Hybrid Method for Short Text Topic Modeling

Jinyuan Chen and Bela Stantic(✉)

School of Information and Communication Technology, Griffith University,
Brisbane, Australia
Jinyuan.chen@griffithuni.edu.au, B.Stantic@griffith.edu.au

Abstract. The rise in social media's popularity has led to a significant increase in user-generated content across various topics. Extracting information from these data can be valuable for different domains, however, due to the nature of the vast volume it is not possible to extract information manually. Different aspects of information extraction have been introduced in literature including identifying what topic is discussed in the text. The challenge becomes even bigger when the text is short, such as found in social media. Various methods for topic modeling have been proposed in the literature that could be generally categorized as unsupervised and supervised learning. However, unsupervised topic modeling methods have some shortcomings, such as semantic loss and poor explanation, and are sensitive to the choice of parameters, such as the number of topics. While supervised machine learning methods based on deep learning can achieve high accuracy they need data annotated by humans, which is time-consuming and costly. To overcome the above mentioned disadvantages this work proposes a hybrid topic modeling method that combines the advantages of both unsupervised and supervised methods. We built a hybrid model by combining Latent Dirichlet Allocation (LDA) and deep learning built on top of the Bidirectional Encoder Representations from the Transformers (BERT) model. LDA is used to identify the optimal number of topics and topic-relevant keywords where the only need for human input, with the aid of ChatGPT, is to identify associated topics based on topic-specific keywords. This annotation is used to train and fine-tune the BERT model. In the experimental evaluation of posts related to climate change, we show that the proposed concept is applicable for predicting topics from short text without the need for lengthy and costly annotation.

Keywords: Topic Modeling · LDA · Deep learning · BERT · ChatGPT

1 Introduction

Social media have provided means for people to share their opinions and observations about different aspects and information deeply hidden in these data can

N. T. Nguyen et al. (Eds.): ACIIDS 2023, CCIS 1863, pp. 157–168, 2023.
https://doi.org/10.1007/978-3-031-42430-4_13

be valuable for different stakeholders. However, due to the volume and complexity of social media data it is impossible and impractical to analyze all of the data manually. Therefore, it is necessary to use methods to extract the information. Identifying the topic in the text is one of the valuable aspects that can be extracted from the text and therefore topic modeling attracted significant attention in the literature, both related to natural language processing and machine learning. It is able to scan documents and, based on word patterns, identify word groups and similar expressions that best characterize a set of documents. The goal of topic modeling is to discover the hidden themes or topics present in a corpus of documents.

Broadly speaking topic modeling algorithms can be categorized into two main groups depending on the method applied: *Unsupervised learning* and *Supervised learning*. When trained on large, high-quality labeled datasets, supervised methods can achieve high accuracy on various tasks [6]. But obtaining labeled data can be time-consuming and expensive, especially when dealing with extensive collections of texts. While unsupervised learning algorithms do not require labeled data, which is making them more scalable and cost-effective, they can be inaccurate and identified topics are general not specific to a certain issue. In addition, unsupervised methods are sensitive to the choice of parameters including the optimal number of topics. The most popular techniques for topic modeling are different variations of Latent Dirichlet Allocation (LDA), an unsupervised learning method.

In this work, to overcome the limitations of both unsupervised and supervised methods we are proposing a hybrid method for topic modeling which combines supervised and unsupervised learning. We first preprocessed the data we collected from Twitter. Then, we identified the optimal number of topics and generated associated keywords for the optimal number of topics. Human involvement in the process of annotation was minimal, it was related to identifying the associated topics based on topic-specific keywords, with the aid of ChatGPT, and in this case, it took less than 30 min. This annotation is used to train and fine-tune the BERT model, which can be used for identifying topics and with all benefits of supervised methods. The remainder of the paper is organized as follows, in the next section considering we harness both unsupervised and supervised method topic modeling we elaborate on both. In Sect. 3 we present the proposed hybrid method for Topic modeling and in individual subsection we present steps of the proposed framework to be performed. Finally, in Sect. 4 we conclude the paper and suggest possible avenues for future work.

2 Literature Review

Topic modeling is a technique that can automatically analyze text and identify the underlying topic discussed in the text and cluster documents with the same topics. Various methods for topic modeling have been proposed in the literature that could be generally categorized as unsupervised or supervised learning methods. In the following subsections we will elaborate on topic modeling methods proposed in the literature.

2.1 Unsupervised Learning Methods

One of the first algorithms to implement topic modeling dates back to the 1990s. The early approaches to cluster and categorize large text datasets to topic modeling were based on Latent Semantic Analysis (LSA), also known as latent semantic indexing (LSI) [12]. Implementing LSA to identify latent document structures is based on Singular Value Decomposition (SVD) that captures a text corpus's underlying semantic structure by building a matrix of term-document frequency counts [17].

As for text classification and information retrieval, LSA is still an algorithm worth considering, but the topic modelling works implemented by LSA now have been largely replaced by other techniques such as Latent Dirichlet Allocation (LDA) [5]. LDA is a three-level hierarchical Bayesian model Fig. 1 [18], in which each item of a collection is modelled as a finite mixture over an underlying set of topics. Each topic is, in turn, modelled as an infinite mixture over an underlying set of topic probabilities. The output of the LDA model is a series of topics. Then the represented topic will be classified by a set of representative words, which can be used to label topics. For example, a topic distribution that is heavily weighted to terms like "soccer", "teammates", "score", and "goal" may be associated with the "sports" topic. Where M is the number of documents, N represents the number of words in a given document. The distribution of the 'θ' and 'β' represent the multinomial distribution. The parameters of the multinomial distribution are 'α' and 'ϕ'. Every row of data in 'θ' is a K-dimensional vector representing K^{th} topics in the corpus. 'Z' is one of a topic from 'θ'. while 'W' is the corresponding word in the topic 'Z'.

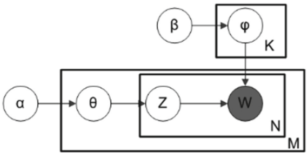


Fig. 1. The structure of Latent Dirichlet Allocation

Because of the simplicity and advantages of LDA, many models based on LDA, to suit different data, have been developed [1]. For instance, the method named *Sparse Topic-Latent Dirichlet Allocation* was the first LDA extension for unsupervised topic modeling without hierarchy regression [15]. In addition, several other models based on LDA have been proposed, for example, a Nonparametric Bayesian Model [3] and the correlated Topic Model [4].

2.2 Supervised Learning Methods

Supervised topic modeling, involves using the labeled dataset to train a model that could classify new unseen documents into topics. These methods require humans to annotate documents into corresponding topics. The topic model is trained with annotated data to categorize new documents accurately. Researchers have developed various supervised topic modeling methods for topic modeling. One such method is the supervised LDA algorithm (sLDA) initially proposed by [15]. This method uses labeled training data to train the model that is optimized for a specific task or application [7]. sLDA could be used with regression to multi-class classification [16], the differences between LDA and sLDA are displayed below in Fig. 2. For each D, the N_D words are generated by drawing a topic t from the distribution of documents and topics θ and then giving a word w from the topic-word distribution φ. The usability of sLDA is demonstrated in [9] where the intention was to check the key attributes influencing Airbnb user satisfaction and dissatisfaction by analyzing online reviews. Authors used LDA to extract positive and negative topics from reviews and combined the statistical results of sLDA to discover topics related to the satisfaction of Airbnb users.

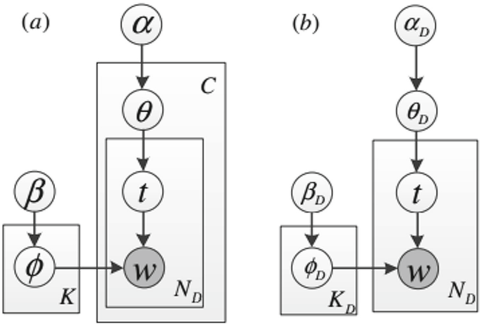


Fig. 2. Structure of sLDA (b) compared with LDA (a), [13]

There are also proposals that combine multiple methods and connecting individuals' strengths to offset the limitations of individual methods and therefore produce more accurate results than individual topic modeling methods. For example, an approach may combine probabilistic topic modeling methods like Latent Dirichlet Allocation (LDA) with clustering methods or word embedding techniques such as *Word2Vec* to enhance the accuracy and interpretability of the extracted topics. One such algorithm is the LDA-W2V method [11]. In addition, Topic Attention Model (TAM) for topic modeling combines a supervised recurrent neural network (RNN) with LDA [19]. TAM has two inputs, one of them is a sequence model and another is a bag-of-words topic model. So, the whole vocabulary is input for RNN, and the word embedding matrix was initialized by the word embedding learned from Word2Vec. In experimental evaluation authors demonstrated the dominance of the TAM method when compared to different unsupervised and supervised methods being applied individually.

3 Methodology

In this section, we present the methodology proposed to harness unsupervised learning as an annotation for supervised learning. The Framework and steps are shown in Fig. 3 and they are listed below:

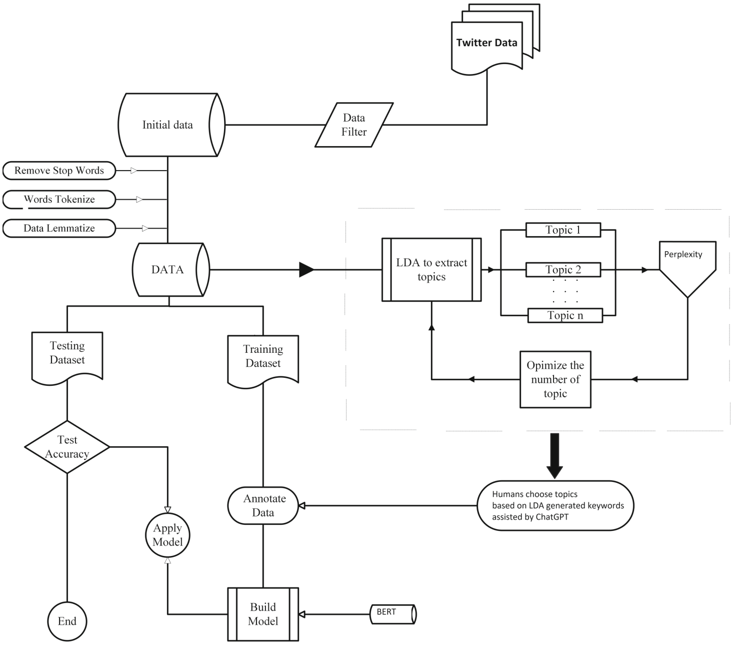


Fig. 3. Framework of Hybrid Model for Topic Modeling

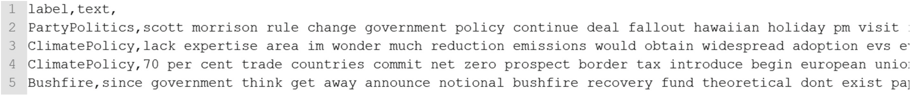
- Data Collection: Collect Twitter posts relevant to climate change,
- Data Preparation: The first step task is to prepare the data, which involves preprocessing the text data, including cleaning, tokenization, and stop-word removal. In supervised topic modeling, the data preparation step is similar but also involves labeling the data with the corresponding topic or category.
- Unsupervised Topic Modeling: Dirichlet Allocation (LDA) is used to identify topics. These techniques generate a set of topics, each represented by a set of words that are highly correlated with each other.
- Identifying the optimal number of Topics.
- Human naming of topics based on the prevalence of words in individual topics. ChatGPT[1] can assist in identifying the most likely topic based on a set of words produced by LDA for an individual topic.

[1] https://openai.com/blog/chatgpt.

- Annotation: Assigning a label to the training dataset for each individual post, as identified by LDA and named by Humans assisted by ChatGPT.
- Building the model: based on the annotated training dataset. We harness BERT embedding with deep learning models for classification.
- Evaluation: apply the trained model to the test set and predict the associated topic of each individual post.

3.1 Data Collection

Social media have become a powerful tool for focusing on various topics and events. On social media, texts are often required to be short. Therefore, this work we will rely on social media and use Twitter as the posts are publicly available. The rise of social media platforms has provided a space for users to share their views on different topics. Many studies relied on this source of data and demonstrated their usefulness, for example, tracing the rise of ‘flightshame’ in social media and analysis of the climate crisis and flying [2]. Due to the significance of climate change in this work we opted to consider climate change related social media posts.



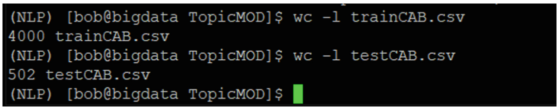
```
label,text,
PartyPolitics,scott morrison rule change government policy continue deal fallout hawaiian holiday pm visit
ClimatePolicy,lack expertise area im wonder much reduction emissions would obtain widespread adoption evs e
ClimatePolicy,70 per cent trade countries commit net zero prospect border tax introduce begin european unio
Bushfire,since government think get away announce notional bushfire recovery fund theoretical dont exist pa
```

Fig. 4. The structure of the training set

To access public Twitter data we relied on an academic-level Twitter application programming interface (API). The API has the option to filter posts by time, keywords, or a geographic area, on this occasion. Considering that we intended to look into opinions about climate change from the Australian public we applied geographic filter in the form ‘location = “-26.117995,134.300207,2200km”’. In addition, we also apply keyword filtering and to avoid the sparsity of data, we selected several diverse keywords such as “Co2”, “Climate change”, and “emission”. Because Twitter data is in JavaScript Object Notation (JSON) format, we stored the collected posts in the *MongoDB* database located in an in-house Big Data cluster. Out of the collected posts, considering that we were interested in topic modeling we randomly selected 4,502 posts of length more than 150 characters (Twitter has a restriction of 280 characters) and split the data set into 4,000 for training sets (the structure is shown in Fig. 4) and 502 for the test set, Fig. 5.

3.2 Data Pre-processing

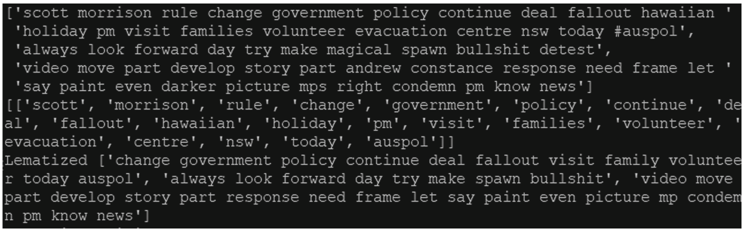
Preprocessing is an important step in preparing text data for topic modeling. It is the process of cleaning and transforming original text data into a format



```
(NLP) [bob@bigdata TopicMOD]$ wc -l trainCAB.csv
4000 trainCAB.csv
(NLP) [bob@bigdata TopicMOD]$ wc -l testCAB.csv
502 testCAB.csv
(NLP) [bob@bigdata TopicMOD]$ █
```

Fig. 5. The number of records in training and test sets

suitable for further processing. In our dataset, the unstructured and noisy nature of Twitter data presents unique challenges for topic modeling and deep learning. To clean the text we have removed punctuations, special characters, Twitter handles, emojis, images, and URLs. Before performing tokenization, to reduce the number of tokens, we also performed lemmatization (grouping together different forms of the same word), and in addition, converted all letters to lowercase Fig. 6.

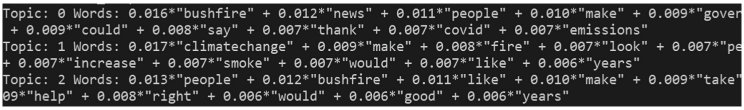


```
['scott morrison rule change government policy continue deal fallout hawaiian '
 'holiday pm visit families volunteer evacuation centre nsw today #auspol',
 'always look forward day try make magical spawn bullshit detest',
 'video move part develop story part andrew constance response need frame let '
 'say paint even darker picture mps right condemn pm know news']
[['scott', 'morrison', 'rule', 'change', 'government', 'policy', 'continue', 'de
al', 'fallout', 'hawaiian', 'holiday', 'pm', 'visit', 'families', 'volunteer', '
evacuation', 'centre', 'nsw', 'today', 'auspol']]
Lematized ['change government policy continue deal fallout visit family voluntee
r today auspol', 'always look forward day try make spawn bullshit', 'video move
part develop story part response need frame let say paint even picture mp condem
n pm know news']
```

Fig. 6. Sample preprocessed, tokenized, and Lemmatized data

3.3 LDA

We used *Gensim*[2] to generate the LDA model. Gensim is an open-source Python library for natural language processing, specifically for topic modeling, document similarity analysis, and text summarization. Sample output from LDA-generated topics by Gensim is shown in Fig. 7.



```
Topic: 0 Words: 0.016*"bushfire" + 0.012*"news" + 0.011*"people" + 0.010*"make" + 0.009*"gover
 + 0.009*"could" + 0.008*"say" + 0.007*"thank" + 0.007*"covid" + 0.007*"emissions"
Topic: 1 Words: 0.017*"climatechange" + 0.009*"make" + 0.008*"fire" + 0.007*"look" + 0.007*"pe
+ 0.007*"increase" + 0.007*"smoke" + 0.007*"would" + 0.007*"like" + 0.006*"years"
Topic: 2 Words: 0.013*"people" + 0.012*"bushfire" + 0.011*"like" + 0.010*"make" + 0.009*"take"
09*"help" + 0.008*"right" + 0.006*"would" + 0.006*"good" + 0.006*"years"
```

Fig. 7. LDA Topic Distribution

The experiments were run on a CPU-based server (Intel(R) Xeon(R) CPU E5-2609 v3 @ 1.90 GHz, 12-Core Processor, 65 GB Memory), enhanced by a GPU (GeForce GTX 1080 with 8 GB of memory).

[2] https://pypi.org/project/gensim/.

3.4 Optimal Number of Topics

Usually, in order to evaluate works, 'Recall, Precision, and F-score' [10] are designed to be the metrics of classification tasks. In comparison, Perplexity is a vital metric to evaluate a language model. It is used to evaluate the predictive power of a language model, that is, whether the model can assign sequences in the test set to a distribution similar to the training set. Perplexity is a measure of how well a probability distribution or probability model predicts a sample (similar to entropy), the smaller value indicates the better performance of the model evaluation, it is shown in the formula below.

$$Perplexity = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 P(w_i)}$$

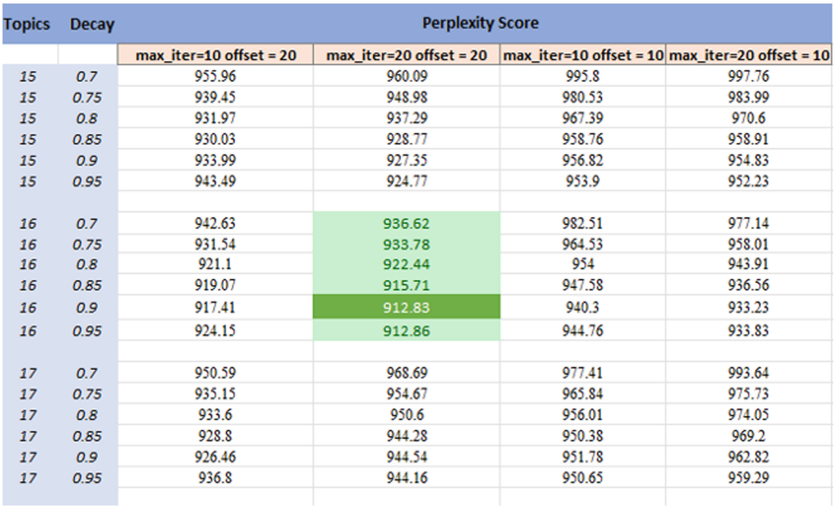
where *N represents the total number of distinct words.* $P(w_i)$ *means the probability of the* i^{th} *word from documents.* To find the optimal number of topics we calculated perplexity for all options with the number of topics between 5 and 25. While testing different numbers of topics, we also looked into different values for learning_decay, learning_offset, and max_iter and performed experiments with all possible combinations. The *max_iter* is the maximum number of passes over the training data (aka epochs), *learning_decay* is a parameter that controls learning rate in the online learning method and to guarantee asymptotic convergence the value should be set between (0.5, 1.0]. The *learning_offset* (also called tau_0)is a parameter that down weights early iterations in online learning, it should be greater than 1.0. For topics, as mentioned earlier we considered the range from 5 to 25. From previous work, we concluded that for learning_decay the value should be between 0.7 to 0.95, and we tested all values with the step of 0.05. Similarly, we considered for *max_iter* and *learning_offset* values 10 and 20, shown in Fig. 8. We have noticed that the enforcing to calculate perplexity for every item did not improve the performance but just slowed down the experiment significantly. Therefore, we left the parameter *evaluate_every=-1*, which indicates to not evaluate perplexity for every item.

From the experiment results, where we tested all possible combinations of topics, *learning_decay*, *max_iter*, and *learning_offset* we found that the lowest perplexity is 912.86 when the number of topics is 16, *learning_decay* is 0.95, and both max_iter and *learning_offset* are equal to 20 (As it is shown and highlighted in Fig. 8). Therefore, we applied these values when generating keywords for 16 topics. We also identified that 12 keywords are sufficient to define the topics.

3.5 ChatGPT Annotation

The only involvement of humans in the annotation is to assign associated topic names for 16 topics produced by LDA. This can be assisted, and we also relied on ChatGTP 3.5[3] to identify associated topic names based on topic-specific keywords and to possibly cluster more LDA-identified topics into one, as it can be seen in Fig. 9. In our case, 16 LDA-generated topics were clustered into

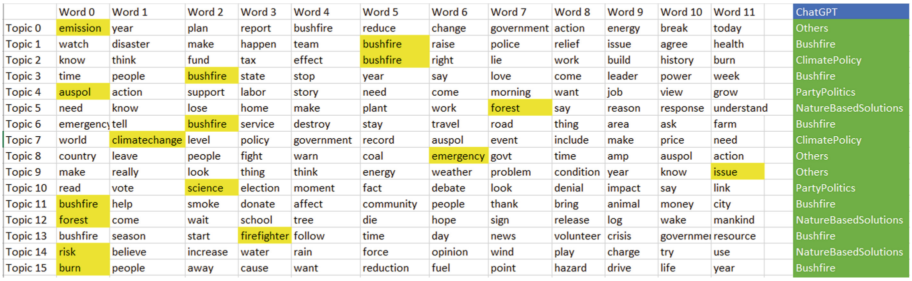
[3] https://chat.openai.com/.



Topics	Decay	Perplexity Score			
		max_iter=10 offset = 20	max_iter=20 offset = 20	max_iter=10 offset = 10	max_iter=20 offset = 10
15	0.7	955.96	960.09	995.8	997.76
15	0.75	939.45	948.98	980.53	983.99
15	0.8	931.97	937.29	967.39	970.6
15	0.85	930.03	928.77	958.76	958.91
15	0.9	933.99	927.35	956.82	954.83
15	0.95	943.49	924.77	953.9	952.23
16	0.7	942.63	936.62	982.51	977.14
16	0.75	931.54	933.78	964.53	958.01
16	0.8	921.1	922.44	954	943.91
16	0.85	919.07	915.71	947.58	936.56
16	0.9	917.41	912.83	940.3	933.23
16	0.95	924.15	912.86	944.76	933.83
17	0.7	950.59	968.69	977.41	993.64
17	0.75	935.15	954.67	965.84	975.73
17	0.8	933.6	950.6	956.01	974.05
17	0.85	928.8	944.28	950.38	969.2
17	0.9	926.46	944.54	951.78	962.82
17	0.95	936.8	944.16	950.65	959.29

Fig. 8. Parameters for optimal number of topics based on lowest perplexity

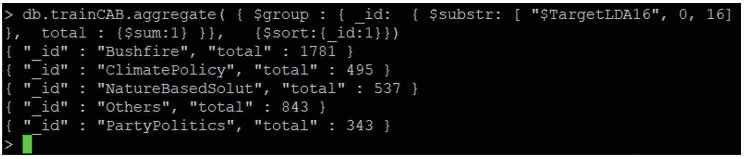
5 categories and named appropriately based on keywords. It is important to mention that this task took only minimal effort and ensured that all training data (4000 posts) for supervised learning were labeled. In contrast, a similar task of annotation performed fully by humans in our previous project required 4,000 min (one minute per tweet) for annotation (about 66 h).



	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11	ChatGPT
Topic 0	emission	year	plan	report	bushfire	reduce	change	government	action	energy	break	today	Others
Topic 1	watch	disaster	make	happen	team	bushfire	raise	police	relief	issue	agree	health	Bushfire
Topic 2	know	think	fund	tax	effect	bushfire	right	lie	work	build	history	burn	ClimatePolicy
Topic 3	time	people	bushfire	state	stop	year	say	love	come	leader	power	week	Bushfire
Topic 4	auspol	action	support	labor	story	need	come	morning	want	job	view	grow	PartyPolitics
Topic 5	need	know	lose	home	make	plant	work	forest	say	reason	response	understand	NatureBasedSolutions
Topic 6	emergency	tell	bushfire	service	destroy	stay	travel	road	thing	area	ask	farm	Bushfire
Topic 7	world	climatechange	level	policy	government	record	auspol	event	include	make	price	need	ClimatePolicy
Topic 8	country	leave	people	fight	warn	coal	emergency	govt	time	amp	auspol	action	Others
Topic 9	make	really	look	thing	think	energy	weather	problem	condition	year	know	issue	Others
Topic 10	read	vote	science	election	moment	fact	debate	look	denial	impact	say	link	PartyPolitics
Topic 11	bushfire	help	smoke	donate	affect	community	people	thank	bring	animal	money	city	Bushfire
Topic 12	forest	come	wait	school	tree	die	hope	sign	release	log	wake	mankind	NatureBasedSolutions
Topic 13	bushfire	season	start	firefighter	follow	time	day	news	volunteer	crisis	governmen	resource	Bushfire
Topic 14	risk	believe	increase	water	rain	force	opinion	wind	play	charge	try	use	NatureBasedSolutions
Topic 15	burn	people	away	cause	want	reduction	fuel	point	hazard	drive	life	year	Bushfire

Fig. 9. Suggestion by ChatGPT based on LDA-generated words and dominant keywords

After identifying suitable topics, based on keywords and assisted by ChatGPT, these values were updated in the MongoDB database and resulted in the distribution shown in Fig. 10. It can be seen that most topics associate with 'Bushfire', followed by 'Others, and 'NatureBasedSolutions'. This is in line with annotation performed by humans on the same data set.



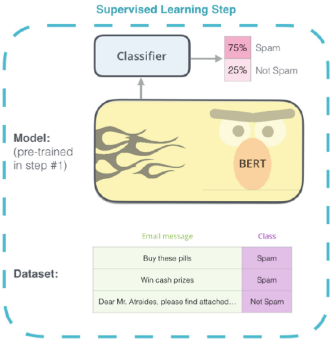
```
> db.trainCAB.aggregate( { $group : { _id:  { $substr: [ "$TargetLDA16", 0, 16]
},  total : {$sum:1} }},   {$sort:{_id:1}})
{ "_id" : "Bushfire", "total" : 1781 }
{ "_id" : "ClimatePolicy", "total" : 495 }
{ "_id" : "NatureBasedSolut", "total" : 537 }
{ "_id" : "Others", "total" : 843 }
{ "_id" : "PartyPolitics", "total" : 343 }
> 
```

Fig. 10. Topics generated by LDA named and clustered by humans based on keywords

3.6 Deep Learning Language Modeling

We have adapted the BERT model [8] to fine-tune our model with our LDA and ChatGPT assisted annotated data. The BERT model is deeply bidirectional and pre-trained using only a plain text corpus, which means it is designed to pre-train deep bidirectional representations from an unlabeled text by joint conditioning on both the left and right context. Bidirectionally trained models can have a deeper sense of language context and flow than single-direction language models and therefore can be fine-tuned with an extra additional output layer to create a domain-specific model. These bidirectional fine-tuned Transformer models can even surpass human performance in this challenging area.

We tested different BERT pre-trained models, as advised in [14], and concluded that in our case the 'bert-base-uncased' model performed the best, therefore we used it for training the model. The concept for fine-tuning of the model using problem-specific application data on top of the BERT pre-trained model is shown in Fig. 11[4]. The training was performed with a 'learning_rate: $4e^{-5}$' and 'num_train_epochs: 5'.



Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached...	Not Spam

Fig. 11. Fine-tuning of pre-trained BERT model with annotated domain-specific data

The proposed concept takes a sample tweet text (pre-processed by removing URL, punctuations, and stop words as well as lemmatized to reduce the number

[4] http://jalammar.github.io/illustrated-bert/.

of tokens) as input and predicts the target label with fine-tuned trained model. It converts a sample tweet input into a feature tensor, which is next classified using a Neural Network to determine its target label. Considering the fine-tuning of the 'bert-base-uncased' model with LDA annotated data leads to more accurate and interpretable models. In the experimental evaluation of posts related to climate change, we showed that the proposed concept is applicable for predicting topics from short text without the need for lengthy and costly annotation.

4 Conclusion

Topic modeling is a useful technique that can automatically analyze text and identify the underlying topic discussed. It can be valuable for different domains, presenting potential advantages for diverse stakeholders. Various methods for topic modeling have been proposed in the literature. However, both main methods (unsupervised and supervised) have shortcomings. To overcome disadvantages this work proposes a hybrid topic modeling method that combines the advantages of both unsupervised and supervised methods.

We built a hybrid model by combining Latent Dirichlet Allocation (LDA) and deep learning built on top of the Bidirectional Encoder Representations from Transformers (BERT) model. LDA is used to identify the optimal number of topics and associated keywords. The only human input needed, with the help of ChatGPT, is to suggest to suggest topic names based on topic-specific keywords and possibly cluster more LDA-defined topics into one. This annotation is used to train and fine-tune the BERT model. In the experimental evaluation on posts related to climate change, we show that the proposed concept is applicable for predicting topics from short text without the need for lengthy and costly annotation. In this work, due to the harnessing of LDA, ChatGPT, and BERT, we completed the annotation of 4,000 posts in about 30 min while the same task required more than 66 h to be fully performed by humans. Testing the accuracy on test data revealed that the proposed concept achieves good accuracy and therefore the proposed concept is applicable for short text topic modeling.

As for future work, it would be interesting to further experiment with parameters for the LDA model to obtain better and maybe more keywords, which will possibly allow better classification and naming of underlying topics. In addition, it is also useful to experiment with deep learning parameters for fine-tuning and different pre-trained models. It is also necessary to devise the method to assess the accuracy of the hybrid method and experiment with other unsupervised topic modeling methods as well as explore the feasibility of applying the proposed method to other domains beyond climate change.

References

1. Albalawi, R., Yeap, T.H., Benyoucef, M.: Using topic modeling methods for short-text data: a comparative analysis. Front. Artif. Intell. **3**, 42 (2020)

2. Becken, S., Friedl, H., Stantic, B., Connolly, R.M., Chen, J.: Climate crisis and flying: social media analysis traces the rise of 'flightshame'. J. Sustain. Tourism **29**, 1450–1469 (2021)
3. Blei, D., Carin, L., Dunson, D.: Probabilistic topic models. IEEE Signal Process. Mag. **27**(6), 55–65 (2010)
4. Blei, D.M., Lafferty, J.D.: A correlated topic model of science (2007)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
6. Chen, J., Stantic, B., Chen, J.: Age prediction of social media users: case study on robots in hospitality. In: Jo, J., et al. (eds.) Robot Intelligence Technology and Applications 7, RiTA 2022. Lecture Notes in Networks and Systems, vol. 642, pp. 426–437. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-26889-2_39
7. Chong, W., Blei, D., Li, F.F.: Simultaneous image classification and annotation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1903–1910 (2009)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL-HLT, pp. 4171–4186 (2019)
9. Ding, K., Choo, W.C., Ng, K.Y., Ng, S.I., Song, P.: Exploring sources of satisfaction and dissatisfaction in Airbnb accommodation using unsupervised and supervised topic modeling. Front. Psychol. **12**, 659481 (2021)
10. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and *F*-score, with implication for evaluation. In: Losada, D.E., Fernández-Luna, J.M. (eds.) ECIR 2005. LNCS, vol. 3408, pp. 345–359. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-31865-1_25
11. Jedrzejowicz, J., Zakrzewska, M.: Text classification using LDA-W2V hybrid algorithm. In: Czarnowski, I., Howlett, R.J., Jain, L.C. (eds.) Intelligent Decision Technologies 2019. SIST, vol. 142, pp. 227–237. Springer, Singapore (2020). https://doi.org/10.1007/978-981-13-8311-3_20
12. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Process. **25**(2–3), 259–284 (1998)
13. Ma, D., Rao, L., Wang, T.: An empirical study of SLDA for information retrieval. In: Salem, M.V.M., Shaalan, K., Oroumchian, F., Shakery, A., Khelalfa, H. (eds.) AIRS 2011. LNCS, vol. 7097, pp. 84–92. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25631-8_8
14. Mandal, R., Chen, J., Becken, S., Stantic, B.: Tweets topic classification and sentiment analysis based on transformer-based language models. Vietnam J. Comput. Sci. **10**, 117–134 (2022)
15. Mcauliffe, J., Blei, D.: Supervised topic models. In: Advances in Neural Information Processing Systems, vol. 20 (2007)
16. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Conference on Empirical Methods in Natural Language Processing, pp. 248–256 (2009)
17. Song, W., Park, S.C.: A novel document clustering model based on latent semantic analysis. In: Third International Conference on Semantics, Knowledge and Grid (SKG 2007), pp. 539–542. IEEE (2007)
18. Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Handbook of Latent Semantic Analysis, pp. 439–460. Psychology Press (2007)
19. Wang, X., Yang, Y.: Neural topic model with attention for supervised learning. In: Conference on Artificial Intelligence and Statistics, pp. 1147–1156 (2020)