


Article

Integrating Text Classification into Topic Discovery Using 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, Mexico² Departamento de Sistemas, Universidad Autonoma Metropolitana, Mexico City 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 involves identifying the main ideas within large volumes of textual data. It indicates recurring topics in documents, providing an overview of the text. Current topic discovery models receive the text, with or without pre-processing, including stop word removal, text cleaning, and normalization (lowercase conversion). A topic discovery process that receives general domain text with or without processing generates general topics. General topics do not offer detailed overviews of the input text, and manual text categorization is tedious and time-consuming. Extracting topics from text with an automatic classification task is necessary to generate specific topics enriched with top words that maintain semantic relationships among them. Therefore, this paper presents an approach that integrates text classification for topic discovery from large amounts of English textual data, such as *20-Newsgroups* and *Reuters* Corpora. We rely on integrating automatic text classification before the topic discovery process to obtain specific topics for each class with relevant semantic relationships between top words. Text classification performs a word analysis that makes up a document to decide what class or category to identify; then, the proposed integration provides latent and specific topics depicted by top words with high coherence from each obtained class. Text classification accomplishes this with a convolutional neural network (CNN), incorporating an embedding model based on semantic relationships. Topic discovery over categorized text is realized with latent Dirichlet analysis (LDA), probabilistic latent semantic analysis (PLSA), and latent semantic analysis (LSA) algorithms. An evaluation process for topic discovery over categorized text was performed based on the normalized topic coherence metric. The *20-Newsgroups* corpus was classified, and twenty topics with the ten top words were identified for each class. The normalized topic coherence obtained was 0.1723 with LDA, 0.1622 with LSA, and 0.1716 with PLSA. The *Reuters* Corpus was also classified, and twenty and fifty topics were identified. A normalized topic coherence of 0.1441 was achieved when applying the LDA algorithm, obtaining 20 topics for each class; with LSA, the coherence was 0.1360, and with PLSA, it was 0.1436.

Keywords: deep learning; natural language processing; topic discovery; text classification



Citation: Lezama-Sánchez, A.L.; Tovar Vidal, M.; Reyes-Ortiz, J.A. Integrating Text Classification into Topic Discovery Using Semantic Embedding Models. *Appl. Sci.* **2023**, *13*, 9857. <https://doi.org/10.3390/app13179857>

Academic Editor: Douglas

O'Shaughnessy

Received: 1 July 2023

Revised: 28 August 2023

Accepted: 29 August 2023

Published: 31 August 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

Natural language is the mechanism that a human being uses to communicate and transmit an idea, opinion, or feeling [1]. Understanding natural language is a complex task and requires time because millions of connections between neurons are necessary to learn it. However, a computer needs structure and logic to understand a programming or natural language. Therefore, a mathematical formula or a predefined pattern will be necessary for the computer to learn the required knowledge [1]. For a computer to recognize the data it receives, it must generate an adequate numerical representation.

Therefore, *natural language processing (NLP)* is an area that is in constant development; it seeks to generate efficient algorithms for a computer to understand the spontaneous language of a human being. Some of the characteristics of natural language involve strict

rules, which facilitate their computerized analysis [1]. Hence, *NLP* encompasses techniques and tools for developing systems that can interpret and utilize *natural language* to perform desired tasks, such as news classification or spam identification [1]. Mechanisms, such as the *extraction of semantic relationships, named entity recognition, topic discovery, and word embedding*, are essential to provide a computer with the necessary knowledge to process information and display inevitable results. Therefore, *topic discovery* is defined as the task of finding specific topics present in a set of documents as input. This task can be applied to any text [2]. The purpose is to identify, without needing a dictionary, the main themes that implicitly exist within a collection of texts [2]. For a computer to understand *natural language*, it is necessary to create vectors of numbers. The embedding vectors or patterns may be subject to operations, such as addition, subtraction, and distance measurements. The literature shows that some word-embedding models are based on neural networks or context matrices [3]. The advancement of technology has made it possible to streamline processes, such as:

1. Searching for the subject of a document.
2. Searching for a specific document.
3. Generating a summary or extracting key phrases from a text.

In the literature, some *word embedding* models are *word2vec, glove, fastText, and BERT*. The embedding or word embedding model came to fruition in 2013 when Tomas Mikolov and his team at Google developed the first embedding model, named *word2vec*. The model has the following sub-models:

1. Continuous bag-of-words (CBOW [4]): receives a context and predicts a target word [4].
2. Skip-gram [5]: each word is represented as a bag of n -grams of [6] characters.

The *GloVe* embedding model was developed in 2014 by Jeffrey Pennington [7]. This model combines the advantages of the two main family models in the literature: *global matrix factorization* and *local context window*. *GloVe* works with non-zero elements in a word–word co-occurrence matrix rather than the entire sparse matrix or separate context windows in a large corpus [7].

On the other hand, in 2015, *Facebook* researchers created the embedding model called *fastText*. The *fastText* model has pre-trained models for 294 languages. The authors relied on the *skip-gram* [8] model. In 2018, *BERT (bidirectional encoder representations from transformers)* was designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning left and right contexts in all layers [9]. In [3], the authors applied *GloVe* and *fastText* for the text classification of two text corpora, and compared the results that were obtained with a semantics embedding model.

Hence, *text classification* involves processing data without a person's intervention. The computer must have access to the knowledge necessary to carry out tasks such as medical diagnoses, analysis of social networks, or the 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 [10].

In addition, *deep learning* is a process carried out with a neural network, for example, a convolutional neural network. It has been adopted for text classification tasks, generating better results than a traditional classification task [11].

A *CNN* is a multi-layer network or hierarchical network and is a high-level feature-based method. *CNN* builds by stacking multiple layers of features. One feature of a *CNN* is the presence of a subsampling or pooling layer [12]. It allows for optimizing the calculation processes to reduce the data size in learning new data, allowing the recognition of different features [3].

Currently, computational approaches need to model knowledge to generate accurate results without the intervention of a person. *Text classification* involves ordering large amounts of documents in short periods. On the other hand, *topic discovery* involves finding the main ideas 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 automatically extracting knowledge from the text. It indicates the recurring topics in the documents, allowing for an overview of the text. A topic discovery model that receives text with or without processing generates general topics since the input data include many unclassified texts.

However, discovering topics from previously classified text allows us to learn specific topics. Hence, the top words that characterize the topics are linked to each other since they come from classified text [13].

This paper presents a text classification process integrated into identified topics with semantic embedding models. This incorporation provides specific topics with specific significance instead of general topics from the complete set of unclassified text since the topics are extracted from each identified class. The input texts are composed of two news domain corpora, previously classified with a convolutional neural network, using three semantic embedding models [3] as semantic features. The proposed topic discovery process aims to obtain specific topics for each class with semantic relationships between their top words. A quality assessment of the identified topics was performed with the normalized topic coherence metric. Therefore, the identified topics in each class provide latent and specific topics depicted by top words with high coherence from each obtained class. Based on the results obtained by integrating text classification with topic discovery, it was concluded that discovering topics in previously classified text generates specific topics from each class.

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

2. Related Works

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

In the literature, some authors created different models for discovering topics. The authors in [14–19] used the normalized topic coherence metric to evaluate the results obtained. In [14], a topic model based on min-hashing was proposed to find sets of matching words, which were subsequently grouped to produce the existing topics in the analyzed text.

On the other hand, in [17,20–23], the authors applied different models, like encoder, LSTM, and matrix factorization in their research. In [17], the authors combined contextualized representations with topic models via neural networks. The combination presents an extension of the neural-named ProLDA model. On the other hand, ref. [18] proposed a variational automatic encoder (VAE) NTM model. The model reconstructs the sentence and word count of the document by using combinations of bag-of-words and word embedding.

Furthermore, in [24], the authors showed the pseudo-document-based topic model (PTM), which introduces the concept of a pseudo-document to add short text against the scarcity of data implicitly. They also proposed a word embedding PTM (WE-PTM). On the other hand, in [21], a hierarchical latent tree analysis was proposed for hierarchical topic modeling, with the extracting and selecting collocations as a preprocessing step. The model was named HLTA. Selected collocations were replaced with unique tokens in the bag-of-words model before running HLTA. In [22], the authors presented the automated extraction of discussions related to COVID-19 and applied an LSTM recurrent neural network for sentiment classification. In [23], they showed two mixed counting neural models, called the negative binomial-neural topic model (NB-NTM) and the gamma negative binomial-neural topic model (GNB NTM). However, in [23], the authors showed two models for discovering scattered topics. The first model involved the negative binomial-neural topic model (NB-NTM) subjects, and the second involved gamma negative binomial-neural topic model (GNB-NTM) subjects. In [20], the authors presented two approaches, NTM-R and

NTM-F, which were based on regularization and factorization constraints. The objective was to incorporate knowledge about topic coherence in formulating topic models.

The authors of [19,25–30] applied word embedding to discover topics in long and short texts, or they applied clustering algorithms. Hence, in [25,26], the authors showed models based on word embeddings. In [25], the authors presented a model named Word2Vec2Graph, based on the Word2Vec model. The authors applied the model to analyze long documents, obtain unexpected word associations, and discover topics in the papers. On the other hand, in [26], the authors presented an approach to topic discovery and extracted text representations of tweets using a word embedding model. They then grouped them into semantically similar groups using the HDBSCAN algorithm, with each representing a topic.

In addition, in [19], the authors presented a hierarchical topic modeling algorithm. The algorithm is based on community mining of word co-occurrence networks, taking advantage of the natural network structure. However, a Bayesian generative model was shown in [27]. The model describes thematic hierarchies organized into taxonomies. The experiments show that the proposed model integrates prior knowledge and improves both the hierarchical discovery of topics and the representation of documents. In [28], the authors presented the use of the kernel principal component analysis (KernelPCA) and K -means clustering in BERTopic architecture. In [29], the authors presented a new method, combining a pre-trained BERT model and a K -clustering algorithm, applying similarity between documents and topics. Furthermore, ref. [30] proposed a polymerization topic sentiment model (PTSM) to conduct textual analysis for online reviews.

The authors of [31] incorporated topic discovery with a long-term memory model (LSTM) to extract patterns in the analyzed comments in crowdfunding campaigns. The proposed model trains with latent Dirichlet allocation (LDA) with word embedding. In [32], the authors presented an analysis of comments about COVID-19 to detect feelings related to the disease. 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.

A dependency SCOR-topic sentiment (DSTS) model was offered in [33]. The authors used online tea sales data as empirical evidence to test the proposed model. The results show that the DSTS model is generally superior to the LDA and PLSA models. In addition, in [15], each document was interpreted as word embeddings and a two-way model for discovering multi-level topic structures. In each layer, it learns a set of topical embeddings. On the other hand, in [16], the authors presented an approach that introduced hyperbolic embeddings to represent words and topics. In 2022, the authors of [34–36] presented their contributions in this field with text classification tasks and topic discovery.

TextNetTopics [34] is an approach that applies feature selection by considering the bag-of-topics (BOT) approach rather than the traditional bag-of-words (BOW) approach. This paper suggested scoring topics to select the top topics for training the classifier, hence reducing dimensionality and preserving the semantic descriptions of documents. On the other hand, in [35], the authors proposed considerations for selecting a suitable topic model based on the predictive performance and interpretability measures for text classification. Using clinical notes, they compared 17 different topic models regarding interpretability and predictive performance in an inpatient violence prediction task. Finally, ref. [36] presented a comprehensive survey of algorithms for short text topic discovery, and performance was evaluated using text classification.

On the other hand, in [37,38], the authors presented their contributions in the same field. In [37], the authors presented a model combining the advantages of unsupervised topic modeling with supervised string kernels for text classification tasks. The top words in the identified topics reduced the document corpus to a topic–word sequence. This reduction was used for text classification with string kernels, significantly improving accuracy and reducing training time. For [38], an approach to discovering topics through cosine similarity brought great results. First, the authors extracted synonyms from a semantic network, and

in this way, relevant topics from datasets such as Yahoo and BBC News were identified. They used text classification models, such as support vector machines, decision trees, and random forests to carry out the text classification task.

The authors of [39] considered open information extraction techniques to integrate into text classification tasks, considering semantic aspects in different languages. Hence, the authors presented an approach to enrich the open information extraction paradigm by exploiting syntactic and semantic analysis and semantic relations from an ontology. The authors used the English Wikipedia as a dataset. On the other hand, in [40], the authors presented an approach based on patterns and ontologies for information extraction and integration with other tasks; hence, in [40], the authors experimented with building an open information extraction system (OIE) for text in the Italian language. The authors proposed an approach that relied on linguistic structures and a set of verbal patterns, combining theoretical linguistic knowledge and corpus-based statistical information. Also, ref. [41] presented an approach to perform open information extraction (OIE) for the Italian language; it was based on linguistic structures to analyze sentences and a set of verbal behavior patterns to extract information from them. The patterns combined a linguistic theoretical framework (such as lexicon-grammar (LG)) and distributional profiles extracted from a contemporary. In addition, the authors of [42] presented a multi-²OIE, which performed open information extraction (OIE) by using multilingual BERT. The model is a sequence-labeling system with an extraction method. On the other hand, in [43], the authors presented an overview of the current situation of neural information extraction models, focusing on the advantages, disadvantages, and future of work in the field. In [44,45], the authors applied machine learning and sentiment analysis techniques to the COVID-19 domain. The authors of [44] proposed a methodology for sentiment analysis based on natural language processing (NLP) and sentiment analysis to obtain insight into opinions on COVID-19 vaccination in Italy. Ref. [45] presented an analysis of the scoping review of AI in COVID-19 research. However, in [46], the authors presented DEFIE, an approach to information extraction (IE) based on the syntactic-semantic analysis of textual definitions and techniques involving semantic aspects; they harvested instances of semantic relations from a corpus of textual descriptions. The aim was to extract as much information as possible by unifying syntactic analysis with state-of-the-art disambiguation and entity linking. An extensive knowledge base was produced against state-of-the-art OIE systems based on much larger corpora. In this analysis, only in reference [19] was text classification and topic discovery used to obtain the sentiment associated with the text. The rest of the works used topic discovery with other techniques or algorithms (but not text classification). Therefore, this work seeks to provide a methodology that classifies text by analyzing the words that make up a document, to decide which class it identifies. This results in discovering specific topics for each class with a higher semantic relationship between their primary words.

This paper proposes integrating a text classification process with semantic embedding models to discover specific topics. Specific topics are identified in each class identified in each corpus. The 20-Newsgroups corpus has twenty classes, and the Reuters Corpus has ninety classes. The results are evaluated with the normalized topic coherence metric to assess the performance of the proposed model.

3. Proposed Approach

This section presents the proposed approach for text classification and integrates it into topic discovery with semantic embedding models and three algorithms.

The proposed approach for integrating text classification in the topic discovery process incorporates the following process:

1. Pre-processing: The text sets are pre-processed; this consists of removing punctuation marks, converting to lowercase, and removing URL marks.
2. The semantic relationship extraction: Extracting semantic relationships from the English Wikipedia corpus is vital for constructing the proposed embedding mod-

els. It is necessary to extract the relations of synonymy, hyponymy, and hyperonymy using lexical–syntactic patterns extracted from the literature for these semantic relationships [3].

3. Development of embedding models: Each word pair of identified relationships is assigned a unique identifier for constructing semantic relationship embeddings [3].
4. Text classification from *20-Newsgroups* and *Reuters* Corpus with a convolutional neural network (CNN) [3].
5. Topic discovery for each class with the *latent Dirichlet analysis (LDA)*, *probabilistic latent semantic analysis (PLSA)*, and *latent semantic analysis (LSA)* algorithms. An evaluation process based on normalized topic coherence using the top words is performed.

The *LDA*, *PLSA*, and *LSA* algorithms are chosen because they are the most used algorithms in the literature. It is possible to contrast with works that only use the input data with traditional preprocessing and discover topics against those works where they perform additional processing to the traditional one, such as text classification and community detection algorithms.

Figure 1 shows the proposed approach. A total of 2 previously preprocessed corpora were classified, from which, 20 classes were obtained for the *20-Newsgroups* corpus and 90 for the *Reuters* Corpus. The classes are the input data set to the *LDA*, *PLSA*, and *LSA* algorithms for topic discovery. The *20-Newsgroups* corpus comprises 20,000 documents, organized into 20 classes, from which, 20, 50, and 100 topics are identified. On the other hand, the *Reuters* Corpus has fewer documents; 90 classes must be obtained, so it is impossible to extract 100 topics due to the corpus size.

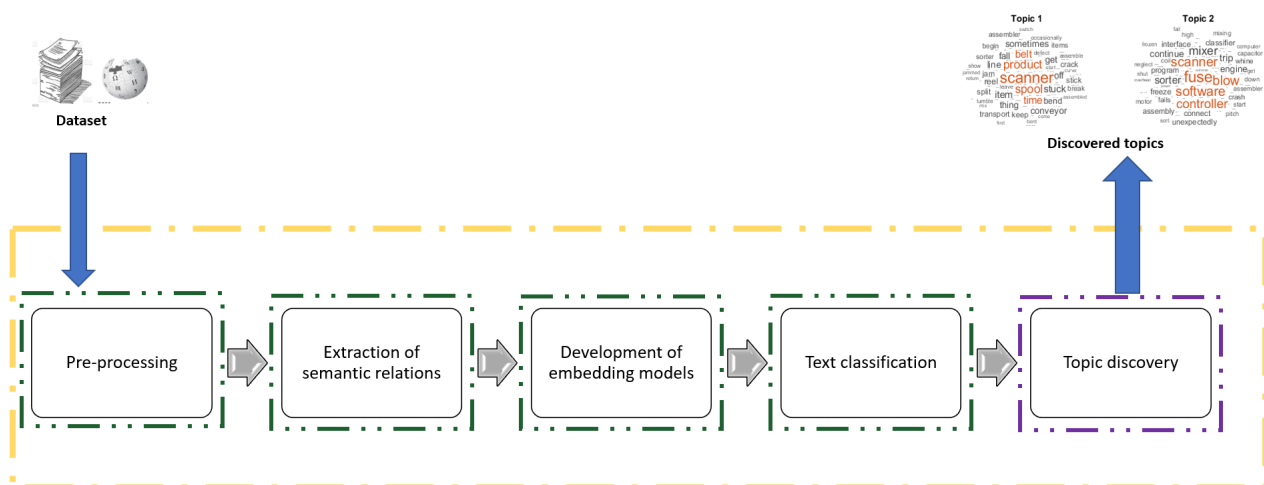


Figure 1. Proposed approach to incorporate text classification into the topic discovery process.

3.1. Text Classification

The text was preprocessed via text cleaning, removing stop words, and converting to lowercase. The text classification process was carried out as proposed in [3]. The method, which includes CNN, was used to assess the performance of three embedding models of semantic relations and to produce a set of classes for each corpus used. The classification task generated the corresponding classes in each corpus used. For the *20-Newsgroups* corpus, 20 classes were obtained, and for the *Reuters* Corpus, 90 classes were obtained. The classes are the basis for topic discovery since they are the input data set to each topic algorithm. In addition, classification was carried out to have an ordered corpus with semantic relationships 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 for topic discovery is the main contribution of this paper.

Semantic Embedding Model

In this paper, the semantic embedding model used in text classification was developed in [3], which involves semantic relationships of synonymy, hyponymy, and hyperonymy.

In [3], the authors presented a novel approach based on relationships extracted from Wikipedia to create embedding models. The creation of embedding models is conditional on the available semantic relations in the text. The process focuses on extracting semantic relationships from an English corpus from Wikipedia. Synonymy, hyponymy, and hyperonymy relationships are extracted with a set of lexical–syntactic patterns from the literature. The relationships are embedded using the procedure proposed by [11] based on matrix factorization. A text classification using CNN was carried out to compare the performance of the relationship-based embeddings and the word-based models, such as *fastText*, *GloVe*, and the WordNet-based model presented in [11]. Therefore, the performance of the semantic embedding model based on three semantic relationships was the best result obtained by [3]. For that reason, in this paper, the semantic embedding model incorporated in text classification for topic discovery is shaped by these three semantic relations.

3.2. Topic Discovery

Topic discovery breaks down a large corpus of text into a small set of interpretable topics, allowing a domain expert to explore and analyze a corpus efficiently [47]. 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 text 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 used for topic discovery. The model is based on the hypothesis that each text contains words or terms from different topics. This model needs to know a priori the text and the number of topics to be found in the text. This model is maintained under the premise that the topics and text are treated through Dirichlet distributions [48]. Figure 2 presents a graphical representation of the LDA model.

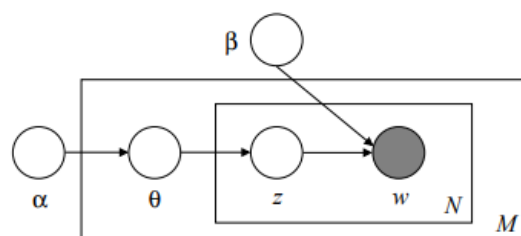


Figure 2. Graphical representation of the LDA model [49].

Where:

M denotes the number of documents.

N denotes the number of words in a given document (document i has N_i words).

α denotes the parameter of the Dirichlet prior to the per-document topic distributions.

β denotes the parameter of the Dirichlet prior to the per-topic word distribution.

θ_i denotes the topic distribution for document i .

φ_k denotes the word distribution for topic k .

z_{ij} denotes the j -th word in document i .

w_{ij} denotes the specific word.

On the other hand, latent semantic analysis (LSA) is another algorithm used for topic discovery. LSA is a mathematical dimensionality reduction procedure denoted as the singular value decomposition (SVD). LSA (or LSI) is an automatic index analysis that projects terms and text in a space of reduced dimensions. The reduction of attributes or dimensions of the text leads to the recovery of the semantics of the original text. The LSA

dimensionality reduction process captures important terms or topics [50]. Figure 3 shows the matrix generated by the LSA model.

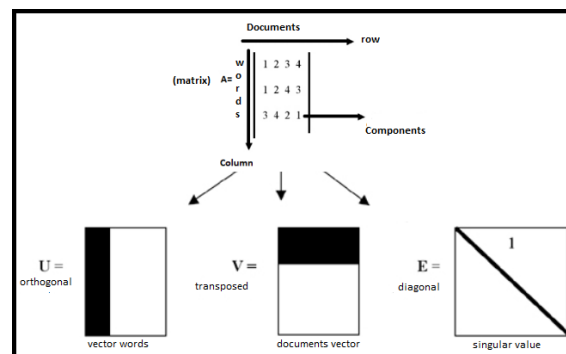


Figure 3. An example of a matrix generated by LSA [51].

A singular value decomposition (SVD) applies to the resulting matrix through a series of elementary linear operations, such as adding and multiplying rows and columns. Therefore, the matrices resulting from applying SVD are as follows:

- Orthogonal matrix (U): obtained by linearly processing the original matrix's number of columns (orthogonal).
- Transposed matrix (V): obtained by swapping the rows with the columns, providing an orthogonal arrangement of the elements of the row.
- Diagonal matrix (E): obtained by linearly processing the original matrix's number of rows, columns, and dimensions (A); the diagonal matrix represents the singular value of (A), and in this, all the elements that do not belong to the diagonal are null or equal to zero.

Finally, probabilistic latent semantic analysis (PLSA) continues the LSA. In PLSA, words are attributed to latent topics or concepts based on the weighted frequency of 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 the 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. In Figure 4, PLSA models the joint probability of seeing a word w and a document (text) d as a mixture of conditionally independent multinomial distributions.

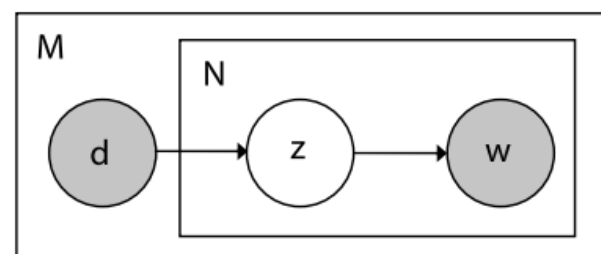


Figure 4. Graphical representation of the PLSA model [52].

Where:

M denotes the number of texts.

N denotes the number of words in a given text.

d indicates a text.

z denotes the latent or hidden variable (topic).

w denotes a specific word.

The evaluation of topic discovery is performed with the normalized topic coherence metric (NPMI) described in Equation (1).

The normalized topic coherence consists of obtaining the *normalized coherence* of each topic (t_i). It 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; this is described below:

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

NPMI scores were then computed from the top- k words for each topic, and lexical probabilities $p(w_i, w_j)$, $p(w_i)$, and $p(w_j)$ were calculated by sampling word counts within a sliding context window over an external corpus, in this case, the English Wikipedia.

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 k top words co-occur within the same paragraph of the set of external text, in this case, the English Wikipedia [53].

4. Results and Discussion

This section presents the results of integrating text classification for topic discovery with semantic embedding models. In addition, the results obtained are compared with those found in the literature and between the datasets used in this work.

4.1. Datasets

A corpus in English from Wikipedia was used as a reference corpus to evaluate the topics. Table 1 shows the text and token numbers for each dataset, i.e., Wikipedia for the evaluation of topics, and *Reuters* (<https://trec.nist.gov/data/reuters/reuters.html>, accessed on 1 May 2020) and *20-Newsgroups* (<http://qwone.com/~jason/20-Newsgroups/>, accessed on 1 May 2020) for topic discovery. The Wikipedia corpus was chosen due to its diverse range of topics, leading to relationships between some words. The *Reuters* and *20-Newsgroups* corpora were chosen because most authors in the literature use these datasets. In addition, the *20-Newsgroups* and *Reuters* corpora are for general purposes.

Table 1. Description of the dataset.

Corpus	Documents	Tokens
Wikipedia	1,000,000	1,560,478,211
<i>20-Newsgroups</i>	20,000	1,800,385
<i>Reuters</i>	18,456	3,435,808

4.2. Experimental Results

The proposed approach was evaluated with the normalized topic coherence metric described in the Equation (1). The corpora used were the classes of *Reuters* and *20-Newsgroups* corpora. Table 2 shows some classes recovered in each corpus.

Table 2. Example of classes obtained in the *20-Newsgroups* and *Reuters* corpora.

Corpus	Class
20-Newsgroups	... <i>Atheism, Sport, Politics, Computing, Cars</i> ...
Reuters	... <i>Aluminium, Barley, Bop, coffee, Cocoa</i> ...

For the topic discovery process, different configurations of parameters and sizes were used. Twenty, fifty, and one hundred topics with the ten top words were identified for the

20-Newsgroups corpus. For the Reuters Corpus, only twenty and fifty topics with the ten top words were identified.

The mean and standard deviations were extracted from the results obtained. The objective was to identify trends in the 20-Newsgroups and Reuters corpora. In this way, it was possible to analyze the results of each corpus.

The number of topics (n), mean (Avg), and standard deviation (std) of the normalized coherence of the topics extracted from each identified class from the 20-Newsgroups and Reuters corpora are presented in Tables 3–5.

Table 3 presents the results of the topics extracted from each class from the 20-Newsgroups and Reuters corpora with the LDA algorithm. The average and standard deviations of the normalized coherence of each identified topic in each class were obtained by extracting 20, 50, and 100 topics from the 20-Newsgroups corpus. The LDA algorithm achieved a highly normalized coherence by extracting 20 topics with 10 top words from the corpus classes. On the other hand, for the Reuters Corpus, only 20 and 50 topics with the ten top words were identified. Also, the LDA algorithm obtained a highly normalized coherence by extracting 20 topics with 10 top words from the corpus classes. However, for each class belonging to the 20-Newsgroups corpus, when 50 and 100 topics were identified with the LDA algorithm, the results did not exceed those obtained when discovering 20 topics with the same algorithm. The same situation occurs with the Reuters corpus in each class when 50 topics were identified.

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

n	20-Newsgroups		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 presents the results obtained when discovering twenty, fifty, and one hundred topics for each class in the 20-Newsgroups corpus. On the other hand, Table 5 shows the results obtained from the twenty and fifty topics identified for each class in the Reuters corpus. The results in both cases were obtained by applying the LSA and PLSA algorithms, respectively. However, they are minor to the results obtained with the LDA algorithm.

Table 4. Average normalized coherence with the LSA algorithm with twenty, fifty, and one hundred topics for the 20-Newsgroups corpus and twenty and fifty topics for the Reuters corpus.

n	20-Newsgroups		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	-	-

In total, 10,200 topics were obtained with the LDA, LSA, and PLSA algorithms for each class in the 20-Newsgroups corpus. For the Reuters corpus, 12,600 topics were obtained from the ninety classes with the LDA, LSA, and PLSA algorithms.

Table 5. Average normalized coherence with the PLSA algorithm with twenty, fifty, and one hundred topics for the *20-Newsgroups* corpus and twenty and fifty topics for the *Reuters* corpus.

<i>n</i>	<i>20-Newsgroups</i>		<i>Reuters</i>	
	<i>Avg</i>	<i>std</i>	<i>Avg</i>	<i>std</i>
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 6 presents classes of the *Reuters* Corpus, and Table 7 shows classes of the *20-Newsgroups* corpus. They only present the three top words obtained with the *LDA*, *LSA*, and *PLSA* algorithms, respectively.

Table 6. Top words of the *Reuters* corpus with the *LDA*, *LSA*, and *PLSA* algorithms with twenty topics and the ten top words.

Class	Topics					
	LDA		LSA		PLSA	
	Company	Disasters	Company	Disasters	Company	Disasters
Aluminum	... operations, dump, ton debris, Panamá, toll dlrs, bank, group debris, Portugal, industries,	... godless, jewish, sabbath ... ,	... group, ... law, love ... ,
	Cultivation	Grain	Cultivation	Grain	Cultivation	Grain
Barley	... acreage, wheat, department acreage, corn, farm maize, wheat, department corn, corn, drum production, system, acres ... ,	... tonnes, february, export ... ,
	Finance	Duty	Finance	Duty	Finance	Duty
Bop	... pressure, country, finance dollar, oil, price singapore, britoil, finance japan, oil, price gas, models, pay battery, wheel, oil ...

Table 7. Top words of the *20-Newsgroups* corpus with the *LDA*, *LSA*, and *PLSA* algorithms, with twenty topics and the ten top words.

Class	Topics					
	LDA		LSA		PLSA	
	Religion	Rituals	Religion	Rituals	Religion	Rituals
Atheism	... town, big, life bible, ceremonial, love goodles, sabbath, ceremonial course, started, hostage,	... godless, jewish, sabbath ... ,	... ceremonial, ... law, love ... ,
	Software	Hardware	Software	Hardware	Software	Hardware
Computing	... virtual, video, file cpu, harddisk, machine software, driver, email circuits, video, pc windows, system, copies ... ,	... mail, acceso, location ... ,
	Elements	Others	Elements	Others	Elements	Others
Cars	... battery, bad, street video, computer, design price, batteries, street video, computer, concrete gas, models, pay software, alert, safety ...

The results obtained with our proposed approach offer insight into integrating classes as input data for each topic discovery algorithm. Although the results are somewhat low,

we hypothesize that the results are relevant by applying additional parameters like the number of epochs in text classification, the number of topics, and the embedding model previously applied. The approach provides consistent topics relevant to the language and domain used. The approach is applied to the LDA, LSA, and PLSA techniques, combining the number of parameters (twenty, fifty, and one hundred) to gauge, respectively, the approach's behavior relative to the number of topics identified.

The results obtained were compared with those existing in the literature; the conclusion is that, in this paper, better results were obtained when considering twenty identified topics with the ten top words.

Table 8 presents the results of different authors in the literature. This work identified topics from each class from which they were extracted. However, even when the authors listed in Table 8 did not discover topics by class, the results obtained by applying the topic coherence metric are higher than those disclosed in this work.

The authors used the normalized topic coherence evaluation metric and the *20-Newsgroups* and/or *Reuters* corpora. However, not all authors applied document classification or clustering algorithms before performing topic discovery. The *Based in* column shows the methods or algorithms applied by the authors in their papers. The results obtained in this paper are higher than those obtained by the authors of [14–18,20]. On the other hand, some authors, such as [18], obtained higher coherence values for the *20-Newsgroups* corpus. The *Reuters* corpus [19,27] obtained higher coherence values than the results obtained in this work. In [18,27], the results are significant to those obtained in this paper because they applied algorithms and methods, such as variational autoencoder community detection and community mining. We deem the methods mentioned previously as beneficial to the authors' results. In this paper, the objective was to integrate text classification into topic discovery. Hence, no additional method was contemplated.

The proposed approach obtained a coherence of 0.1723 for the *20-Newsgroups* corpus using the LDA algorithm with 20 topics, and 0.1441 for the *Reuters* Corpus with the LDA algorithm with 20 topics.

It is evident that with 20 extracted topics, optimal topic coherent results were obtained. The *20-Newsgroups* corpus has 20,000 documents, and it was necessary to obtain 20 classes, which allowed experiments to extract 20, 50, and 100 topics, resulting in specific topics and coherent results. On the other hand, the *Reuters* Corpus has a smaller number of documents, and 90 classes must be extracted. The dispersion will affect the coherent results by adding repeated words to each topic. On the other hand, the LDA algorithm weighs the results better by assigning a higher value to a word; therefore, this algorithm has better results.

Table 8. Comparison of the average results obtained with normalized topic coherence for both corpora.

Author	20-Newsgroups	Reuters	Based in
[14]	0.10	0.04	Min-hashing
[15]	0.103	0.152	LDA
[16]	0.39	-	Hyperbolic embeddings
[20]	0.28	-	Neural Variational Document Model
[17]	0.102	-	Contextualized embeddings
[18]	0.042	-	Community detection algorithm
[18]	0.279	-	Community detection algorithm
[19]	0.044	0.182	Co-occurrence Networks
[27]	0.17	0.18	Principal Component Analysis, BERTopic
This work	0.172	0.144	LDA, LSA, PLSA

5. Conclusions and Future Work

This paper presents an approach that integrates text classification with the task of discovering specific topics from large amounts of classified texts in English. The approach

starts by integrating automatic text classification before applying a topic discovery process to obtain specific topics with a semantic relationship between the primary words for each class. The text classification process analyzes the words via semantic relationship embedding—which makes up each document—to decide which class it identifies. The proposed integration provides latent and specific topics of each class represented by top words.

The results are compared with the literature, demonstrating that integrating text classification into the process of discovering specific topics in each class generates significant relationships within the comprising texts.

The main contribution of this work lies in the integration of classification with the discovery of specific topics. We compare the results obtained on the discovery of topics with works in the literature. The results show the importance of discovering specific topics in each identified class. The approach is a valuable resource for natural language, demonstrating that adding semantics to a classification process will yield favorable results with high coherence.

The identified specific topics, together with their top words, help data analysts organize the latent themes present in the text with a specific vision of each corpus class.

In future work, the objective will be to label the identified topics using an ontology of the same domain, and then compare the results obtained from both experiments.

Author Contributions: Conceptualization, M.T.V. and J.A.R.-O.; methodology, A.L.L.-S., M.T.V. and J.A.R.-O.; software, A.L.L.-S., M.T.V. and J.A.R.-O.; validation, M.T.V. and J.A.R.-O.; investigation, A.L.L.-S. and J.A.R.-O.; resources, J.A.R.-O.; data curation, A.L.L.-S.; writing—original draft, A.L.L.-S.; writing—review & editing, M.T.V. and J.A.R.-O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Universidad Autonoma Metropolitana, Azcapotzalco. The present work was funded by the research project SI001-18 at UAM Azcapotzalco and by the Consejo Nacional de Humanidades de Ciencia y Tecnologia (CONAHCYT) with scholarship number 788155. The authors thankfully acknowledge the computer resources, technical advice, and support provided by Laboratorio Nacional de Supercómputo del Sureste de México (LNS), a member of the CONAHCYT national laboratories, with project No 202103090C. The authors would like to thank Instituto Nacional de Astrofísica, Óptica y Electrónica for the computer resources, technical advice, and support provided by the Laboratorio Nacional de Supercómputo, a deep learning platform for language technologies and by project VIEP 2023 at BUAP.

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

References

1. 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: Tandil, Argentina, 2016.
2. López López, A. *Descubrimiento de Tópicos a Partir de Textos en Español Sobre Enfermedades en México*; Universidad Autonoma Metropolitana: Ciudad de Mexico, Mexico, 2022.
3. 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. [[CrossRef](#)]
4. Orkphol, K.; Yang, W. Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet. *Future Internet* **2019**, *11*, 114. [[CrossRef](#)]
5. Zhou, Z.; Fu, B.; Qiu, H.; Zhang, Y.; Liu, X. Modeling medical texts for distributed representations based on Skip-Gram model. In Proceedings of the 2017 3rd International Conference on Information Management (ICIM), Chengdu, China, 21–23 April 2017; pp. 279–283.
6. 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.

7. Church, K.W.; Kordoni, V. Emerging trends: SOTA-chasing. *Nat. Lang. Eng.* **2022**, *28*, 249–269. [[CrossRef](#)]
8. Athiwaratkun, B.; Wilson, A.G.; Anandkumar, A. Probabilistic fasttext for multi-sense word embeddings. *arXiv* **2018**, arXiv:1806.02901.
9. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
10. Vásquez, A.C.; Quispe, J.P.; Huayna, A.M. Procesamiento de lenguaje natural. *Rev. Investig. Sist. Inform.* **2009**, *6*, 45–54.
11. Saedi, C.; Branco, A.; Rodrigues, J.; Silva, J. Wordnet embeddings. In Proceedings of the Third Workshop on Representation Learning for NLP, Melbourne, Australia, 20 July 2018; pp. 122–131.
12. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. *Information* **2019**, *10*, 150. [[CrossRef](#)]
13. Lezama Sánchez, A.L.; Tovar Vidal, M.; Reyes Ortiz, J.A. A Behavior Analysis of the Impact of Semantic Relationships on Topic Discovery. *Comput. Sist.* **2022**, *26*, 149–160. [[CrossRef](#)]
14. Fuentes-Pineda, G.; Meza-Ruiz, I.V. Topic discovery in massive text corpora based on min-hashing. *Expert Syst. Appl.* **2019**, *136*, 62–72. [[CrossRef](#)]
15. 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. In Proceedings of the ICLR, Kigali, Rwanda, 1–5 May 2023; p. 16.
16. Xu, Y.; Wang, D.; Chen, B.; Lu, R.; Duan, Z.; Zhou, M. HyperMiner: Topic Taxonomy Mining with Hyperbolic Embedding. *arXiv* **2022**, arXiv:2210.10625.
17. Bianchi, F.; Terragni, S.; Hovy, D. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv* **2020**, arXiv:2004.03974.
18. Jin, Y.; Zhao, H.; Liu, M.; Du, L.; Buntine, W. Neural attention-aware hierarchical topic model. *arXiv* **2021**, arXiv:2110.07161.
19. Austin, E.; Trabelsi, A.; Langeron, C.; Zaïane, O.R. Hierarchical Topic Model Inference by Community Discovery on Word Co-occurrence Networks. In Proceedings of the Data Mining: 20th Australasian Conference, AusDM 2022, Western Sydney, Australia, 12–15 December 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 148–162.
20. Ding, R.; Nallapati, R.; Xiang, B. Coherence-aware neural topic modeling. *arXiv* **2018**, arXiv:1809.02687.
21. Poon, L.K.; Zhang, N.L.; Xie, H.; Cheng, G. Handling collocations in hierarchical latent tree analysis for topic modeling. *arXiv* **2020**, arXiv:2007.05163.
22. 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 J. Biomed. Health Inform.* **2020**, *24*, 2733–2742. [[CrossRef](#)] [[PubMed](#)]
23. 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.
24. Zuo, Y.; Li, C.; Lin, H.; Wu, J. Topic modeling of short texts: A pseudo-document view with word embedding enhancement. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 972–985. [[CrossRef](#)]
25. Romanova, A. Semantics graph mining for topic discovery and word associations. *Int. J. Data Mining Knowl. Manag. Process (IJDKP)* **2021**, *10*, 1–14. [[CrossRef](#)]
26. Stanik, C.; Pietz, T.; Maalej, W. Unsupervised topic discovery in user comments. In Proceedings of the 2021 IEEE 29th International Requirements Engineering Conference (RE), Notre Dame, IN, USA, 20–24 September 2021; pp. 150–161.
27. Wang, D.; Xu, Y.; Li, M.; Duan, Z.; Wang, C.; Chen, B.; Zhou, M. Knowledge-aware Bayesian deep topic model. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 14331–14344.
28. Ogunleye, B.; Maswera, T.; Hirsch, L.; Gaudoin, J.; Brunson, T. Comparison of Topic Modelling Approaches in the Banking Context. *Appl. Sci.* **2023**, *13*, 797. [[CrossRef](#)]
29. 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. *Appl. Sci.* **2021**, *11*, 11897. [[CrossRef](#)]
30. Huang, L.; Dou, Z.; Hu, Y.; Huang, R. Textual analysis for online reviews: A polymerization topic sentiment model. *IEEE access* **2019**, *7*, 91940–91945. [[CrossRef](#)]
31. Shafqat, W. A Hybrid Approach for Topic Discovery and Recommendations Based on Topic Modeling and Deep Learning. Ph.D. Thesis, Graduate School of Jeju University, Jeju, Republic of Korea, 2020.
32. Pandey, C. redBERT: A topic discovery and deep sentiment classification model on COVID-19 online discussions using BERT NLP model. *Int. J. Open Source Softw. Process.* **2021**, *12*, 32–47. [[CrossRef](#)]
33. Huang, L.; Dou, Z.; Hu, Y.; Huang, R. Online sales prediction: An analysis with dependency scor-topic sentiment model. *IEEE Access* **2019**, *7*, 79791–79797. [[CrossRef](#)]
34. Yousef, M.; Voskergian, D. TextNetTopics: Text classification based word grouping as topics and topics' scoring. *Front. Genet.* **2022**, *13*, 893378. [[CrossRef](#)] [[PubMed](#)]
35. Rijcken, E.; Kaymak, U.; Scheepers, F.; Mosteiro, P.; Zervanou, K.; Spruit, M. Topic modeling for interpretable text classification from EHRs. *Front. Big Data* **2022**, *5*, 846930. [[CrossRef](#)]
36. Murshed, B.A.H.; Abawajy, J.; Mallappa, S.; Saif, M.A.N.; Al-Ghuribi, S.M.; Ghanem, F.A. Enhancing Big Social Media Data Quality for Use in Short-Text Topic Modeling. *IEEE Access* **2022**, *10*, 105328–105351. [[CrossRef](#)]
37. Chandran, N.V.; Anoop, V.; Asharaf, S. Topicstriker: A topic kernels-powered approach for text classification. *Results Eng.* **2023**, *17*, 100949. [[CrossRef](#)]

38. Kaur, A.; Singh, B.; Nandi, B.P.; Jain, A.; Tayal, D.K. Enhancing Topic Prediction Using Machine Learning Techniques and ConceptNet-based Cosine Similarity Measure. 2023. Available online: https://assets.researchsquare.com/files/rs-3172758/v1_covered_b085adba-6dc9-4b33-9c28-aa72287bc4f8.pdf?c=1689778843 (accessed on 19 July 2023).
39. Moro, A.; Navigli, R. Integrating syntactic and semantic analysis into the open information extraction paradigm. In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013.
40. Guarasci, R.; Damiano, E.; Minutolo, A.; Esposito, M. When Lexicon-Grammar Meets Open Information Extraction: A Computational Experiment for Italian Sentences. In Proceedings of the CLiC-it, Bari, Italy, 13–15 November 2019.
41. Guarasci, R.; Damiano, E.; Minutolo, A.; Esposito, M.; De Pietro, G. Lexicon-grammar based open information extraction from natural language sentences in Italian. *Expert Syst. Appl.* **2020**, *143*, 112954. [CrossRef]
42. Ro, Y.; Lee, Y.; Kang, P. Multi²OIE: Multilingual Open Information Extraction Based on Multi-Head Attention with BERT. *arXiv* **2020**, arXiv:2009.08128.
43. Zhou, S.; Yu, B.; Sun, A.; Long, C.; Li, J.; Yu, H.; Sun, J.; Li, Y. A survey on neural open information extraction: Current status and future directions. *arXiv* **2022**, arXiv:2205.11725.
44. Catelli, R.; Pelosi, S.; Comito, C.; Pizzuti, C.; Esposito, M. Lexicon-based sentiment analysis to detect opinions and attitude towards COVID-19 vaccines on Twitter in Italy. *Comput. Biol. Med.* **2023**, *158*, 106876. [CrossRef]
45. Guo, Y.; Zhang, Y.; Lyu, T.; Prospero, M.; Wang, F.; Xu, H.; Bian, J. The application of artificial intelligence and data integration in COVID-19 studies: A scoping review. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 2050–2067. [CrossRef] [PubMed]
46. Bovi, C.D.; Telesca, L.; Navigli, R. Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Trans. Assoc. Comput. Linguist.* **2015**, *3*, 529–543. [CrossRef]
47. May, C.C. Topic Modeling in Theory and Practice. Ph.D. Thesis, Johns Hopkins University, Baltimore, MD, USA, 2022.
48. Valero Moreno, A.I. *Técnicas estadísticas en Minería de Textos*; Universidad de Sevilla: Sevilla, Spain, 2017.
49. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
50. Venegas, R. La similitud léxico-semántica en artículos de investigación científica en español: Una aproximación desde el Análisis Semántico Latente. *Rev. Signos* **2006**, *39*, 75–106.
51. Torres López, C. Segmentación y Detección de Tópicos Enfocado a la Minería de Opinión. Ph.D. Thesis, Universidad Central “Marta Abreu” de Las Villas, Santa Clara, Cuba, 2016.
52. Niebles, J.C.; Wang, H.; Fei-Fei, L. Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vis.* **2008**, *79*, 299–318. [CrossRef]
53. Wales, J.; Sanger, L. Available online: <https://dumps.wikimedia.org/enwiki/20230101/> (accessed on 15 January 2001).

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.