculty of Computer Science, Benemerita Universidad Autonoma de Puebla, Puebla 72570, Mexico; yumita1102@gmail.com
[2] Departamento de Sistemas, Universidad Autonoma Metropolitana, Azcapotzalco 02200, Mexico
* Correspondence: mireya.tovarvidal@viep.com.mx (M.T.V.); jaro@azc.uam.mx (J.A.R.-O.)
† These authors contributed equally to this work.

Abstract: Topic discovery is finding the main idea of large amounts of textual data. It indicates the recurring topics in the documents, allowing an overview of the texts. Current topic discovery models receive the texts, with or without pre-processing of Natural Language Processing. The processing consists of stopwords removal, text cleaning and normalization (lowercase conversion). A topic discovery model that receives texts with or without processing generates general topics since the input data is many uncategorized texts. The general topics do not offer a detailed overview of the input texts, and manual text categorization is a time-consuming and tedious task. Accordingly, it is necessary to integrate an automatic text classification task in the topic discovery process to obtain specific topics with their top words that contain relevant relationships based on belonging to a class. Text classification performs a word analysis that makes up a document to decide what class or category is being identified; then, integrating the text classification before a topic discovery process will provide latent topics depicted by top words with a high coherence in each topic based on the previously obtained classes. Therefore, this paper exposes a approach that integrates text classification into topic discovery from large amounts of English textual data, such as *20-Newsgroup* and *Reuters* corpora. The text classification is accomplished with a Convolutional Neural Network(CNN) incorporating three embedding models based on semantic relationships. The topic discovery over categorized texts is realized with Latent Dirichlet Analysis(LDA), Probabilistic Latent Semantic Analysis(PLSA), and Latent Semantic Analysis(LSA) algorithms. An evaluation process was performed based on the normalized topic coherence metric. The *20-Newsgroup* corpus was classified, and twenty topics with ten top words were discovered for each class, obtaining 0.1723 normalized topic coherence when applying LDA, 0.1622 with LSA, and 0.1716 with PLSA. The *Reuters* corpus was also classified, obtaining 0.1441 normalized topic coherence when applying the LDA algorithm to obtain 20 topics for each class.

Keywords: deep learning; topic discovery; latent dirichlet allocation; latent semantic analysis; probabilistic latent semantic analysis

1. Introduction

Text classification involves the processing of data without the intervention of a person. The computer must have access to the knowledge necessary to carry out tasks such as medical diagnoses, analysis of social networks, or search for fake news. Therefore, a system that works with many documents requires algorithms or methods to provide the computer with the necessary knowledge to generate the results expected by a user [1].

Natural Language refers to the mechanism that the human being uses to communicate and transmit an idea, opinion, or feeling [2]. Understanding natural language is a complex task and requires time because millions of connections between neurons and bodily processes are required to learn it. A characteristic of Natural Language is not about sentences, opinions, or feelings with previous training or pre-test. However, a computer needs structure and logic to understand a programming

 doi:10.20944/preprints202305.0908.v1

language. A mathematical formula or a predefined pattern will be necessary to learn the necessary knowledge. If this pattern does not follow, the computer cannot understand any data [2]. For a computer to recognize the data it receives, it must generate an adequate numerical representation.

Natural Language Processing (*NLP*) is an area in constant development that seeks to generate efficient algorithms for a computer to understand the spontaneous language of the human being. Some of the characteristics of natural language are strict rules, which facilitate their computerized analysis, like grammar. On the contrary, the meaning of the texts is a more complex structure for the computer to understand [2]. It has developed techniques and tools to implement systems capable of interpreting and using natural language to perform the desired tasks, such as a news classifier or a spam identifier [2].

Text Classification is a task performed with a neural network model or a traditional [3] classification algorithm. A deep learning model for text classification involves using a GPU (Graphics Processing Unit) to train models. The data sets have the manual review of an expert in the field, but it becomes tedious and requires much time [3]. The technology that supports deep learning and the libraries that allow these techniques to does implemented are rapidly evolving [4].

Deep Learning is a process carried out with Convolutional Neural Networks (CNN). They have been adopted for text classification tasks, generating successful results [5]. A CNN is a multi-layer or hierarchical network and is a high-level feature-based method. CNN does build by stacking multiple layers of features. A feature of a CNN is the presence of a subsampling or pooling layer [6]. It allows optimizing the calculation processes to reduce the data size in learning new data, allowing the recognition of different features [7].

Semantic relationship extraction, named entities, topic discovery, and word embedding are areas of study in Natural Language Processing (NLP). They are tasks responsible for providing a computer with the necessary knowledge to process the information and expose a particular result. For the computer to understand data from the real world, it is necessary to use computational algorithms.

For a computer to understand natural language, it is necessary to create vectors of numbers. The embedding vectors or patterns be subject to operations such as addition, subtraction, and distance measurements. On the other hand, word embeddings have become a widely used research area in recent years. In general, they are used as features in PLN tasks. The literature shows that some word embedding models are based on neural networks or context matrices [7]. The advancement of technology has made it possible to streamline processes, for example:

1. Search for the subject of a document.
2. Search for a specific document.
3. Generate a summary or extract key phrases from a text.

Currently, computational approaches need to model knowledge to generate accurate results without the intervention of a person. The text classification allows ordering large amounts of documents in short periods. On the other hand, topic discovery is finding the main idea from large amounts of textual data; it is presented as a recurring topic. The objective of topic discovery in text documents is to extract the central idea by imitating human capacity without human intervention extracting knowledge from the texts automatically. It indicates the recurring topics in the documents, allowing an overview of the texts. A topic discovery model that receives texts with or without processing generates general topics since the input data is many uncategorized texts. However, discovering topics in previously obtained classes will allow us to learn more specific topics. Hence the top words of the discovered topics will be more related to them [4].

This paper extracts existing classes in text documents. Therefore, integrating the texts of classification in the discovery of topics will provide topics with more specific topics significant relationships since the topics to be extracted will be on each existing class rather than on the complete texts.

This work presents a document classification integrated into discovered topics with semantic embedding models. Two news domain corpora were previously classified with a convolutional neural

 doi:10.20944/preprints202305.0908.v1

network using three semantic embedding models [7]. The topic discovery process is conditioned to the existence of previously obtained classes. The quality assessment of the discovered topics was performed with the normalized topic coherence metric. Therefore, the discovered topics in each class obtained will generate topics that belong to the class from which they were extracted. The main contribution of this work is the integration of text classification in the topic discovery task based on the incorporation of semantic relationships. In addition to a comparison of topic discovery performance with existing approaches in the literature; and a model that discovers the topics in classes obtained with a Convolutional Neural Network(CNN), it is observed that the results obtained are promising in classifying the texts with a CNN since it provides the topic discovery algorithm with texts grouped by class.

The rest of the paper is organized as follows: In Section 2 explores works related to this research. Section 3 shows the proposed approach to incorporate text classification in topic discovery process. The experimental results are presented in Section 4. The conclusions and future work are presented in Section 5.

2. Related Work

This section exposes related works in the same field. Some authors incorporated additional algorithms into their approaches to discover topics, such as text classification through deep learning models. They also apply sentiment analysis and clustering algorithms for the same purpose.

In [8], discuss developing an approach to topic discovery. The authors rely on min Hashing to generate multiple random corpus partitions. In this way, it was possible to find sets of matching words, which were subsequently grouped to produce the existing topics in the analyzed text. The data sets used were *20-Newsgroup, Reuters* and *Wikipedia* in English and Spanish. The evaluation metric used was normalized topic coherence. The authors concluded that their proposed approach could produce coherent topics with an extensive vocabulary demonstrating its robustness against rare words.

A model for topic discovery is outlined in [9]. The authors interpreted each document as word embeddings. They later proposed a two-way model for the discovery of multi-level topic structures. In each layer, it learns a set of topical embeddings. The authors proposed learning topic hierarchies with a bidirectional transport chain, which generated topics in two adjacent layers approaching each other. The authors carried out experiments on learning hierarchical visual topics from images. The evaluation metric used was normalized topic coherence.

In [10], they present an approach that introduces hyperbolic embeddings for representing words and topics. The authors used the tree property of hyperbolic space to extract more interpretable topics. The experiments demonstrated that the proposed approach achieved improved performance compared to existing topical models. The data sets used were *20-Newsgroup* and *WikiText-103*. The evaluation metric used was normalized topic coherence.

[11] exposes two models for discovering scattered topics. The first model is negative binomial neural subjects (NB-NTM), and the second is gamma negative binomial neural subjects (GNB-NTM). The experiments indicated that both models outperformed the existing models in the literature. The experiments were performed on large-scale data sets through the backpropagation algorithm and GPU acceleration. The models can model random variables with overdispersed and hierarchically dependent characteristics. The results indicate that distribution families can adequately characterize text data. It is due to its conformity with the sparse properties of natural language. The data sets used in the experiments were *MXM song lyric, Reuters* and *20-Newsgroup*.

In [12], they present two approaches based on regularization and factorization constraints. The objective was to incorporate knowledge about topic coherence in formulating topic models. The authors named their approaches NTM-R and NTM-F, respectively. NTM-R substantially improves topic coherence. The approaches take advantage of pre-trained word embeddings. The intent is to provide word embeddings with contextual similarity information related to the NPMI calculation.

 doi:10.20944/preprints202305.0908.v1

Therefore, the authors constructed a topic coherence regularization term. They also used the word embedding matrix as a factorization constraint. The data set used was *20-Newsgroup*.

In [13], they combine contextualized representations with topic models with neural networks. The combination is performed to present an extension of the Neural ProdLDA model. In addition, its approach produces meaningful and coherent topics. One of the evaluation metrics used was normalized topic coherence. The data sets used were *20-Newsgroup*, *Wiki20K*, *StackOverflow*, *Tweets2011* and *GoogleNews*. The proposed approach, the *20-Newsgroup* corpus, obtained an average of 0.1025 over five topic numbers.

[14] proposed a Variational Automatic Encoder (VAE) NTM model. The model reconstructs the sentence and word count of the document using combinations of a bag of words and word embedding. The experiments carried out demonstrated the effectiveness of the proposed model. The model reduced reconstruction errors at both the sentence and document levels. In addition to discovering topic coherence from real-world data sets. The data sets used were *wikitext-103*, *20-Newsgroup*, *COVID-19* and *NIPS*. The evaluation metric used was normalized topic coherence. The model comprises two types of encoders. The internal *BoW* data encoder and external knowledge encoder. It captures latent topics of documents, sentences, and words from internal and external sources. In addition, an attention-aware hierarchical divergence that regularizes the topical embedding of sentences within its documents. The results obtained for the *20-Newsgroup* corpus indicate that the encoder of external knowledge can obtain higher results.

In [15], they expose a novel model called the Pseudo-Document-Based Topic Model (PTM). This model introduces the concept of a pseudo-document to add short texts against the scarcity of data implicitly. They also proposed a word embedding-enhanced PTM (WE-PTM). The goal is to take advantage of pre-trained word embeddings essential to alleviate data scarcity. The authors conducted experiments with baselines based on autoaggregation or word embedding. The data sets used were *20-Newsgroup*, *tweets*, *DBLP* and *question*. The experiments demonstrated that the topics discovered by the proposed model are of high quality. The results were evaluated with the topic coherence metric.

In [16], they expose a method for extracting and selecting collocations as a preprocessing step. The model was named HLTA. Selected collocations are replaced with unique tokens in the bag of words model before running HLTA. The empirical evaluation showed that the proposed method improved HLTA performance on three data sets. The data sets used were *NIPS*, *AAN*, *JRC and Reuters*. The metric used to evaluate the method was the coherence of the topic.

In [17], they implement a recommendation system based on user comments in crowdfunding campaigns. The authors incorporated topic discovery with a long-term memory model (LSTM) to extract patterns in the analyzed comments. The proposed model is trained with Latent Dirichlet Allocation (LDA) with word embedding. The exposed model is capable of capturing long-range contextual and temporal dependencies. The author aimed to suggest safe and optimized recommendations to investors using their feedback. The metrics used in the evaluation process were prediction accuracy and topic coherence.

In [18], the authors present the automated extraction of discussions related to COVID-19. The social network called Reddit was used in this work. The authors investigated using an LSTM recurrent neural network for sentiment classification of COVID-19 comments. The authors concluded that it is essential to use public opinion and appropriate computational techniques to understand the problems related to COVID-19 and guide decision-making.

In [19], they expose two mixed counting neural models called Negative Binomial Neural Topic Model (NB-NTM) and Gamma Negative Binomial Neural Topic Model (GNB NTM) based on NB and GNB processes, respectively. The overall motivation of the authors was to combine the advantages of NVI and mixed count models. On the one hand, NVI-based models are fast and easy to estimate but challenging to interpret. The document modelling through mixed count models is easy to interpret but difficult to infer. Authors NB-NTM and GNB-NTM developed NVI algorithms to infer parameters using the reparameterization of the Gamma distribution and the Gaussian approximation of the

 doi:10.20944/preprints202305.0908.v1

Poisson distribution. Experiments on real-world data sets validated the effectiveness of their proposed models on perplexity, topic coherence, and sparse topic learning.

In [20], the authors present an analysis of comments about COVID-19 to detect feelings related to the disease. The data set was extracted from the Reddit social network. They then applied preprocessing and determined the polarity of each comment. For this, they used the VADER lexicon, which associates a sentiment rating to each word, followed by TextBlob. The discovery of the topics was carried out using the LDA algorithm. The number of extracted topics was five, with 10 top words. The words of each topic were tokenized with the gensim library and grouped into bigrams and trigrams. Based on the results, they mapped each word with Gensim's id2word and the different interpretations of the keywords with pyLDAvis. In addition, they applied a deep learning model called BERT to classify the topics.

In [21], they expose a model named Word2Vec2Graph. This model is based on the Word2Vec word embedding model. The authors applied the model to analyze long documents, obtain unexpected word associations, and discover topics in the documents. The discovered topics were validated, transformed with Gramian Angular Field into images, and used a convolutional neural network (CNN) for image classification. Semantic graphs were built using an unweighted method, and the connected components were calculated with weight threshold parameters. They then compared the high-weight and low-weight clusters—two domain-specific data corpora: a data corpus on Creativity and Aha Moments and another on psychoanalysis.

In [22], they present an approach to topic discovery. The data set used were comments from Twitter users. The approach takes a set of tweets as input. They then extracted the text representation for each tweet using an embedding model. They then grouped semantically similar tweets into groups using the HDBSCAN algorithm, each representing a topic. The authors indicated that they evaluated this through semantic cohesion. In addition, they applied a manual review of the topics obtained to verify that the topics were clear to the users.

In [23], they expose a hierarchical topic modelling algorithm. The algorithm is based on community detection. Furthermore, it is based on community mining of word co-occurrence networks taking advantage of the natural network structure. The algorithm has three steps: First, they build a network from the corpus of documents with terms as vertices. The edge weights were derived from the frequency of co-occurrence. The data sets used were *20-Newsgroup*, *Reuters* and *BBC*. The experiments showed that the exposed model produced topics with greater coherence and a more cohesive hierarchy. The features offered by the proposed model make it an ideal tool to guide the conversation of a chatbot. The authors applied the normalized topic coherence metric obtaining a coherence of 0.044 for the *20-Newsgroup* corpus and 0.182 for the *Reuters* corpus.

A Bayesian generative model is exposed in [24]. The proposed model describes thematic hierarchies that are organized into taxonomies. The experiments showed that the proposed model efficiently integrates prior knowledge and improves both the hierarchical discovery of topics and the representation of documents. The model was able to learn the latent distribution of document topics and the deterministic hierarchical embeddings of the document. The data sets used were *20-Newsgroup* and *Reuters*. The evaluation metric used was topic coherence.

In [25], expose the use of Kernel Principal Component Analysis (KernelPCA) and K-means Clustering in the BERTopic architecture. They have prepared a new dataset using tweets from customers of Nigerian banks, and they use this to compare the topic modelling approaches. The metric used is topic coherence. The experimental results showed that BERTopic with BERT components for embeddings, KernelPCA for dimension reduction, and K-means clustering achieved the highest consistency score than reported in the literature. The topics produced were bank inquiries, transaction problems, mobile applications and ATMs.

In [26], expose a new method of topic discovery. The method combines a pre-trained Bert model and a k-means spherical clustering algorithm and applies similarity between documents and topics. The method proposed was applied to literature abstracts related to geospatial data. The

 doi:10.20944/preprints202305.0908.v1

method exposes geospatial data technology features and research application development trends. The evaluation metric used was the coherence of the topic. The results showed that the proposed method could produce highly coherent topics. In addition, the research provided new ideas for analysing trends in technologies and applications related to geospatial data.

This paper proposes integrating document classification in topic discovery with semantic embedding models. Topic discovery is on classes previously obtained of each corpus used during the experiments. The first *20-Newsgroup* with 20 classes, and the second *Reuters* with 90 classes. The results were evaluated with the normalized topic coherence metric to evaluate the performance of the proposed model.

3. Proposed Approach

This section outlines the proposed procedure when performing document classification and integrating it into topic discovery with semantic embedding models. The proposed approach included the following process: classes of the *20-Newsgroup* and *Reuters* corpora obtained with a convolutional neural network (CNN) [7]; discovered specific topics in each of the classes previously obtained with the CNN with the Dirichlet Latent Analysis, Latent Probabilistic Semantic Analysis, and Latent Semantic Analysis algorithms; and an evaluation process. This work uses document classification, and it integrates on-topic discovery. The results show that using a previous classification for topic discovery results has a more significant relationship.

Figure 1 shows the proposed approach in this paper. Two previously preprocessed corpora were classified, from which 20 classes are obtained for the *20-Newsgroup* corpus and 90 for the *Reuters* corpus. The classes are the input data set to the Latent Dirichlet Analysis, Probabilistic Latent Semantic Analysis, and Latent Semantic Analysis algorithms for the topic discovery.

It was observed that with 20 extracted topics, good topic coherence results were obtained. The *20-Newsgroup* corpus has 20,000 documents, and it was only necessary to remove 20 classes, which allowed experiments to extract twenty, fifty, and 100 topics. On the other hand, the *Reuters* corpus has a smaller number of documents, and 90 classes must be extracted, so it is impossible to extract 100 topics because the more topics are extracted, the dispersion of the issues will affect the coherence results.

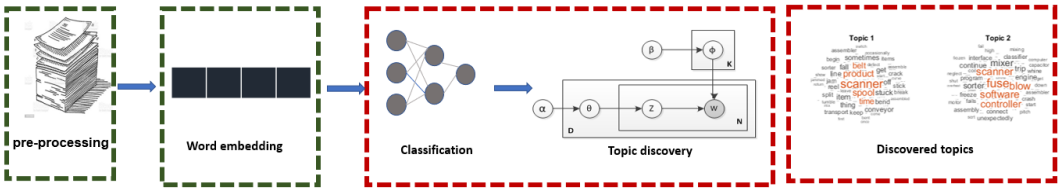

Figure 1. Proposed approach to incorporate text classification into topic discovery process.

3.1. Text classification

The texts were preprocessed by applying text cleaning, removing stop words, and converting them to lowercase. The text classification process does carry out as proposed in [7], which includes CNN, was used to evaluate the performance of three embedding models of semantic relations and generate a set of classes belonging to each of the corpora to does use. The classification task generated the corresponding classes in each corpus used. For the *20-Newsgroup* corpus, 20 classes were obtained and for the *Reuters* corpus, 90 classes. The classes are the basis for topic discovery since they are the input data set to each algorithm for topics. In addition, the classification was carried out to have an ordered corpus with a more significant relationship between the texts. The hypothesis focuses on how finding topics in a classified corpus will improve the coherence of the retrieved topics. Therefore integrating the classification of two corpora with semantic embedding models in the discovery of topics is the main contribution of this paper. Table 1 shows some classes recovered in each corpus.

 doi:10.20944/preprints202305.0908.v1

Table 1. Example of classes obtained in the corpus *20-Newsgroup* and *Reuters*.

Corpus	*Class*
20-Newsgroup	*...Atheism, Sport, Politics, Computing, Cars...*
Reuters	*...Aluminum, Barley, Bop, coffee, Cocoa...*

3.2. Topic Discovery

In the literature, algorithms for topic discovery are Latent Dirichlet Analysis (LDA), Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA). However, some authors have incorporated additional procedures into their approaches, such as classifying the texts before topic discovery. On the other hand, topic discovery has been an essential part of different tasks of the NLP, for example, sentiment analysis and decision-making.

Latent Dirichlet Analysis is an algorithm for topic discovery. The model is based on the hypothesis that each document contains words or terms from different topics. The distribution of topics in each document is different. This model needs to know a priori the texts and the number of topics to be found in the texts. This model is maintained under the premise that topics and documents are treated through Dirichlet distributions [27]. Figure 2 shows the LDA model graphically.

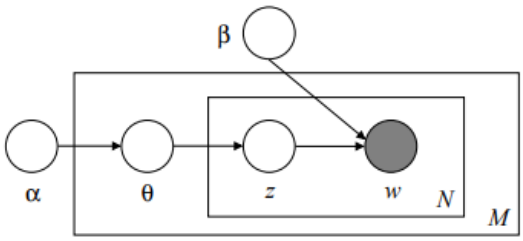

Figure 2. Graphical representation of the LDA model [28].

On the other hand, Latent Semantic Analysis (LSA) is another algorithm for topic discovery. LSA is a mathematical dimensionality reduction procedure called singular value decomposition (SVD). LSA, or LSI, is an automatic index analysis that projects terms and documents in a space of reduced dimensions. The reduction of attributes or dimensions of a document produces a recovery of the semantics of the original documents. The LSA dimensionality reduction process captures more important terms or topics [29]. Figure 3 shows the matrix generated by LSA model.

Finally, Probabilistic Latent Semantic Analysis (PLSA) continues LSA. In PLSA, words are attributed to latent topics or concepts based on the weighted frequency of some document terms. It interprets frequencies in terms of probability. PLSA is a descriptive statistical technique. The probability that a term forms part of the set of terms belonging to a topic or concept depends on different parameters obtained. These parameters are obtained by counting the given frequencies in a matrix based on a multinomial probability calculation. The objective of PLSA is to estimate the multinomial probability distribution of some words in a topic. Figure 4 shows the PLSA model graphically.

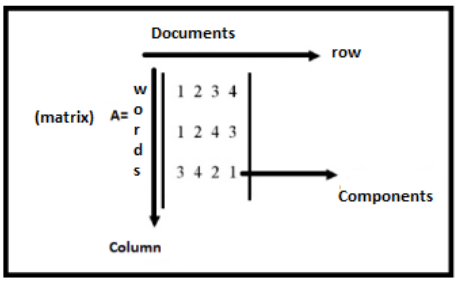

Figure 3. An example of a matrix generated by LSA [30].

 doi:10.20944/preprints202305.0908.v1

The topic discovery has as its starting point the integration of the classes previously obtained in [7]. The classes obtained by CNN from each corpus were the input set to the topic discovery algorithms. The goal is to provide each topic discovery algorithm with an ordered corpus with no unrelated text. The data sets used were *20-Newsgroup* and *Reuters* exposed in Section 4.1.

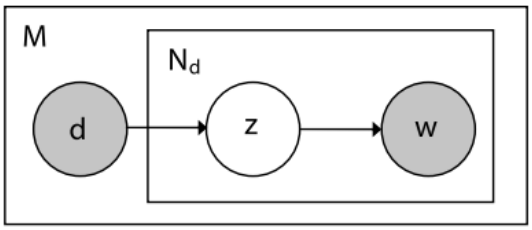

Figure 4. Graphical representation of the PLSA model [31].

The classes are the input data set to the Latent Dirichlet Analysis, Probabilistic Latent Semantic Analysis, and Latent Semantic Analysis algorithms. For each algorithm, 20, 50 and 100 topics with 10 top words are extracted for the *20-Newsgroup* corpus, which is generated for the *20-Newsgroup* corpus 4,000, 10,000 and 20,000 topics. On the other hand, for the *Reuters* corpus, 20 topics with 10 top words were extracted, generating 18,000 topics [29].

The evaluation of topic discovery is performed with the normalized topic coherence metric described below.

The normalized topic coherence consists of obtaining the *Normalized Coherence* of each topic (t_i). Measures the semantic relevance of the most important words of a topic, which is computed by the normalized pointwise mutual information (NPMI) over the selected words of each topic. Which is described below:

$$f(w_i, w_j) = \frac{[\log \frac{p(w_i,w_j)}{p(w_i)p(w_j)}]}{[-\log p(w_i, w_j)]} \quad (1)$$

Normalized Coherence is based on obtaining the normalized mutual point information (NPMI) of each pair of words belonging to the k top words representing each topic. The metric is based on calculating the probabilities that the k top words co-occur within the same paragraph of the set of external texts, in this case, Wikipedia in English [32].

4. Results and Discussion

This section presents the results of integrating document classification in topic disco-very with semantic embedding models. In addition, the results obtained are compared with those found in the literature.

The results provided a vision of integrating the classification of two corpora with semantic embedding models(exhibited in [7]) and the topic discovery.

The following sections present the results obtained and evaluated with the normalized topic coherence metric, as well as the data sets used in the development of this work.

4.1. Datasets

A corpus in English from Wikipedia was used as a reference corpus to evaluate the topics. Table 2 exposes each dataset's number of documents and tokens. That is Wikipedia for the evaluation of the topics and *Reuters* (https://trec.nist.gov/data/reuters/reuters.html, accessed on 1 May 2020) and *20 -Newsgroup* (http://qwone.com/~jason/20-Newsgroups/, accessed on 1 May 2020) for topic discovery.

 doi:10.20944/preprints202305.0908.v1

Table 2. Description of dataset.

Corpus	Documents	Tokens
Wikipedia	1,000,000	1,560,478,211
20-Newsgroup	20,000	1,800,385
Reuters	18,456	3,435,808

The LDA, LSA and PLSA algorithms were applied. 20, 50 and 100 topics with 10 top words were extracted for the *20-Newsgroup* corpus. For the *Reuters* corpus, only 20 and 50 topics with 10 top words were extracted. The median, mean, and standard deviation were extracted from the results obtained. The objective was to identify trends or where the corpora *20-Newsgroup* and *Reuters* lean or group more. In this way, it was possible to analyze and visualize the results of each corpus, in addition to inferring about the results obtained. The number of topics (n), mean (Avg), and standard deviation (std) are presented. The algorithm that obtained a high normalized topic coherence was obtained by extracting 20 topics with 10 top words from the *20-Newsgroup* and *Reuters* corpus classes. Table 3 exposes the number of topics (n), the mean (Avg), and the standard deviation (std) of the normalized coherence of the topics extracted from the classes from the *20-Newsgroup* and *Reuters* corpus with the LDA algorithm. The average and the standard deviation of the normalized coherence of each discovered topic in each class are obtained from extracted 20, 50 and 100 topics from *20-Newsgroup* corpus. It is observed that the algorithm that obtained a high normalized coherence was obtained by extracting 20 topics with 10 top words from the corpus classes. Hence in each of the tables presented in this paper, only three top words are shown as an example. Also Table 3 exposes the number of topics (n), the mean (Avg) and the standard deviation (std) of the normalized topic coherence of the discovered topics from the classes from the *Reuters* corpus with the LDA algorithm. The average and the standard deviation of each discovered topic's normalized coherence in each class are obtained with 20 topics. Therefore, Table 4 presents only three top words shown as an example from *Reuters* corpus. Table 5 lists the discovered topics in three corpus classes from the *20-Newsgroup* corpus.

The results are coherent since 20 topics are extracted in each of the 20 previously recovered classes; 400 topics are obtained for the 20 existing classes. They generate topics that relate to the class from which they were extracted and therefore had a relationship between the words that represent them.

Table 3. Average normalized topic coherence for the LDA algorithm with 20, 50, and 100 topics for the corpus *20-Newsgroup* and *Reuters*.

n	*20-Newsgroup*		*Reuters*	
	Avg	*std*	*Avg*	*std*
LDA_20	**0.1723**	0.0104	**0.1441**	0.0472
LDA_50	0.1572	0.0116	0.1394	0.0165
LDA_100	0.1453	0.0097	-	-

Table 4. Top words of the corpus *Reuters* with the LDA algorithm with 20 topics with 10 top words.

	Topics		
Class	**Company**	**Disasters**	**Prices**
Aluminum	*...operations, dump, ton...*	*...debris, Panamá, toll...*	*...automotive, coin, finance...*
Barley	**Cultivation** *...acreage, wheat, department...*	**Grain** *...acreage, corn, farm...*	**Sales** *...wheat, export, tonnes...*

 doi:10.20944/preprints202305.0908.v1

Table 4. *Cont.*

Bop	**Finance** *...pressure, country, finance...*	**Duty** *...dollar, oil, price...*	**Trade** *...billion, export, deficit...*

Table 5. Top words of the *20-Newsgroup* corpus for the LDA algorithm with 20 topics and 10 top words.

	Topics		
Class	**Religion**	**Rituals**	**Items**
Atheism	*...town, big, life...*	*...bible, ceremonial, love...*	*...believe, officer, children...*
Computing	**Sofware** *...virtual, video, files...*	**Hardware** *...cpu, harddisk, machine...*	**Several** *...company, host, unix...*
Cars	**Elements** *...battery, bad, street...*	**Characteristics** *...ford, color, models...*	**Others** *...video, computer, design...*

Table 6 exposes the number of topics (n), the mean (Avg) and the standard deviation (std) of the normalized coherence of the topics extracted from the classes from the *20-Newsgroup* corpus with the LSA algorithm. The average and the standard deviation of the normalized topic coherence of each discovered topic in each class are obtained by extracting 20, 50 or 100 topics. It is observed that the algorithm obtained a high normalized topic coherence was obtained by extracting 20 topics with 10 top words from the corpus classes.

Table 6 also shows the results obtained by the LSA algorithm with 20 and 50 topics for the *Reuters* corpus. It is observed that the algorithm obtained a high normalized topic coherence was again obtained by extracting 20 topics with 10 top words from the corpus classes.

Therefore, Table 7 presents only three top words shown as an example for *20-Newsgroup* corpus and Table 8 for *Reuters*. As in LDA and PLSA for the *20-Newsgroup* corpus, the input data to LSA are strictly 20 classes previously obtained with a neural network. In this case, the search with 20 topics was sufficient and coherent since we are working with 20 different classes and extracting 20 topics into each class. However, for the *Reuters* corpus, the input data to LSA are strictly 90 classes previously obtained with the same neural network. In this case, the search with 20 topics was sufficient and coherent since extracting more topics will affect the coherence of results.

Table 6. Average normalized coherence with the LSA algorithm with 20, 50 and 100 topics for the corpus *20-Newsgroup and Reuters*.

n	***20-Newsgroup***		***Reuters***	
	Avg	***std***	***Avg***	***std***
LSA_20	**0.1622**	0.0158	**0.1360**	0.0176
LSA_50	0.1556	0.0095	0.1342	0.0170
LSA_100	0.1462	0.0098	-	-

 doi:10.20944/preprints202305.0908.v1

Table 7. Top words of the corpus *20-Newsgroup* with the LSA algorithm with 20 topics and 10 top words.

	Topics		
Class	**Religion**	**Rituals**	**Items**
Atheism	*...goodless, sabbath, ceremonial...*	*...course, started, hostage...*	*...individuals, officer, children...*
Computing	**Sofware** *...software, driver, email...*	**Hardware** *...circuits, video, pc...*	**Several** *...companies, work, zip...*
Cars	**Elements** *...price, batteries, street...*	**Characteristics** *...price, wheel, oil...*	**Others** *...video, computer, concrete...*

Table 8. Top words of the corpus *Reuters* with the LSA algorithm with 20 topics with 10 top words.

	Topics		
Class	**Company**	**Disasters**	**Prices**
Aluminum	*...dlrs, bank, group...*	*...debris, Portugal, industries...*	*...pct, coin, rise...*
Barley	**Cultivation** *...maize, wheat, department...*	**Grain** *...corn, corn, drum...*	**Sales** *...sunflowers, export, tonnes...*
Bop	**Finance** *...singapore, britoil, finance...*	**Duty** *...japan, oil, price...*	**Trade** *...legislation, export, president...*

Table 9 exposes the number of topics (n), the mean (Avg) and the standard deviation (std) of the normalized topic coherence of the discovered topics from the classes from the *20-Newsgroup* corpus with the PLSA algorithm. The average and the standard deviation of the normalized topic coherence of each discovered topic in each class are obtained by extracting 20, 50 or 100 topics. It is observed that the algorithm obtained a high normalized topic coherence obtained by extracting 20 topics with 10 top words from the corpus classes since its mode has a value of 0.1735, the highest result. However, the most significant result for this experiment's corpus *Reuters* was when 20 discovered topics with the PLSA algorithm on each input class. It is observed that the algorithm that generated a high normalized topic coherence was the one obtained with 20 discovered topics with 10 top words from the corpus classes with the PLSA algorithm.

Therefore, Table 10 presents only three top words shown as an example of the *20-Newsgroup* corpus and Table 11 for *Reuters* corpus.

 doi:10.20944/preprints202305.0908.v1

Table 9. Average normalized coherence with the PLSA algorithm with 20, 50 and 100 topics for the corpus *20-Newsgroup* and *Reuters*.

n	*20-Newsgroup*		*Reuters*	
	Avg	*std*	*Avg*	*std*
PLSA_20	0.1716	0.0099	0.1436	0.0160
PLSA_50	0.1559	0.0095	0.1409	0.0531
PLSA_100	0.1457	0.0095	-	-

Table 10. Top words of the corpus *20-Newsgroup* with the PLSA algorithm with 20 topics with 10 top words.

	Topics		
Class	**Religion**	**Rituals**	**Items**
Atheism	*...godless, jewish, sabbath...*	*...ceremonial, law, love...*	*...people, childs, court...*
Computing	**Sofware** *...windows, system, copies...*	**Hardware** *...mail, acceso, location...*	**Several** *...machines, work, cpu...*
Cars	**Elements** *...gas, models, pay...*	**Characteristics** *...battery, wheel, oil...*	**Others** *...software, alert, safety...*

Table 11. Top words of the corpus *Reuters* with the PLSA algorithm with 20 topics with 10 top words.

	Topics		
Class	**Company**	**Disasters**	**Prices**
Aluminum	*...godless, jewish, sabbath...*	*...group, law, love...*	*...corp, commodity, monetary...*
Barley	**Cultivation** *...production, system, acres...*	**Grain** *...tonnes, february, export...*	**Sales** *...machines, work, cpu...*
Bop	**Finance** *...gas, models, pay...*	**Duty** *...battery, wheel, oil...*	**Trade** *...carney, countries, government...*

4.2. Experimental Results

The proposed approach was evaluated with the normalized topic coherence metric described in the Section 3.2. The corpora used were *Reuters* and *20-Newsgroup*. For this evaluation, different configurations of parameters and sizes were used. The results obtained were compared with those existing in the literature, and the conclusion was that the results in this paper obtained better results when 20 discovered topics with 10 top words.

The results obtained with the proposed approach provided a vision of integrating classes as input data for each algorithm of topic discovery. Although the results are somewhat low, we have the hypothesis about applying additional parameters; the results obtained will be more significant. The

 doi:10.20944/preprints202305.0908.v1

approach can provide consistent topics about the language and domain used. The approach is applied to the LDA, LSA and PLSA techniques, combining the number of parameters 20, 50 and 100 to discover, respectively, to observe the approach's behaviour according to the number of discovered topics.

Table 12 shows the results of different authors in the literature. The authors used the normalized topic coherence evaluation metric and the corpora *20-Newsgroup* or/and *Reuters*. However, not all authors apply document classification or clustering algorithms before performing topic discovery. The results obtained in this paper are higher than the results obtained by the authors [8–10,12–14]. On the other hand, authors such as [14] obtain higher coherence values for the *20-Newsgroup* corpus. The corpus *Reuters* [23,24] obtain higher coherence values than the results obtained in this work. In [14]b and [24], the results are significant to the obtained in this paper because they apply algorithms and methods like variational automatic encoder and community detection and community mining. We consider the methods mentioned before to benefit the authors' results. In this paper, the objective was the integration of classification texts into topic discovery. Hence no additional method was contemplated.

The proposed approach obtained a coherence of 0.172 for the *20-Newsgroup* corpus with the LDA algorithm with 20 topics and 0.147 for the *Reuters* corpus with the LSA algorithm with 20 topics.

Table 12. Comparison of the average of the results obtained with normalized topic coherence for both corpus.

Author	20-Newsgroup	Reuters
[8]	0.10	0.04
[9]	0.103	0.152
[10]	0.39	-
[12]	0.28	-
[13]	0.102	-
[14]a	0.042	-
[14]b	0.279	-
[23]	0.044	0.182
[24]	0.17	0.18
This work	0.172	0.147

5. Conclusions and Future Work

This paper presents the integration of document classification into topic discovery with semantic embedding models. The objective is to discover topics with a more significant relationship between the words that form the topics. In document classification, a convolutional neural network was used in addition to three previously developed semantic relationship embedding models. The classes obtained results from the classification with the semantic relations model that involves synonymy, hyponymy, and hyperonymy relations integrated into the topic discovery. The topic discovery was carried out with the algorithms LDA, LSA and PLSA. In the three algorithms, it is strictly necessary that the data set they receive are the classes obtained by classifying with the model that involves the three embedding models of semantic relationships. The *20-Newsgroup* corpus has twenty classes, and 20, 50 and 100 discovered topics with ten top words in each case. However, the corpus *Reuters* has 90 classes and only 20 discovered topics with ten top words. The results are compared with the literature, demonstrating that topic discovery in classes generates results with a more significant relationship between the texts that form it. The discovered topics were evaluated with the normalized topic consistency metric (NPMI). The results showed that integrating text classification provided mostly related texts in each retrieved class; therefore, topics with top words were more related. The main contribution of this work is an integration of classification in topic discovery with semantic embedding models. We compare

 doi:10.20944/preprints202305.0908.v1

the performance of integrating the classification of documents with the topic discovery with works of literature. The results express the importance of having a classified text to discover relevant topics. The approach becomes a helpful resource in natural language, demonstrating that adding semantics to a classification process will bring positive results. In future work, it is proposed to label the discovered topics with the help of an ontology of the same domain and then compare the results obtained in both experiments.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Vásquez, A.C.; Quispe, J.P.; Huayna, A.M.; others. Procesamiento de lenguaje natural. *Revista de investigación de Sistemas e Informática* **2009**, *6*, 45–54.
2. Ramos, F.; Vélez, J. Integración de técnicas de procesamiento de lenguaje natural a través de servicios web. *Universidad Nacional del Centro de la provincia de Buenos Aires* **2016**.
3. Almeida, F.; Xexéo, G. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069* **2019**.
4. Lezama Sánchez, A.L.; Tovar Vidal, M.; Reyes Ortiz, J.A. A Behavior Analysis of the Impact of Semantic Relationships on Topic Discovery. *Computación y Sistemas* **2022**, *26*, 149–160.
5. Saedi, C.; Branco, A.; Rodrigues, J.; Silva, J. Wordnet embeddings. Proceedings of the third workshop on representation learning for NLP, 2018, pp. 122–131.
6. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. *Information* **2019**, *10*, 150.
7. Lezama-Sánchez, A.L.; Tovar Vidal, M.; Reyes-Ortiz, J.A. An Approach Based on Semantic Relationship Embeddings for Text Classification. *Mathematics* **2022**, *10*, 4161.
8. Fuentes-Pineda, G.; Meza-Ruiz, I.V. Topic discovery in massive text corpora based on min-hashing. *Expert Systems with Applications* **2019**, *136*, 62–72.
9. Wang, D.; Zhao, H.; Guo, D.D.; Liu, X.; Li, M.; Chen, B.; Zhou, M. BAT-Chain: Bayesian-Aware Transport Chain for Topic Hierarchies Discovery.
10. Xu, Y.; Wang, D.; Chen, B.; Lu, R.; Duan, Z.; Zhou, M. HyperMiner: Topic Taxonomy Mining with Hyperbolic Embedding. *arXiv preprint arXiv:2210.10625* **2022**.
11. Wu, J.; Rao, Y.; Zhang, Z.; Xie, H.; Li, Q.; Wang, F.L.; Chen, Z. Neural mixed counting models for dispersed topic discovery. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 6159–6169.
12. Ding, R.; Nallapati, R.; Xiang, B. Coherence-aware neural topic modeling. *arXiv preprint arXiv:1809.02687* **2018**.
13. Bianchi, F.; Terragni, S.; Hovy, D. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974* **2020**.
14. Jin, Y.; Zhao, H.; Liu, M.; Du, L.; Buntine, W. Neural attention-aware hierarchical topic model. *arXiv preprint arXiv:2110.07161* **2021**.
15. Zuo, Y.; Li, C.; Lin, H.; Wu, J. Topic modeling of short texts: A pseudo-document view with word embedding enhancement. *IEEE Transactions on Knowledge and Data Engineering* **2021**.
16. Poon, L.K.; Zhang, N.L.; Xie, H.; Cheng, G. Handling collocations in hierarchical latent tree analysis for topic modeling. *arXiv preprint arXiv:2007.05163* **2020**.
17. Shafqat, W.; others. A Hybrid Approach for Topic Discovery and Recommendations Based on Topic Modeling and Deep Learning. PhD thesis, 2020.
18. Jelodar, H.; Wang, Y.; Orji, R.; Huang, S. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics* **2020**, *24*, 2733–2742.

19. Wu, J.; Rao, Y.; Zhang, Z.; Xie, H.; Li, Q.; Wang, F.L.; Chen, Z. Neural mixed counting models for dispersed topic discovery. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 6159–6169.
20. Pandey, C. redBERT: A topic discovery and deep sentiment classification model on COVID-19 online discussions using BERT NLP model. *International Journal of Open Source Software and Processes (IJOSSP)* **2021**, *12*, 32–47.
21. Romanova, A. Semantics graph mining for topic discovery and word associations. *Int. J. Data Mining Knowl. Manag. Process (IJDKP)* **2021**, *10*.
22. Stanik, C.; Pietz, T.; Maalej, W. Unsupervised topic discovery in user comments. 2021 IEEE 29th International Requirements Engineering Conference (RE). IEEE, 2021, pp. 150–161.
23. Austin, E.; Trabelsi, A.; Largeron, C.; Zaïane, O.R. Hierarchical Topic Model Inference by Community Discovery on Word Co-occurrence Networks. Data Mining: 20th Australasian Conference, AusDM 2022, Western Sydney, Australia, December 12–15, 2022, Proceedings. Springer, 2022, pp. 148–162.
24. Wang, D.; Xu, Y.; Li, M.; Duan, Z.; Wang, C.; Chen, B.; Zhou, M.; others. Knowledge-aware Bayesian deep topic model. *Advances in Neural Information Processing Systems* **2022**, *35*, 14331–14344.
25. Ogunleye, B.; Maswera, T.; Hirsch, L.; Gaudoin, J.; Brunsdon, T. Comparison of Topic Modelling Approaches in the Banking Context. *Applied Sciences* **2023**, *13*, 797.
26. Cheng, Q.; Zhu, Y.; Song, J.; Zeng, H.; Wang, S.; Sun, K.; Zhang, J. Bert-Based Latent Semantic Analysis (Bert-LSA): A Case Study on Geospatial Data Technology and Application Trend Analysis. *Applied Sciences* **2021**, *11*, 11897.
27. Valero Moreno, A.I. Técnicas estadísticas en Minería de Textos **2017**.
28. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *Journal of machine Learning research* **2003**, *3*, 993–1022.
29. Venegas, R. La similitud léxico-semántica en artículos de investigación científica en español: Una aproximación desde el Análisis Semántico Latente. *Revista signos* **2006**, *39*, 75–106.
30. Torres López, C. Segmentación y detección de tópicos enfocado a la minería de opinión. PhD thesis, Universidad Central "Marta Abreu" de Las Villas, 2016.
31. Niebles, J.C.; Wang, H.; Fei-Fei, L. Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision* **2008**, *79*, 299–318.
32. Wales, J.; Sanger, L. `https://dumps.wikimedia.org/enwiki/20230101/`, 15 de enero de 2001.