Journal of Integrated
SCIENCE & TECHNOLOGY

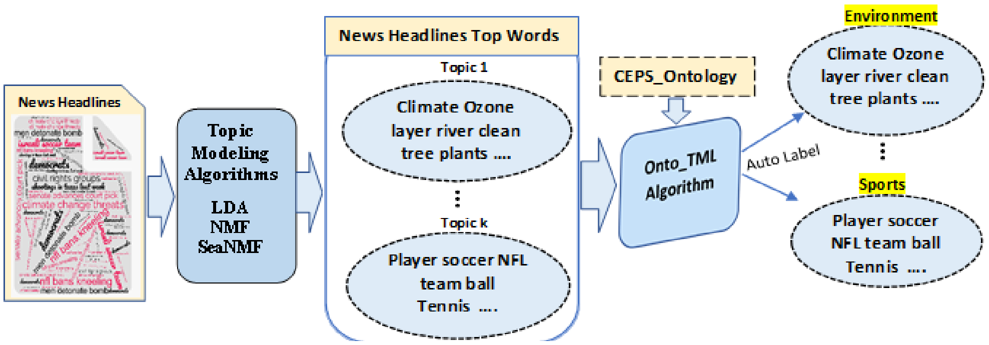
Onto_TML: Auto-labeling of topic models

Supriya Kinariwala, Sachin Deshmukh

Department of computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Auranganad, Maharashtra, India

Received on: 07-Oct-2021, Accepted and Published on: 02-Dec-2021

ABSTRACT


News Headlines
Topic Modeling Algorithms
LDA
NMF
SeaNMF
News Headlines Top Words
Topic 1
Climate Ozone layer river clean tree plants ....
Topic k
Player soccer NFL team ball Tennis ....
CEPS_Ontology
Onto_TML Algorithm
Auto Label
Environment
Climate Ozone layer river clean tree plants ....
Sports
Player soccer NFL team ball Tennis ....

Text mining is a new branch of AI that employs natural language processing techniques to convert unstructured text into a structured format for easier comprehension. It is becoming increasingly significant in every field since it allows users to extract information from large amounts of text or unstructured data. Topic modelling is one of the most important tools in text mining. Topic modelling aids in the discovery of hidden topics, which are the patterns of co-occurring words. Its purpose is to uncover hidden topics in massive amounts of unstructured data. However, because the topics detected are a list of the top n words in a topic, they may not give the viewer a highly coherent image of the document. As a result, automatic topic labelling has been investigated in order to improve understanding of the topics. In this article, we propose a novel method, Onto_TML, ontology-based auto-labelling for topic modelling algorithms and domain-specific CEPS_Ontology. The CEPS-Ontology, which consists of four domains: crime, environment, politics, and sports, is designed using the Protégé tool. Onto_TML uses CEPS_Ontology to assign an appropriate generic label to the top words generated by topic modelling algorithms. For experimentation, we have used two datasets: the News Headline dataset and the News Category V-2 dataset, and LDA, NMF, and SeaNMF topic modelling algorithms. Empirical evaluation shows that Onto_TML has generated appropriate labels for the top words given by topic modelling algorithms. The results have been analyzed using a Normalized Google Distance (NGD) score.

Keywords: Topic Modeling, Auto-Labeling, Ontology, Text Mining, Natural Language Processing

INTRODUCTION

Topic modelling is a widely used unsupervised method for detecting underlying topics in document collections, with several applications in information retrieval.[1,2] To reveal latent structure in regular sized text corpora, traditional algorithms of topic modeling such as Non-negative Matrix Factorization (NMF),[3] Latent Dirichlet Allocation (LDA),[4] Interpretive Structural Modelling (ISM)[5] and Probabilistic Latent Semantic Analysis (PLSA)[6] are used and work on the assumption that a single document may be related to multiple topics. This assumption is not suitable for short text as it contains limited number of words in a single document; hence conventional methods are ineffective for short text data.[7] Due to the fact that short texts lack in contextual information, they suffer from the problem of sparsity.[8,9] Short text topic modeling techniques based on global word co-occurrence is also used for experimentation. Semantic Assisted Non-negative Matrix Factorization (SeaNMF) is a method for determining the themes of short text by revealing the meaningful relationship between keywords and their context. Skip gram is used to discover semantic associations between keywords and their context during the training phase of the word embedding approach. This helps to mitigate the

Supriya Kinariwala,
Email: supriya.kinariwala@mit.asia

Sachin Deshmukh
Email: sndeshmukh@hotmail.com

Cite as: *J. Integr. Sci. Technol., 2021, 9(2), 85-91.*

©ScienceIN ISSN: 2321-4635 http://pubs.iscience.in/jist

sparseness of short text. Here each page is considered as a single sliding window.[10]

Finding one or a few sentences to adequately express the meaning of a topic is known as topic labelling. This activity, which may be time-consuming when dealing with hundreds of themes, has recently gotten a lot of press.[11]

A topic is often represented as a list of terms ranked by their likelihood, but because this can be difficult to read, many ways of assigning labels to topics have been devised.[12,13]

The top 'n' number of terms with the highest marginal probabilities are used to represent the discovered topics in a generic way. However, the top words returned by topic modelling algorithms may not always provide the user with a complete picture of the topic they are trying to portray.[14] As a result, it would be incredibly beneficial if the topics returned could be titled with a label expressing the top topic words. These labels will be extremely valuable since they will allow the user to get a better picture of what the primary themes are in a given corpus of data without having to seek up the top terms in each topic list. Another problem with not labelling subjects is that different people will interpret the top n topic words in different ways, whereas naming the topics will make it easier to understand the topics in a universal way. The problem of autonomously labelling topics is concerned with producing candidate words to represent the labels.[15]

To find relationships between words, we can use ontology for it. An ontology is a collection of concepts and categories in a subject area or domain that displays their attributes and relationships. Protégé is an open source Web-based ontology editor and knowledge acquisition tool. Our contributions are

i) The Onto_TML algorithm is designed to automatically assign labels to the top words clustered by topic model algorithms.

ii) We built an ontology named CEPS_Ontology using the existing Protégé tool infrastructure, focusing on four domains: Sports, Crime, Politics and Environment.

The paper is arranged as follows; section 2 contains a review of labeling techniques for topic models. The proposed method is explained in section 3 followed by a conclusion in section 4.

Literature Survey

In this section, we have endeavored to review all past work in order to do subject labelling. Jeyhan Lau et al[16] suggested a method for automatically classifying topics generated by LDA by combining top ranking topic terms with Wikipedia titles, including the top topic terms, and subphrases retrieved from Wikipedia articles to create a candidate label set. They use a supervised ranking model that incorporates a combination of association measurements and lexical characteristics.

A probabilistic strategy to automatically label topics was investigated by Qiaozhu Mei et al.[17] They present probabilistic ways of objectively categorizing multinomial topic models in their work. They treated the challenge of labelling as an optimization problem, requiring decreasing Kullback-Leibler divergence between word distributions and maximizing mutual information between a label and a topic model. Their results are effective in generating labels that are useful and can be utilized to interpret the topic models discovered. This approach can be applied to a variety of topic models, including LDA, PLSA, and their variants.

Mehdi Allahyari et al.[18] developed a topic model that integrates ontological ideas and topic models into a unified framework, with subjects and concepts represented as a multinomial distribution across concepts and words. They used the semantic relatedness of the concepts and their ontological categories to choose the most appropriate topic labels. Their method includes (1) onto/LDA, which is an ontology-dependent topic model that includes an ontology. (2) A topic labelling method based on the semantics of the concepts in the discovered subjects, as well as the ontological ties between the concepts in the ontology. They demonstrated that by utilizing topic-concept relationships, their model enhances labelling accuracy and can automatically generate labels that are meaningful to users while analyzing topics.

Xian-Ling Mao et al[19] provided a method to address the difficulty of accurately interpreting the meaning of each subject when applying topic models to other knowledge management problems. They offer two methods in their work that exploit the sibling and parent-child relationships among topics to automatically assign accurate labels to each topic in a hierarchy. They have demonstrated that by utilizing their method, the inter-topic relationship is particularly effective in improving topic labelling accuracy and that they can provide meaningful topic labels that can be used to analyze hierarchical subjects.

Davide Magatti et al[20] suggested an approach for automatic topic labelling that uses a hierarchy. A collection of similarity metrics and a set of topic labelling criteria are the two basic measures. These labelling criteria are intended to extract the most widely agreed-upon labels for the given topic and hierarchy. The hierarchy is built using the Google Directory service, as well as themes retrieved via an ad-hoc designed software technique and enlarged using the Open Office English Thesaurus.

The use of text summaries for topic labelling was proposed by Xiaojun Wan et al.[21] To build the summary for each topic, sentences are collected from the most related documents. The study was based on sub modular optimization in order to obtain high relevance summaries. They have shown that the summaries they produce are superior to those produced by existing summarization algorithms, and that summaries are more effective at classifying subjects than words or phrases.

Lau et al.[22] proposed a strategy in which they employ phrases as topic labels and then rank candidate phrases using supervised learning techniques to find the most appropriate label. The top five topic terms, as well as a few nouns from relevant Wikipedia pages, are used as potential labels. Mao et al[23] utilized two algorithms to automatically assign labels to each topic in a hierarchy based on sibling and parent-child relationships between topics. Kou et al.[24] offered another innovative method in which they employed a method to associate topics and phrases with word vectors and letter trigram vectors to determine which label is semantically more comparable to the topics created using data from DBpedia.

Hulpus et al.[25] suggested a new approach to topic categorization that leverages graph centrality measurements. A way of using graphics to represent themes has also been developed. Candidate

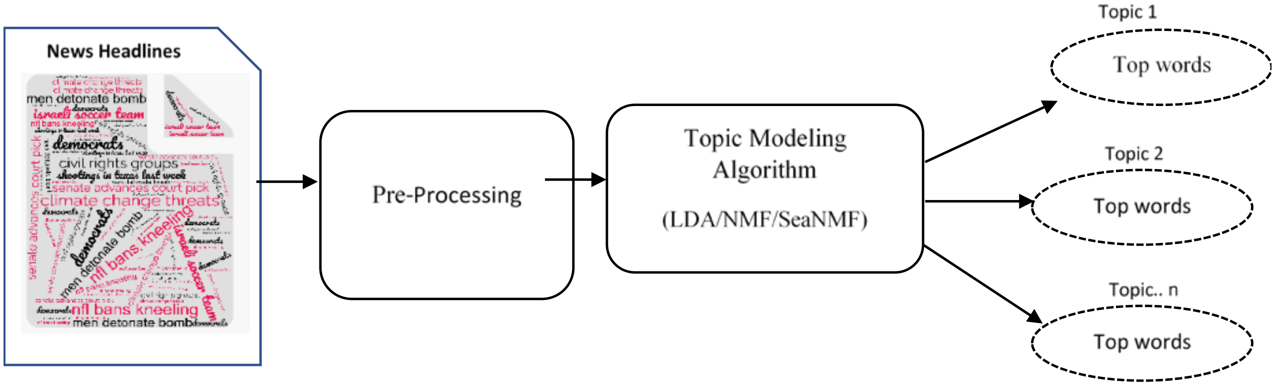


Figure 1: Process of Extracting Top Words

images are extracted instead of labels and used to name the themes.[26] A graph-based technique is used to choose candidate photos.

The primary shortcoming of topic modeling methods is their inability to generate topic labels. In a few articles, the topic labels were chosen manually or based on the term that appeared most frequently in the output cluster. Assigning labels to topics manually is time consuming and requires a subject expert.[27] Here in this paper, we have proposed the CEPS_Ontology and the Onto_TML algorithm to assign labels automatically to the topics based on selected top words.

Proposed Methodology

CEPS_Ontology :

An ontology for the domains of sports, crime, politics, and the environment has been created using the open source tool -Protégé.[28] Protégé is a free, open-source application for creating and managing terminologies and ontologies.[29] It's more than just a terminology editor; it also serves as a platform for developers to incorporate terminologies into end-user applications. CEPS-Ontology comprises 500 words and can be expanded further. For each domain, a list of words has been collected by surveying the online dictionaries. Figure2 shows a portion of CEPS_ontology.

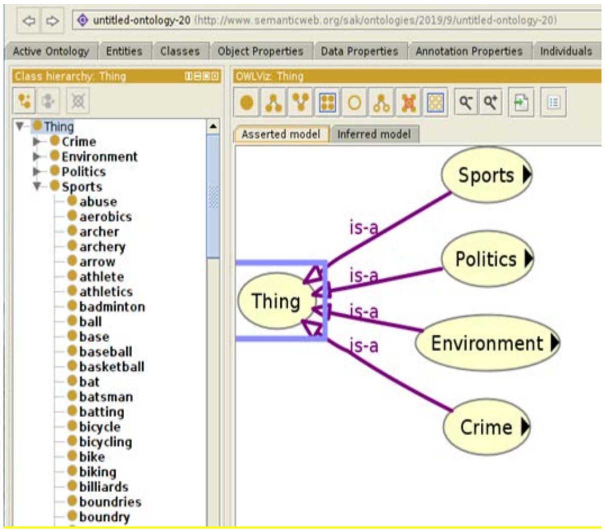


Figure 2 Portion of CEPS_Ontology

Onto_TML Algorithm :

The main objective of the Onto_TML algorithm is to assign appropriate labels automatically to the top words by using CEPS_Ontology.

Following are the steps of Onto_TML algorithm:

Algorithm Onto_TML

```
-------------------------------------------
Input: Topics with top words
Output: Topics with Label
-------------------------------------------
 for each topic
    Ancestor_List={NULL}
    for each word $w_i$
       if word $w_i$ present in CEPS_Ontology
            Append Ancestor_Name of word $w_i$ in Ancestor_List
     label= Ancestor_Name with maximum occurrence
  return label as Topic Label
```

Topics with top words: Top words plays an important role in topic modelling. The Selection of top words varies from algorithm to algorithm. Figure 1 describes the process of extracting top words from the News Headline corpus using topic modelling algorithms. Pre-processing is an important step in cleaning the raw corpus.[30] It includes the removal of stop words, stemming, and lemmatization.

LDA is probabilistic generative topic model used to identify topics in a given document. It works on the assumption that similar words are used to represent similar topics, and each document is a mixture of topics that represent the whole corpus. The model considers that each word is mapped to at least one of the topics of document.[31]

Non-negative matrix factorization is a linear algebraic model used for dimensionality reduction. This method is suitable where underlying factors are non-negative. As NMF yields good clustering results for high dimensional data, it is used for topic modeling.[32] Semantic-assisted NMF (SeaNMF) is a method that reveals semantic relationships between keywords and their context to discover topics from short text.[33] During training, word embedding is used to find semantic associations between top words and their contexts. A semantic correlation matrix(S) between word-context is obtained from the vocabulary of words (V) using the skip gram model.[34]

A lexical group of words created by topic modeling algorithms is fed to a Onto_TML as an input. It finds the word's ancestor using

CEPS_Ontology. Ancestor lists are produced for all words within each topic. For each topic, the ancestor's occurrence count is computed; the ancestor with the highest occurrence count is chosen as the topic label. The Onto_TML algorithm's general flow has been described in the following Figure 3.

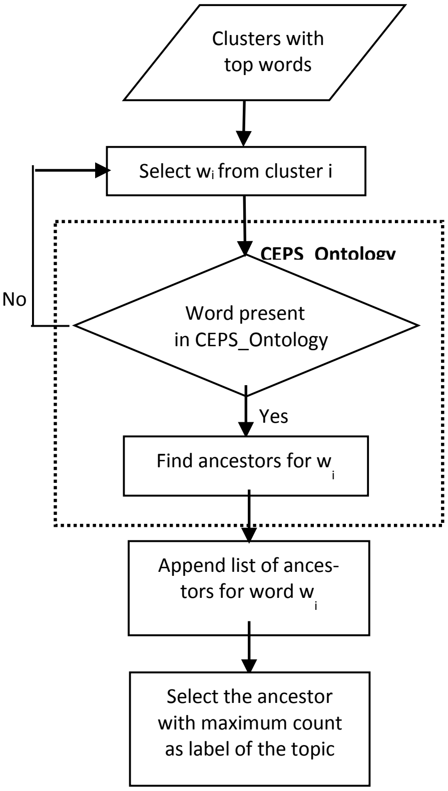

Clusters with top words
Select wᵢ from cluster i
CEPS_Ontology
Word present in CEPS_Ontology
No
Yes
Find ancestors for wᵢ
Append list of ancestors for word wᵢ
Select the ancestor with maximum count as label of the topic

Figure 3: General Flow of Proposed Method : Onto_TML

Results and Discussion

The proposed methodology has been evaluated on the News_Headline dataset and the News_Category_V2 dataset from Kaggle. Both the datasets contain news headlines from various domains. News headlines contain fewer words, so they are sparse in nature. For experimentation, news headlines from four domains such as Sports, Environment, Crime and Politics have been considered.

Experiment 1: Topic Labeling with CEPS_Ontology on News Headline Dataset

Algorithms used are Latent Dirichlet Allocation (LDA), Nonnegative Matrix Factorization, Semantic assisted Nonnegative Matrix Factorization. Dataset used is News Headlines Dataset with 4000 Headlines (Domain - Sports, Crime, Politics and Environment). Output of above-mentioned algorithms is passed to customized ontology to get topic labels. Labels generated by CEPS_Ontology has been verified by using dictionary labels. Dictionary labels are created manually from a domain specific vocabulary generated from the corpus.

Labels generated by Onto_TML algorithm using CEPS_Ontology has an accuracy of 83.33% for News Headlines dataset. Appropriateness of labels generated by Onto_TML algorithm depends upon top words produced by algorithm. Result shows that words produced by LDA under single topic are from multiple domains in comparison with NMF and SeaNMF. LDA is good for regular sized text whereas SeaNMF is better choice for short text. The relevance score is used to verify the topic label. Normalized Google distance is used to calculate the relevance score

Table 1: Labels Generated for Output Produced by LDA

	Topic 0	Topic 1	Topic 2	Topic3
Ontology Label	**Politics**	**Sports**	**Crime**	**Environment**
DictionaryLabel	**Crime**	**Sports**	**Crime**	**Crime**
	shoot	crme	crime	shoot
	house	week	shoot	new
	says	year	olympics	black
	climate	animal	national	hate
	hate	team	day	super
	white	player	dead	dog
	police	national	shooter	people
	school	weather	world	don
	gun	mass	winter	winter
	gop	report	new	bowl

Table 2: Labels Generated for Output Produced by SeaNMF

	Topic 0	Topic 1	Topic 2	Topic3
Ontology Label	**Politics**	**Sports**	**Crime**	**Environment**
Dictionary Label	**Politics**	**Sports**	**Crime**	**Environment**
	gop	olympics	police	dog
	senate	gold	man	baby
	house	women	shoot	animals
	oil	team	shooter	day
	bill	hockey	black	climate
	big	usa	fatally	zoo
	health	open	officer	world
	pruitt	winter	fatal	animal
	democrats	player	suspect	photo
	republican	u.s.	shoots	hurricane

Table 3: Labels Generated for Output Produced by NMF

	Topic_0	Topic_1	Topic_2	Topic_3
Ontology Label	**Sports**	**Crime**	**Crime**	**Environment**
Dictionary Label	**Sports**	**Crime**	**Crime**	**Environment**
	olympics	shoot	crime	week
	winter	police	hate	climate
	team	mass	new	animal
	gold	school	white	change
	hockey	suspect	man	weather
	medal	hate	house	extreme
	skier	killed	york	baby
	house	gun	police	world
	ceremony	dead	black	white
	player	victim	victim	house

between words and the label assigned by ontology. The Normalized Google Distance (NGD) is an extrinsic semantic similarity measure calculated from the number of hits received by the Google search engine for a given collection of top terms. The NGD score is calculated by the formula:

$$NGD(x,y)=\frac{max\{\log f(x),\log f(y)\}-\log(x,y)}{\log M-min\{\log f(x),\log f(y)\}}$$

where f(x) and f(y) are frequency of term x and term y, M is overall number of web pages indexed by Google. NGD(x,y) is a nonnegative score. The terms having NGD greater than 0 and less than 1 are said to be correlated terms.[35] The aggregate NGD score per topic for all described approaches is used to calculate the degree of correlativity of top words with label generated by Onto TML. LDA, NMF, and SeaNMF have average relevance scores of 67 percent, 70 percent, and 89.5 percent, respectively. Ontology labels are based on the results of a topic modelling algorithm.

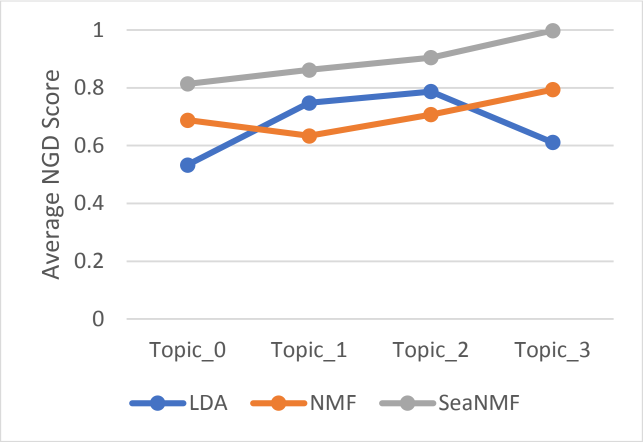


Figure 4: Relevance Score of labels with top words

Experiment 2: Topic Labeling with CEPS_Ontology on News_Category V-2 Dataset

1500 news headlines from each domain are further subdivided into 6 files, each file contains 250 News headlines. On each file LDA, NMF and SeaNMF topic modelling algorithms have been applied to get list of ten top words. Output of above-mentioned algorithms is passed to Onto_TML algorithm to get topic labels. Following tables shows the labels generated by CEPS_Ontology based on domain-wise news headlines.

Table 4: Label Generated by CEPS_Ontology for Crime Domain

File No	LDA	NMF	SeaNMF
1	Crime	Crime	Crime
2	Crime	Crime	Crime
3	Crime	Crime	Crime
4	Crime	Crime	Crime
5	Crime	Crime	Crime
6	Politics	Crime	Crime
X	5	6	6
Y	1	0	0

Table 5: Label Generated by CEPS_Ontology for Politics Domain

File No	LDA	NMF	SeaNMF
1	Politics	Crime	Politics
2	Politics	Politics	Politics
3	Politics	Politics	Politics
4	Politics	Politics	Politics
5	Environment	Politics	Crime
6	Politics	Politics	Politics
X	5	5	5
Y	1	1	3

Table 6: Label Generated byCEPS_Ontology:EnvironmentDomain

File No	LDA	NMF	SeaNMF
1	Environment	Environment	Environment
2	Environment	Environment	Environment
3	Environment	Sports	Environment
4	Environment	Environment	Environment
5	Environment	Environment	Environment
6	Crime	Politics	Environment
X	5	3	6
Y	1	2	0

Table 7: Label Generated by CEPS_Ontology for Sports Domain

File No	LDA	NMF	SeaNMF
1	Sports	Sports	Crime
2	Sports	Sports	Sports
3	Sports	Environment	Sports
4	Sports	Sports	Crime
5	Sports	Sports	Sports
6	Sports	Sports	Sports
X	6	5	4
Y	0	1	2

where X is number of correct label prediction and Y is number of wrong predictions. Accuracy of labels is calculated by number of correctly predicted labels(X) out of total number of predicted labels(N). Figure 5 shows accuracy of label predictions by Onto_TML algorithm on LDA, NMF and SeaNMF topic modelling algorithms.

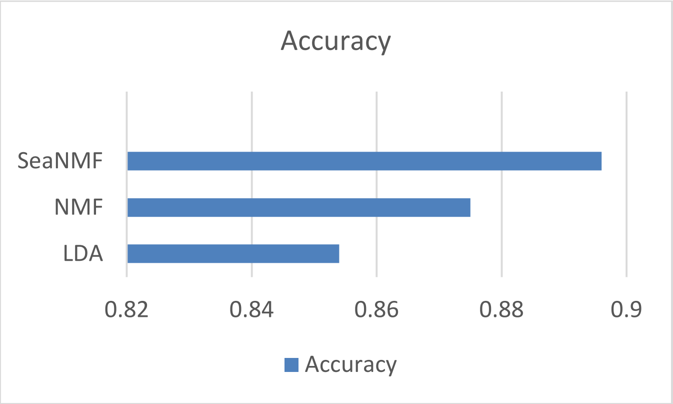


Figure 5: Topic Label Accuracy

Onto-TML algorithm applied on top words generated by topic modelling algorithms. It is observed that the topic labels generated by Onto_TML has average accuracy of 87%.

Conclusion

Traditional topic modelling algorithms extracts top words but do not generate labels for topics. Manual way of labelling the topics takes time for large corpus with multiple domains. Onto_TML algorithm addresses this shortcoming by autolabeling the topics using CEPS_Ontology, which represents set of domain-specific concepts and their relationship. On the basis of this structure, accurate labels are produced for the extracted topics. In this paper experimentation has been done on mixed domain corpus of News headline dataset and domain specific corpus of News Category datsetV2 . The accuracy of labels obtained by Onto_TML algorithm is 83.33% on mixed domain and 87% on domain specific corpus. LDA, NMF, and SeaNMF have average relevance scores of labels with top words is 67 %, 70%, and 89.5% respectively The proposed method Onto_TML can work with any topic modelling algorithms using specific ontology for auto-labeling of topics.

Conflict of Interest

Authors declared no conflict of interest.

References

1. J. Qiang, P. Chen, W. Ding, et al. Topic Discovery from Heterogeneous Texts. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*; IEEE, **2016**; pp 196–203.
2. L. Liu, L. Tang, W. Dong, S. Yao, W. Zhou. An overview of topic modeling and its current applications in bioinformatics. *Springerplus* **2016**, 5 (1), 1608.
3. D. Kuang, J. Choo, H. Park. Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering. In *Partitional Clustering Algorithms*; Springer International Publishing, Cham, **2015**; pp 215–243.
4. D.M. Blei, A.Y. Ng, M.I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, 3 (4–5), 993–1022.
5. R. Singh, S. Kalas, K. Bamarah, A. Singh, R. Kumar. Challenges for Indian Machine Tool Small and Medium Enterprises (SMEs). *J. Integr. Sci. Technol.* **2018**, 6 (2), 25–32.
6. T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*; ACM Press, New York, New York, USA, **1999**; pp 50–57.
7. S.A. Kinariwala, S.N. Deshmukh. Short text topic modeling with empirical learning. *Indian J. Comput. Sci. Eng.* **2020**, 11 (5), 510–516.
8. R. Albalawi, T.H. Yeap, M. Benyoucef. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Front. Artif. Intell.* **2020**, 3, 0042.
9. Q. Jipeng, Q. Zhenyu, L. Yun, Y. Yunhao, W. Xindong. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. *IEEE Trans. Knowl. Data Eng.* **2019**, 1–1.
10. T. Shi, K. Kang, J. Choo, C.K. Reddy. Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*; ACM Press, New York, New York, USA, **2018**; pp 1105–1114.
11. V. Rolim, R. Ferreira Leite de Mello, V. Kovanovic, D. Gasevic. Analysing Social Presence in Online Discussions Through Network and Text Analytics. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*; IEEE, **2019**; pp 163–167.
12. A.E. Cano Basave, Y. He, R. Xu. Automatic Labelling of Topic Models Learned from Twitter by Summarisation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*; Association for Computational Linguistics, Stroudsburg, PA, USA, **2014**; Vol. 2, pp 618–624.
13. S. Hingmire, S. Chougule, G.K. Palshikar, S. Chakraborti. Document classification by topic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*; ACM, New York, NY, USA, **2013**; pp 877–880.
14. C.B. Asmussen, C. Møller. Smart literature review: a practical topic modelling approach to exploratory literature review. *J. Big Data* **2019**, 6 (1), 93.
15. M. Allahyari, K. Kochut. Automatic Topic Labeling Using Ontology-Based Topic Models. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*; IEEE, **2015**; pp 259–264.
16. J.H. Lau, K. Grieser, D. Newman, T. Baldwin. Automatic labelling of topic models. In *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*; **2011**; Vol. 1, pp 1536–1545.
17. Q. Mei, X. Shen, C. Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; **2007**; pp 490–499.
18. M. Allahyari, S. Pouriyeh, K. Kochut, H. Reza. A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling. *Int. J. Adv. Comput. Sci. Appl.* **2017**, 8 (9), 80947.
19. X.-L. Mao, Z.-Y. Ming, Z.-J. Zha, et al. Automatic labeling hierarchical topics. In *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*; ACM Press, New York, New York, USA, **2012**; p 2383.
20. D. Magatti, S. Calegari, D. Ciucci, F. Stella. Automatic Labeling of Topics. In *2009 Ninth International Conference on Intelligent Systems Design and Applications*; IEEE, **2009**; pp 1227–1232.
21. X. Wan, J. Yang, J. Xiao. Manifold-ranking based topic-focused multi-document summarization. In *IJCAI International Joint Conference on Artificial Intelligence*; **2007**; pp 2903–2908.
22. J.H. Lau, D. Newman, S. Karimi, T. Baldwin. Best topic word selection for topic labelling. In *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*; **2010**; Vol. 2, pp 605–613.
23. X.L. Mao, Y.J. Hao, Q. Zhou, et al. A novel fast framework for topic labeling based on similarity-preserved hashing. In *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*; **2016**; pp 3339–3348.
24. W. Kou, F. Li, T. Baldwin. Automatic Labelling of Topic Models Using Word Vectors and Letter Trigram Vectors. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; **2015**; Vol. 9460, pp 253–264.
25. I. Hulpuș, C. Hayes, M. Karnstedt, D. Greene. An eigenvalue-based measure for word-sense disambiguation. In *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, FLAIRS-25*; **2012**; pp 226–231.
26. N. Aletras, M. Stevenson. Representing topics using images. In *NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference*; **2013**; pp 158–167.
27. Z.S. Syed, T. Finin, A. Joshi. Wikipedia as an ontology for describing documents. In *ICWSM 2008 - Proceedings of the 2nd International Conference on Weblogs and Social Media*; **2008**; pp 136–144.
28. M. Strohmaier, S. Walk, J. Pöschko, et al. How ontologies are made: Studying the hidden social dynamics behind collaborative ontology engineering projects. *J. Web Semant.* **2013**, 20, 18–34.
29. N.F. Noy, D.L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology; **2001**.
30. V.I. Nithya. Preprocessing Techniques for Text Mining Preprocessing Techniques for Text Mining. *Int. J. Comput. Sci. Commun. Networks* **2016**, 5, 7–16.
31. Z. Tong, H. Zhang. A Text Mining Research Based on LDA Topic Modelling. In *Computer Science & Information Technology (CS & IT)*;

Academy & Industry Research Collaboration Center (AIRCC), **2016**; pp 201–210.

32. A. Hassani, A. Iranmanesh, N. Mansouri. Text mining using nonnegative matrix factorization and latent semantic analysis. *Neural Comput. Appl.* **2021**, 33 (20), 13745–13766.
33. F. Yi, B. Jiang, J. Wu. Topic Modeling for Short Texts via Word Embedding and Document Correlation. *IEEE Access* **2020**, 8, 30692–30705.
34. T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*; **2013**.
35. A.R. Cohen, P.M.B. Vitanyi. Normalized Google Distance of Multisets with Applications. *ArXiv* **2013**.