Classification based topic extraction using domain-specific vocabulary: a supervised approach

Vandana Kalra[1], Indu Kashyap[1], Harmeet Kaur[2]
[1]Department of Computer Science and Engineering, FET, MRIIRS, Faridabad, India
[2]Department of Computer Science, Hansraj College, University of Delhi, Delhi, India

Article Info

Article history:

Received Oct 26, 2021
Revised Feb 9, 2022
Accepted Feb 16, 2022

Keywords:

Classification
Domain-specific vocabulary
N-Gram
TFIDF
Topic Modeling

ABSTRACT

Recently, a probabilistic topic modelling approach, latent dirichlet allocation (LDA), has been extensively applied in the arena of document classification. However, classical LDA is an unsupervised algorithm implemented using a fixed number of topics without prior domain knowledge and generates different outcomes with the change in the order of documents. This article presents a comprehensive framework to evade the order effect and unsupervised probabilistic nature. First, the framework creates the vocabulary specific to the category using a weight-dependent model that extracts distinctive features suitable for supervised classification. Then, it transforms a classified cluster of documents from the domain corpus to the relevant topic making it more robust to noise. The framework was tested on a comprehensive collection of benchmark news datasets that vary in sample size, class characteristics, and classification tasks. In contrast to the conventional classification methods, the proposed framework achieved 95.56% and 95.23% accuracy when applied on two datasets, indicating that the proposed algorithm has a better classification capability. Furthermore, the topics extracted from the classified clusters are highly relevant to domain categories.

Corresponding Author:

Harmeet Kaur
Department of Computer Science, Hansraj College, University of Delhi
Delhi, India
Email: hkaur@hrc.du.ac.in

1. INTRODUCTION

The automated classification of text documents into categories is a significant urbane task under the umbrella of machine learning [1]. A variety of approaches has been offered in the literature for this task [2]. Presently, document classification algorithms involve preprocessing of text and advanced feature engineering methods [3], [4], such as bag of words (BOW), N-Gram, term frequency (TF), Word2Vec, term frequency-inverse term frequency (TF-IDF), global vectors for word representation (glove and fastText) [5]. These robust feature engineering schemes convert the free-flowing unstructured text into mathematical representations that can be conveniently comprehended by machine learning classification algorithms [6]. The introduction of automated feature engineering capabilities has also become possible to apprehend the core themes in this legible form of data like humans.

Topic modelling is one of the latest techniques with a collection of algorithms that reveal, learn and annotate thematic structure in a group of documents [7]. TF-IDF, N-Gram, bag of words (BOW), and vector space model are the foundations for topic models. It is an unsupervised model which does not require any prior annotation of the documents [8]. It has been implemented in many domains since the first proposed model. In the paper, Onan [9] implemented topic models in the biomedical field motivated by the speedy growth of biological datasets. In the paper, Garcia-Zorita and Pacios [10] analyze the semantic information underlying the collection of documents consisting of the titles on a style unique to Spanish art history known as Mudejar art. In the business domain, topic modelling-based sentiment analysis on tourists' perception toward Indonesia tourism destination Bali is given in [11]. In the paper, Luo *et al.* [12] implemented topic modelling to analyze online reviews for theme park Disneyland to determine visitor behaviour and experiences in the entertainment domain. Real-estate related trends using topic modelling and sentiment analysis on German real estate for market research explored by Ploessl *et al.* [13]. In the paper, Ancin-Murguzur and Hausner [14] analyzed research gaps and trends in the Arctic tundra with the help of a topic-modelling approach. A researcher profile is constructed from non-discernible variables based on research articles by the LDA topic modelling technique in the research domain. It allows the system to capture awareness about his area of proficiency and expertise [15]. There are many more other implementations of topic models. All features in the form of words are not equally important for modelling a document's topic. Therefore, each feature in a document needs to be assigned an appropriate weight. Typically, the weight is based on the number of word occurrences in a document. The perception of weighting is to give more weight to unique terms and less weight to frequently occurring terms. This weighting function is popularly known as term frequency-inverse term frequency (TF-IDF) from the family of weighting functions [16].

According to Tang *et al.* [17], TF-IDF suffers from two drawbacks. First, the IDF inside TF-IDF equals zero if a specific feature appears in all documents. Secondly, the weighting factor IDF moves towards infinity if a particular feature does not appear in a text. Many improved term weighting schemes were proposed by Tang *et al.* to work over these drawbacks. One such improved TF-IDF is the term frequency-inverse exponential frequency (TF-IEF) that characterizes the log-like IDF, a global weighting factor in the corpus. Another term weighting supervised technique was proposed, combining the proportional distortion function and cumulative residual entropy for better accuracy in text classification [18]. Topic modelling algorithms are designed to find hidden semantics in the corpus of documents and cluster the themes as topics. These are also useful to identify latent patterns in the text data. Various methodologies for topic modelling involve matrix decomposition [19]. The method latent semantic indexing (LSI) use singular value decomposition (SVD), and latent dirichlet allocation (LDA) uses a generative probabilistic model for matrix decomposition. Each word is assigned to an explicit topic. LDA model is applied to a document term matrix based on TF-IDF or BOW feature matrix [20]. This method will help calculate the vector distances to classify the documents according to their related topic. Classification with topic modelling means that if a domain corpus consists of numerous known topics and a considerable collection of the documents associated with it, there will be a requirement for classifying the documents according to their allied topics [21], [22]. Therefore, this technique will be supportive for classification [23]. In the paper, Mutanga and Abayomi [24] used the LDA method for topic modelling on COVID-19 tweets. They observed that tracing daily statistics, sale and consumption of alcohol, police brutality, 5G, staying home, and conspiracy theories against vaccines were among the most debated topics around citizens' attitudes and perceptions.

Despite the popularity of the LDA method for topic modelling, various limitations can't be overlooked [25]. First, the number of topics is fixed beforehand without any knowledge of the context of the documents. The second one is it is unsupervised, and sometimes weak supervision is desirable, as in the case of semantic analysis. The third one is that this technique cannot capture correlations among topics. Therefore, a new supervised topic model, twin labeled LDA (TL-LDA), has also been proposed for document classification problems. The researchers executed two independent parallel topic modelling processes. One uses hierarchical Dirichlet distributions having the preliminary label information. And the other is the regrading grouping of tags which incorporate past knowledge about the correlation in labels [26]. In the paper, Lubis *et al.* [27] proposed the novel topic modelling approach on helpful subjective reviews for online courses by extracting learners' experience. In the paper, Mohammed and Al-Augby [28] portrays a comparison study on the classification of e-books using the topic modelling approach LDA and LSA to cluster the words into a set of topics. This provides the opportunity to explore more for a new and relatively unexplored area of supervised topic modelling.

Researchers from various arenas' mathematicians, biologists, computer scientists, software engineers, neuroscientists, and statisticians have considered the topic modelling from diverse outlooks. Therefore, to overcome the shortcomings of previous studies outlined in this paper, a complete topic modelling framework is proposed that involves executing three algorithms in a phased manner. The framework is a supervised mechanism with several separable classified topics known in advance. The algorithms are implemented to create significant domain-specific word vocabulary with their calibrated refined weights, classification methodology using vocabulary, and extraction of the topic from the classified clusters of documents. The procedure will bring light to the potential supervised mechanism responsible for topic modelling with the following contributions: i) Creation of weighted domain-specific vocabulary with significant words, ii) Estimate the threshold value for TF-IDF to control the vocabulary size and removal of insignificant words, iii) Classification of domain-based documents using weighted domain-specific http://mlg.ucd.ie/datasets/bbc.html This framework has made a substantial contribution to topic modelling with the help of vocabulary specific to the domain and classified domain documents. The performance of this automated framework is measured by classification accuracy and observation of the topic, based on the predicted category assigned to the cluster of similar domain documents. The classification accuracy procured using the proposed framework is validated using a classical Naïve Bayes classifier performed over the same set of documents. The rest of the paper is organized as follows: Section 2 describes the datasets and proposes a methodology for creating domain-specific vocabulary, classification-based topic modelling, section 3 addresses the results with discussion, while section 4 presents the main conclusions.

2. MATERIAL AND METHODOLOGY

Experiments were performed on three datasets. The first two were BBC-News and BBC-Sports. These are benchmark sets generated and benefitted for machine learning. BBC-News (http://mlg.ucd.ie/datasets/bbc.html) consists of 2225 news articles under entertainment, politics, business, sports, and technical fields of the standard news domain. BBC-Sports (http://mlg.ucd.ie/datasets/bbc.html), composed of 737 text articles corresponding to sports news in five contemporary areas, were downloaded from the BBC Sport website. The five categories of news articles are athletics, cricket, football, rugby, and tennis under the sports domain. The third dataset of Research-Abstracts consisting of 178 text articles was prepared by extracting abstracts from research papers available online related to five categories: Computer science (https://www.Springer.com/journal/11042), management (https://fbj.Springeropen.com, https://www.Springer.com/journal/10551), psychology (https://journalssagepub.com/description/PSR), social sciences (https://www.Sciencedirect.com/journal/social-sciences-and-humanities-open) and economics (https://www.Journals.elsevier.com/journal-of-international-economics). These three datasets were considered due to their different sizes, domains, and characteristics.

A visual representation of the framework for classification-based topic extraction using weighted domain-specific vocabulary can be seen here in Figure 1. A simple flow of modules has been employed here to study the performance of this framework. This system incorporates three necessary modules weighted vocabulary generation, document classification, clustering, and topic extraction. The algorithms were implemented using Python with Scikit Learn, NLTK, and Matplotlib libraries.

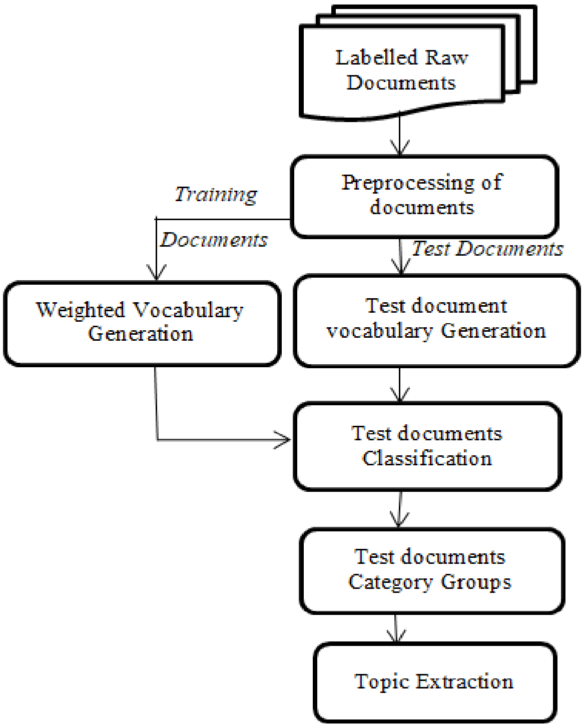

Labelled Raw Documents
Preprocessing of documents
Training Documents
Test Documents
Weighted Vocabulary Generation
Test document vocabulary Generation
Test documents Classification
Test documents Category Groups
Topic Extraction

Figure 1. Framework for classification-based topic extraction using weighted domain-specific vocabulary

The raw text documents from domain corpus are polished through various preprocessing techniques such as tokenization, removal of stop words and punctuation, stemming (PorterStemmer), lemmatization (WordNet), and removal of conjunctions, determiners, pronouns, preposition, articles using POS tagging. Preprocessed documents are partitioned into training and test documents in the ratio of 80%:20%. The algorithm designed by Kalra *et al.* [29] for the generation of domain-specific vocabulary and classification is used here as the basis of topic modelling. Here, the training documents are utilized to design a weighted domain-specific vocabulary comprising significant unigram and bigram features with improved weights (Algorithm 1). Threshold α (1) restricts the vocabulary size by having only significant words [29]. It is calculated as the average TF-IDF scores of every category vocabulary computed at step 6 of Algorithm 1. TF-IDF weight is refined with the total number of categories whose vocabulary does not include the feature (2) [29].

$$\alpha \approx \frac{\Sigma_{category}\left[\left(\Sigma_{features}\, max(TfIdf\,)\right) / \, numberOfFeatures\ \right]}{numberOfCategories} \tag{1}$$

Algorithm 1: Generation of weightedDSVSet

```
Input: Training set (Original_Category, Preprocessed_Text)
Output: Weighted Domain-Specific Vocabulary Set (weightedDSVSet)
 1. Initialize tempVocabularySet and weightedDSVSet
 2. Extract Preprocessed_Text of a particular category from the training set.
 3. Extract all features unigram and bigrams) from the Preprocessed_Text.
 4. Compute TfIdf for all features.
 5. Find a sorted list of the largest TfIdf for every feature.
 6. Find important features(vocabFeature) having TfIdf (feature)> α (threshold)
 7. Append (vocabFeature, TfIdf) with its Original_Category in tempVocabularySet.
 8. Repeat the process 2 to 7 for each category to create the final tempVocabularySet.
 9. Find the weight of every feature ($w_{vocabFeature}$ )of tempVocabulary set (2).

        $\boldsymbol{w_{vocabFeature} = (T_c - \rho_{vocabFeature})}$        (2)

    Where $T_c$ represent the total number of categories, $\rho_{vocabFeature}$ is the number of categories whose
    vocabulary set includes the feature, vocabFeature.
10. Create weightedDSVSet with all (vocabFeature, $w_{vocabFeature}$ ) pairs for every category.
```

The obtained weightedDSVSet with essential features and refined weight for all categories is the standard domain vocabulary set for classifying test documents. First, vocabulary with non-zero values of TF-IDF from preprocessed test document is extracted and stored as testVocabularySet along with its test document [29]. Then, classification of test documents is performed based on matching features from testvocabularySet with that of weightedDSVSet of every category [29]. The total weight is computed per category by adding the weights of matched features (matchedVocabFeature) from a category. The category corresponding to the highest weight is treated as the predicted category of the test document. Once every test document is classified, all test documents with the same predicted category are clubbed together to form a group (classifiedCategoryGroups). The number of groups created here depends upon the number of categories available in the corpus of domain documents. These category groups are then fed into the topic extraction module for identifying the context of the set of documents belonging to a particular domain category, as shown in Algorithm 2.

Algorithm 2: Topic Extraction

```
Input: Classified groups of test documents (classifiedCategoryGroups)
Output: Topic for every classified group
 1.  Combine matched vocabulary Features (matchedVocabFeature) of all test documents belonging to
     a particular category group as combinedCategoryFeature
 2.  Find the frequency of every matchedVocabFeature from the combinedCategoryFeature.
 3.  Store a list of (matchedVocabFeature, frequency) pairs.
 4.  Sort the pairs according to frequency.
 5.  Extract the top $n$ matchedVocabFeature as a topic for a group of test records of a particular
     category.
```

3. RESULTS AND DISCUSSION

This section summarizes and discusses the main findings of this proposed framework. Here, some experimental evaluation of the new algorithms for creating weighted n-gram domain-specific vocabulary, its use for classification of test documents, and finally, extraction of topics from the classified groups are discussed. The classification accuracy of the proposed framework is compared with the results obtained from three types of Naïve Bayes classifier models: multinomial, gaussian and bernoulli. The topics of the classified groups are visually illustrated as wordcloud.

3.1. Creation of weighted n-gram domain-specific vocabulary

The vocabulary generated for all domain categories comprises unigram and bigram significant feature words with their weights as per their importance and the average threshold value. Final weights associated with words are refined further over their basic TF-IDF scores. This refinement in weights is based on a word's uniqueness for a particular category. For example, a few words of vocabulary with the final weight of the category *Computer Science* of Research-Abstract, *Entertainment* of BBC-News, and *Cricket* of BBC-Sports dataset are listed in Table 1. The words with weight 4 (5-1) are significant for the respective category as the total number of categories in these domains is five. Thus, it signifies that these words are members of only one class vocabulary within the domain and can easily classify the test document associated with the respective category.

Table 1. Weighted N-gram vocabulary

Dataset	Category	Features
Research Abstract	Computer Science	**(‘text', 4), ('document', 4), ('accuraci', 4), ('keyphras', 4), ('classif', 4)**, ('**vector machin', 4)**, ('**support vector', 4)**, ('**heterogen data', 4)**, (**cloud comput', 4)**, ('**distanc measur', 4), (**'**artifici intellig', 4)**,…, ('word', 3), ('paper discus', 3),…, ('techniqu', 2),… ,('research', 1)….
BBC-News	Entertainment	**('actor', 4), ('actress' ,4), ('oscar', 4), ('singer', 4), ('album', 4), ('comedi', 4), ('rock' , 4), ('pop', 4), ('nomin', 4), ('drama', 4), ('hollywood', 4), ('theatr', 4), ('concert', 4), ('film releas',4), ('oscar nomine', 4), ('pop star', 4), ('pop song', 4), ('comedi award', 4),** …, ('audienc', 3), ('hit', 3), ('film actor', 3), …, ('stori', 2), ('news websit', 2), …,('watch', 1),…, ('organis', 0), …
BBC-sports	Cricket	**('bowler', 4), ('batsman', 4), ('bat', 4), ('wicket', 4), ('cricket', 4), ('inning', 4), ('icc', 4), ('centuri', 4), ('spinner', 4), ('test cricket', 4), ('cricket council', 4), ('cricket team', 4), ('twenty20 cricket', 4), ('half centuri', 4), ('spin bowler', 4), ('warm up', 4),…,** '(team face', 3),…(,'match', 1),..('play', 0), ('team', 0), …

3.2. Validation of classification using weighted domain-specific vocabulary

Classification of test documents is performed using the weighted domain-specific vocabulary by matching vocabulary words of test documents and categories. Next, the weights of the matched words were added to predict the test document category. Then, the category with maximum weight is allocated as the predicted category of the test document. Finally, a word cloud for a test document of matched words is generated to observe the match's quality visually.

This classification framework records the matched vocabulary words of test documents and domain-specific vocabulary for every category. For example, matched statistics of a test document-1 of BBC-Sports dataset, indicating words and sum of weights of matched words from all category vocabularies are shown in Table 2. It is observed that the total weight of matched words for category *Athletics* is the highest (93) amongst all categories of sports domain (*highlighted in Table 2*). Therefore, the predicted category of test document-1 is *Athletics.*

Table 2. Matching statistics of a test document-1 from BBC-Sports dataset with all categories of vocabulary

Category	Sum of Weights	Matched words
Athletics	**93**	**[('aim', 4), ('indoor season', 4), ('indoor championship', 4), ('trial', 4), ('championship year', 4), ('time minut', 4), ('world championship', 4), ('shape', 4), ('base', 4), ('maintain', 4), ('runner', 4), ('indoor', 4), ('metr', 4), ('mcilroy', 4), ('race race', 4), ('madrid', 3), ('stand', 3), ('ambit', 3), ('compet', 3), ('produc', 3), ('race', 2), ('cours', 2), ('jame', 2), ('line', 1), ('weekend', 1), ('championship', 1), ('retir', 1), ('titl', 1), ('favourit', 1), ('summer', 1), ('nation', 1), ('move', 1), ('ireland', 1), ('earlier', 1), ('admit', 1), ('rest', 0), ('minut', 0),…]**
Football	59	[('fund', 4), ('succeed', 4), ('earlier season', 4), ('spain', 3), ('year career', 3), ('ambit', 3), ('front', 3), ('madrid', 3), ('whose', 3), ('situat', 3), ('night', 3), ('bring', 3), ('race', 2), ('jame', 2), ('cours', 2), ('championship', 1), ('nation', 1), ('line', 1), ('weekend', 1), ('retir', 1), ('chang', 1), ('favourit', 1), ('earlier', 1), ('ireland', 1), ('move', 1), ('admit', 1), ('everi', 1), ('summer', 1), ('titl', 1), ('win', 0), ('place', 0),…]
Tennis	58	[('coach toni', 4), ('form week', 4), ('bogdanov', 4), ('focus', 4), ('list', 4), ('night', 3), ('venu', 3), ('compet', 3), ('becam', 3), ('produc', 3), ('roger', 3), ('spain', 3), ('year career', 3), ('cours', 2), ('race', 2), ('earlier', 1), ('retir', 1), ('everi', 1), ('weekend', 1), ('line', 1), ('favourit', 1), ('titl', 1), ('chang', 1), ('admit', 1), ('championship', 1), ('feel', 0), ('coach', 0),…]
Rugby	34	[('style', 4), ('opinion', 4), ('leadership', 4), ('front', 3), ('becam', 3), ('situat', 3), ('weekend', 1), ('chang', 1), ('ireland', 1), ('nation', 1), ('move', 1), ('retir', 1), ('everi', 1), ('line', 1), ('titl', 1), ('summer', 1), ('admit', 1), ('championship', 1), ('favourit', 1), ('get', 0), ('win', 0), …]
Cricket	36	[('benefit', 4), ('middl', 4), ('richardson', 4), ('stand', 3), ('roger', 3), ('venu', 3), ('bring', 3), ('whose', 3), ('jame', 2), ('nation', 1), ('earlier', 1), ('chang', 1), ('summer', 1), ('ireland', 1), ('move', 1), ('everi', 1), ('start', 0), ('season', 0), ('coach', 0),…]

Test Document-1

mcilroy aim ireland man jame confid titl weekend indoor championship form week metr favourit believ trial race race front middl coach toni lester help get career metr runner promis perform believ decis chang move base bring reward mcilroy windsor feel career leadership style armi sergeant lester lester better work runner roger mark richardson guidanc mcilroy win indoor season mcilroy claim shape sinc ireland championship year temporarili retir return sport lester shrewd guidanc everi race climb mount tri succeed stand start line becam bit mcilroy compar coe cram day compet benefit nation lotteri fund situat chang maintain form repeat time produc race erfurt stuttgart earlier season dmitriy bogdanov madrid venu week claim championship race dutchman arnoud reina spain mcilroy admit look win opinion look beaten season mcilroy whose time minut erfurt elev place list madrid focus year career team import cours weekend end get summer world championship ambit night sinc rest time rebuild career

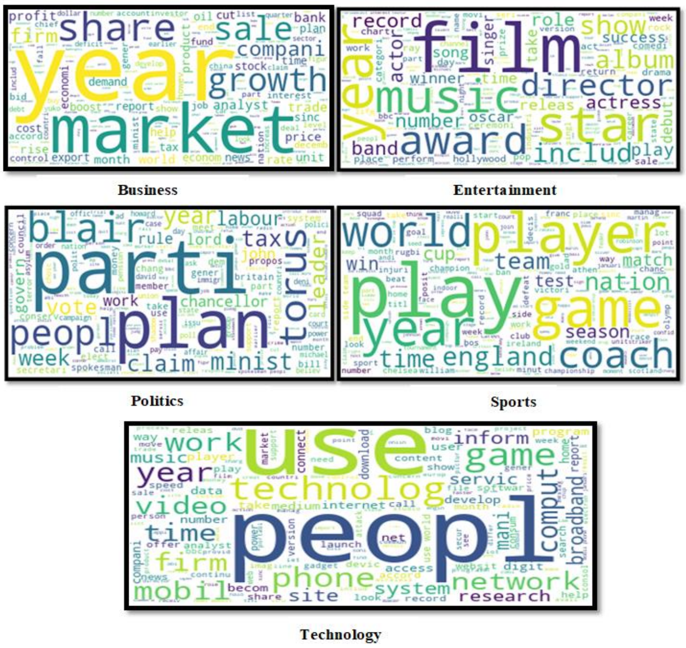
Further, the classification accuracy is computed based on the original and predicted category of the test documents. Finally, three variants of the Naïve Bayes classifier were also applied to the same training and test documents to validate the classification results obtained from the proposed classification methodology. It was found that the classification using weighted domain-specific vocabulary *(highlighted)* produces better, more consistent results compared to the classical Naïve Bayes classifier, as depicted in Table 3.

Table 3. Comparison of classification accuracy

Dataset	Classification of test documents using weighted Domain-Specific Vocabulary		Naïve Bayes (Accuracy)		
	Threshold(α)	Accuracy	Multinomial	Gaussian	Bernoulli
Research-Abstracts	0.021498	**76.493**	65.12	72.10	55.81
BBC-Sports	0.008708	**95.565**	77.09	91.06	64.80
BBC-News	0.004334	**95.233**	91.46	85.18	88.02

3.3. Topic extraction from classified category groups

Classified test documents are grouped according to their predicted category. For further experimentation, matched words of all the test documents in a group are combined to form the matched feature vocabulary. The context from the category group is retrieved by mining the frequent words from the combined matched features vocabulary. The topmost frequent words are considered as the topic of the category group. Table 4 depicts the topics of five classified groups of BBC-News, comprising the top 20 words with their frequency. The word clouds for all category groups in three domains are created to showcase their keywords. For example, the word cloud for the BBC-News domain is shown in Figure 2. It provides perfect five classified groups with all words establishing the topic of the category groups.



Business

Entertainment

Politics

Sports

Technology

Figure 2. Topic word clouds of categories in BBC-news dataset

Table 4. Topics of five categories of BBC-News

Category	Topic (Top 20 words)
Business	[('year', 193), ('market', 105), ('firm', 88), ('compani', 88), ('share', 86), ('growth', 76), ('sale', 75), ('rate', 72), ('month', 70), ('bank', 65), ('price', 65), ('profit', 60), ('rise', 55), ('analyst', 54), ('economi', 53), ('trade', 51), ('interest', 51), ('report', 51), ('product', 49), ('stock', 46)]
Entertainment	[('film', 191), ('year', 124), ('star', 94), ('music', 70), ('award', 62), ('show', 51), ('director', 44), ('album', 41), ('record', 41), ('actor', 41), ('includ', 41), ('world', 34), ('number', 34), ('band', 33), ('play', 32), ('oscar', 32), ('actress', 32), ('role', 31), ('success', 30), ('time', 29)]
Politics	[('plan', 104), ('parti', 103), ('minist', 103), ('peopl', 100), ('year', 93), ('home', 85), ('spokesman', 85), ('blair', 81), ('secretari', 78), ('torus', 74), ('leader', 69), ('michael', 67), ('govern', 64), ('nation', 62), ('work', 59), ('claim', 59), ('campaign', 58), ('labour', 58), ('gener', 57), ('week', 57)]
Sports	[('game', 133), ('play', 131), ('player', 105), ('world', 99), ('year', 99), ('team', 74), ('coach', 73), ('win', 70), ('time', 69), ('nation', 68), ('england', 65), ('cup', 63), ('side', 57), ('champion', 55), ('season', 54), ('get', 54), ('match', 51), ('test', 50), ('end', 49), ('week', 49)]
Technology	[('use', 179), ('peopl', 168), ('game', 124), ('technolog', 113), ('year', 105), ('phone', 89), ('work', 80), ('mobil', 79), ('get', 78), ('way', 74), ('time', 74), ('firm', 67), ('comput', 67), ('system', 64), ('broadband', 62), ('site', 62), ('video', 61), ('servic', 59), ('network', 59), ('websit', 57),]

4. CONCLUSION

The state-of-the-art probabilistic topic modelling model LDA has been suffered from various shortcomings. The LDA algorithm for text classification depends on optimization techniques and is executed without using any historical information stored in the text data. Its implementation required a fixed number of topics to be extracted from the text as initial input without knowing the contextual information in the text. This model has also had an order effect, leading to non-deterministic results. The proposed framework for classification-based topic extraction utilizes prior knowledge in terms of label information to overcome these limitations. It works efficiently for any domain. The four phases of hierarchical execution in the framework: the creation of weighted domain-specific vocabulary, classification of documents using vocabulary, clustering of similar category documents, and topic extraction, makes it a supervised approach. The refined weight assigned to significant features plays a vital role in creating robust vocabulary. Classification of documents was performed well by matching their features with the rich vocabulary of the domain categories. The top significant matched features from a classified group of similar documents were recorded as topics. The resultant topics are not affected by the change in the order of the documents. The classification accuracy under the proposed framework is improved compared with the accuracy obtained from the classical Naïve Bayes classifier when applied on the same split of datasets. The methodology can also be further compared with other classification methods such as support vector machine, and random forest, to prove its strength. In future, there is a need to apply the framework using a large size domain corpus for its better performance.

REFERENCES

[1] E. D. Canedo and B. C. Mendes, “Software requirements classification using machine learning algorithms,” *Entropy*, vol. 22, no. 9, p. 1057, Sep. 2020, doi: 10.3390/E22091057.

[2] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. Mohi Ud Din, “Machine learning based approaches for detecting COVID-19 using clinical text data,” *International Journal of Information Technology (Singapore)*, vol. 12, no. 3, pp. 731–739, Jun. 2020, doi: 10.1007/s41870-020-00495-9.

[3] V. Ganesh, V. Viswanathan, H. S. Kumar, and E. Sivasankar, “Financial sentiment analysis: a study of feature engineering methodologies,” *Soft Computing and Signal Processing*, vol. 1325, pp. 225–240, 2021, doi: 10.1007/978-981-33-6912-2_21.

[4] R. S. Babu and R. Nagarajan, “Improved feature set extraction from documents using modified bag of words,” *Ictact Journal on Soft Computing*, p. 1, 2020, doi: 10.21917/ijsc.2020.0315.

[5] M. W.Habib and Z. N. Sultani, “Twitter sentiment analysis using different machine learning and feature extraction techniques,” *Al-Nahrain Journal of Science*, vol. 24, no. 3, pp. 50–54, Sep. 2021, doi: 10.22401/anjs.24.3.08.

[6] F. Nargesian, H. Samulowitz, U. Khurana, E. B. Khalil, and D. Turaga, “Learning feature engineering for classification,” in *IJCAI International Joint Conference on Artificial Intelligence*, vol. 0, pp. 2529–2535, Aug. 2017, doi: 10.24963/ijcai.2017/352.

[7] C. B. Asmussen and C. Møller, “Smart literature review: a practical topic modelling approach to exploratory literature review,” *Journal of Big Data*, vol. 6, no. 1, Oct. 2019, doi: 10.1186/s40537-019-0255-7.

[8] M. Gerlach, T. P. Peixoto, and E. G. Altmann, “A network approach to topic models,” *Science Advances*, vol. 4, no. 7, Jul. 2018, doi: 10.1126/sciadv.aaq1360.

[9] A. Onan, “Biomedical text categorization based on ensemble pruning and optimized topic modelling,” *Computational and Mathematical Methods in Medicine*, vol. 2018, pp. 1–22, Jul. 2018, doi: 10.1155/2018/2497471.

[10] C. Garcia-Zorita and A. R. Pacios, “Topic modelling characterization of Mudejar art based on document titles,” *Digital Scholarship in the Humanities*, vol. 33, no. 3, pp. 529–539, Oct. 2018, doi: 10.1093/LLC/FQX055.

[11] H. Irawan, G. Akmalia, and R. A. Masrury, “Mining tourist’s perception toward Indonesia tourism destination using sentiment analysis and topic modelling,” in *ACM International Conference Proceeding Series*, Sep. 2019, pp. 7–12, doi: 10.1145/3361821.3361829.

[12] J. M. Luo, H. Q. Vu, G. Li, and R. Law, “Topic modelling for theme park online reviews: analysis of Disneyland,” *Journal of Travel and Tourism Marketing*, vol. 37, no. 2, pp. 272–285, Feb. 2020, doi: 10.1080/10548408.2020.1740138.

[13] F. Ploessl, T. Just, and L. Wehrheim, “Cyclicity of real estate-related trends: topic modelling and sentiment analysis on German real estate news,” *Journal of European Real Estate Research*, vol. 14, no. 3, pp. 381–400, Jul. 2021, doi: 10.1108/JERER-12-2020-0059.

[14] F. J. Ancin-Murguzur and V. H. Hausner, “Research gaps and trends in the arctic tundra: A topic-modelling approach,” *One Ecosystem*, vol. 5, pp. 1–12, Sep. 2020, doi: 10.3897/oneeco.5.e57117.

[15] S. Boussaadi, D. H. Aliane, and P. O. Abdeldjalil, “The researchers profile with topic modeling,” Dec. 2020, doi: 10.1109/ICECOCS50124.2020.9314588.

[16] X. Ao, X. Yu, D. Liu, and H. Tian, “News keywords extraction algorithm based on TextRank and classified TF-IDF,” in *2020 International Wireless Communications and Mobile Computing, IWCMC 2020*, Jun. 2020, pp. 1364–1369, doi: 10.1109/IWCMC48107.2020.9148491.

[17] Z. Tang, W. Li, and Y. Li, “An improved term weighting scheme for text classification,” *Concurrency and Computation: Practice and Experience*, vol. 32, no. 9, May 2020, doi: 10.1002/cpe.5604.

[18] Z. Tang, W. Li, and Y. Li, “An improved supervised term weighting scheme for text representation and classification,” *Expert Systems with Applications*, vol. 189, p. 115985, Mar. 2022, doi: 10.1016/j.eswa.2021.115985.

[19] R. Albalawi, T. H. Yeap, and M. Benyoucef, “Using topic modeling methods for short-text data: a comparative analysis,” *Frontiers in Artificial Intelligence*, vol. 3, Jul. 2020, doi: 10.3389/frai.2020.00042.

[20] M. Rani, A. K. Dhar, and O. P. Vyas, “Semi-automatic terminology ontology learning based on topic modeling,” *Engineering Applications of Artificial Intelligence*, vol. 63, pp. 108–125, Aug. 2017, doi: 10.1016/j.engappai.2017.05.006.

[21] S. Seifollahi, M. Piccardi, and A. Jolfaei, “An embedding-based topic model for document classification,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 3, pp. 1–13, May 2021, doi: 10.1145/3431728.

[22] B. Subeno, R. Kusumaningrum, and Farikhin, “Optimisation towards Latent Dirichlet Allocation: Its Topic Number and Collapsed Gibbs Sampling Inference Process,” *International Journal of Electrical and Computer Engineering*, vol. 8, no. 5, pp. 3204–3213, Oct. 2018, doi: 10.11591/IJECE.V8I5.PP3204-3213.

[23] T. Akilan, D. Shah, N. Patel, and R. Mehta, “Fast detection of duplicate bug reports using LDA-based topic modeling and classification,” in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, Oct. 2020, vol. 2020-October, pp. 1622–1629, doi: 10.1109/SMC42975.2020.9283289.

[24] M. B. Mutanga and A. Abayomi, “Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach,” *African Journal of Science, Technology, Innovation and Development*, vol. 14, no. 1, pp. 163–172, Oct. 2020, doi: 10.1080/20421338.2020.1817262.

[25] R. Joshi, R. Prasad, P. Mewada, and P. Saurabh, “Modified LDA approach for cluster based gene classification using k-mean method,” *Procedia Computer Science*, vol. 171, pp. 2493–2500, 2020, doi: 10.1016/j.procs.2020.04.270.

[26] W. Wang, B. Guo, Y. Shen, H. Yang, Y. Chen, and X. Suo, “Twin labeled LDA: a supervised topic model for document classification,” *Applied Intelligence*, vol. 50, no. 12, pp. 4602–4615, Jul. 2020, doi: 10.1007/s10489-020-01798-x.

[27] F. F. Lubis, Y. Rosmansyah, and S. H. Supangkat, “Topic discovery of online course reviews using LDA with leveraging reviews helpfulness,” *International Journal of Electrical and Computer Engineering*, vol. 9, no. 1, pp. 426–438, Feb. 2019, doi: 10.11591/ijece.v9i1.pp426-438.

[28] S. H. Mohammed and S. Al-Augby, “LSA & LDA topic modeling classification: comparison study on e-books,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 353–362, Jul. 2020, doi: 10.11591/ijeecs.v19.i1.pp353-362.

[29] V. Kalra, I. Kashyap, and H. Kaur, “Generation of domain-specific vocabulary set and classification of documents: weight-inclusion approach,” *International Journal of Information Technology (Singapore)*, Jan. 2022, doi: 10.1007/s41870-021-00830-8.

BIOGRAPHIES OF AUTHORS

Vandana Kalra is an Associate Professor at Sri Guru Gobind Singh College of Commerce, University of Delhi, Delhi, India. She is M.Sc, M.Phil (Computer Science), PhD Scholar (Computer Science and Engineering). She holds the post of Head of Department of Computer science from 2006 to 2019. She has been teaching in this field for 21 years and has been active in the multi-disciplinary field of research and innovation. She has a deep passion for innovation related to technology and advanced areas of computer science. She has been involved in many innovation projects and achieved the best societal impact award for one of the projects. She can be contacted at email: vandana.kalra@sggscc.du.ac.in.

Dr. Indu Kashyap has 15 years of teaching, administration, and research experience. Currently, she is a Professor with Manav Rachna International Institute of Research and Studies. She has done her PhD from Chaudhary Charan Singh University, Meerut. She has to her credit more than 60 publications in reputed journals and conferences, including Elsevier, Springer, and Taylor & Francis. Her research areas include Wireless Networks, Machine Learning, Data Analytics, and Recommender Systems. She can be contacted at email: indu.kashyap82@gmail.com.

Dr. Harmeet Kaur is an Associate Professor in the Department of Computer Science, Hansraj College, University of Delhi. Her teaching experience spans over around 24 years. She completed her PhD from the Department of Computer Science, at the University of Delhi, and her research interests lie in the field of Recommender Systems and Crowdsourcing. She has published around 45 research papers in national and international journals and conferences of repute. She can be contacted at email: hkaur@hrc.du.ac.in.