Information Retrieval Using Effective Bigram Topic Modeling

Vrishali A. Chakkarwar[1](✉) and Sharvari C. Tamane[2]

[1] Department of Computer Science and Engineering, Government Engineering College, Aurangabad, Aurangabad, India
vrush.a143@gmail.com

[2] I.T. Department, University Department of Information and Communication Technology, MGM University, Aurangabad, Aurangabad, India

Abstract. Many fields, such as film reviews, recommendation systems, and language processing, have effectively adopted and utilized topic modeling with Latent Dirichlet Allocation (LDA). Many texts analysis tasks, however, rely heavily on sentence construction and words to capture the meaning of text. However, word coexistence plays an important role in retrieving significant data. In this paper we present a novel method which discovers topics and topical phrases using language modeling. Proposed bigram extended LDA gives promising results to discover latent research areas in research articles and efficient classification of research articles. Experimental results are carried out to test the efficiency of proposed method.

Keywords: Natural language Processing · topic modeling · bigram

1 Introduction

Researchers nowadays who are interested in determining the social traits and dynamics of a specific setting have access to a vast quantity of digital data from which to derive answers to their inquiries. The amount of written research publication content available online alone gives a great significance of knowledge, context, and evidence that may be utilized to solve social scientific problems. Even at the computational level, sifting through this amount of data and highlighting conversations and patterns of special interest becomes difficult, let alone for individual scholars to explore.

The task of collecting good phrases from a corpus is known as phrase mining. Rather than a lengthy technical explanation of a quality term, we shall begin with a basic definition. A quality term in the topic of interest indicates a notion, idea, approach, or procedure. A number of applications require the mining of high-quality phrases [1–5]. There are now two categories of phrase mining techniques: (1) Supervised approaches [6, 7] that make use of hand-crafted or manually acquired lists of high-quality phrases, and (2) unsupervised methods [8–11] that do not use any curated phrase instances. For example, if we want to identify good phrases for the any domain, we may utilize Auto

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 784–791, 2023.
https://doi.org/10.2991/978-94-6463-136-4_68

phrase like ‘neural network’, ‘computer vision’, ‘network security’ and ‘production planning’.

Topic modeling is the most powerful tools for text mining, latent data identification, and detecting links between data and text documents, which is an unsupervised machine learning technology that can recognize words and phrases text corpus. It is primarily described as a statistical text mining approach for identifying possible hidden patterns in a text corpus and categorizing key words in a corpus as themes. Topic modeling has a wide range of applications. Many publications have been published by researchers in a variety of subjects, including software engineering, political science, medical science, and linguistics. Topic modelling based on social media analysis, for example, aids in the comprehension of online reactions and discussions; it pulls important information from comments and material published on social media sites such as Twitter and Facebook [12]. Topic modelling is a scientific approach for identifying and monitoring word clusters, or subjects, in huge volumes of text. A subject may be thought of statistically as a multinomial distribution of words that emerge at the same time. Topics may be learnt from a document collection by running the topic model on repeatedly on it.

2 Literature Survey

Topic modeling identifies hidden semantic information from huge text corpus. Various researchers implemented LDA model for numerous applications.

In this paper, [1] propose to enhance topic modeling by enable a system to dynamically search for word concurrence text that are significant to identifying a target word. Phrase-based system using this new strategy BL LDA incorporates the bigram concept for supervised generative modeling for multilabel text input. Proposed method outperforms the simple LDA.

Biterm Topic Model (BTM) [2] generated the topics by directly modeling word co-occurrence patterns (i.e., biterm) in a text corpus. This algorithm extracts word pair that coexist frequently together and semantically linked to each other, whereas the suggested model makes use of the corpus-level adjacent word and is based on both bigrams and unigrams.

The bag-of-words assumption, on which many traditional topic modeling techniques are predicated, states that each word has a topic distribution for each topic. Numerous in-depth investigations have been conducted to make sure the material is comprehended [2–5]. Instead of being a single word, word units are N-grams or word pairs having co-occurrence patterns gives better information retrieval.

Text mining has emerged as one of the most crucial fields in the internet age as a result of the tremendous expansion of textual content. In recent years, topic models have grown in favorably. A theme is made up of a collection of words that frequently occur together. Topic models outperform other methods for extracting semantic information from data. LSA (Latent Semantic Analysis), PLSA (Probabilistic Latent Semantic Analysis), and LDA are the many approaches utilized for topic models (Latent Dirichlet Allocation). These techniques have become useful for detecting hidden themes in documents (corpus). To investigate, summaries, and reveal hidden conceptual patterns from big corpora, various topic modeling methods have been created. In this paper, [6] offer a

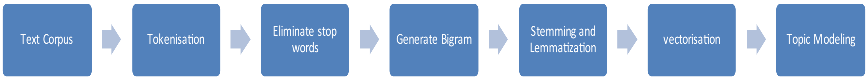

Fig. 1. Proposed model for Bigram topic Model

comprehensive overview of the numerous subject modeling strategies proposed over the last decade in this work. Furthermore, researchers are concentrating on various ways for extracting ideas from social media material, with the purpose of finding and aggregating the topic inside brief texts. It also outlines the many uses and quantitative evaluations of the various methodologies, using computational and numerical knowledge to forecast future convergent.

3 Proposed Methodology

The goal of conventional topic models is to identify the latent subjects of text based on patterns of word co-occurrence at the corpus level. These models also make use of the premise of a bag of words. We suggest a supervised generating model for multi-label corpora by combining them with the bigram notion to Simple LDA. We introduce the Bigram LDA-generation process.

In this proposed work, topic modeling is applied to huge number of research articles. Simple LDA model is implemented using unigram language modeling. Proposed model is implemented using bigram. Word ordering plays very important role for finding contextual meaning in text corpus. Bigram Terms like “Artificial Intelligence” carries more information as like unigrams “artificial” and “intelligence”. This is also referred as “Phrase Modeling”. Model complexity increased in phrase modeling but it improves contextual information. Following flow diagram indicates system flow (Fig. 1).

In this research dataset of 16000 IEEE publications abstracts are used. Each abstract contains roughly 200 words. This itself the big text corpus. Topic models cannot “read” the sentence since they lack any genuine semantic understanding of the terms. Topic models instead rely on mathematics. Statistics suggest that the tokens or words that frequently occur together are connected. Following pre-processing or “cleaning” steps are carried out to ensure that the topic model is detecting interesting or significant patterns rather than noise.

3.1 Preprocessing

Text data with English language properties is used as input data. Before using any clustering or machine learning algorithm, this data must be preprocessed. Stop word removal, stemming, and lemmatization, as well as part of speech tagging, are common preprocessing steps. The following are typical Natural Language Processing steps:

Tokenization is the process of dividing text into units known as tokens. Every token is known as a word. Stop word removal: Frequently, the most common terms don’t imply much. “The, a, of, for, in,” as examples Such a word must be deleted. This process is known as stop word elimination. English nouns like “view” can be stemmed to become the words “viewing, seen” by adding a morphological suffix. They have the same root noun, “view.” Verb omission: Verbs can be used to describe actions but are not thought of as topics. Therefore, these verbs must be eliminated.

3.2 Formation of Bigrams Topic Modelling

For finding and extracting the underlying structure themes in text data, LDA is a popular type of unsupervised and probabilistic topic modeling technique. LDA’s core tenet is that each document is essentially represented as a probability distribution or as a combination of topics, whereas each topic is represented as a probability distribution over a collection of words. The LDA model is based on the BoW assumption.

In this work we assume word co-occurrence plays is very important role. Each document is sequence of words. Consecutive sequence of word form bigram. For any document any three words generates bigrams as follows

$$\{w1,\ w2,\ w3\} = \{(w1,\ w2),\ (w2,\ w3)\} \tag{1}$$

For every document, consecutive word pairs are formed that are referred as bigrams.

Proposed Bigram model accomplishes this goal by simulating the development of bigrams, in contrast to most topic models, which learn the latent topic components in a corpus by simulating the generation of texts. The main concept is that two words are more likely to relate to the same topic if they co-occur more frequently. This notion leads us to believe that the two words in a bigram which are more frequent are independently selected from a subject, where a topic is sampled from a topic mixture throughout the entire corpus.

A new set of random variables was introduced by the LDA Bigram Model x (xi = 1: $\mathrm{w_{i-1}}$ and $\mathrm{w_i}$ form a bigram; xi = 0) it indicates that the previous word token can be made with a bigram, z and w are two sets of random variables in addition. As a result, it may choose whether to produce a bigram or a unigram. After all, unigrams are crucial part of a text. The status variable x1 is assumed to be present, and only unigrams are permitted at the start of a document. A bigram can receive a subject after the fact: the first word of a phrase is always formed from the LDA component that has a topic assignment, and this can be used as the phrase’s topic. As one might assume, this procedure does not always identify a plausible topic to a sentence. The generative process of the LDA Bigram model can be described as follows:

1. Draw multinomial ϕ_{z} from a Dirichlet prior β;
2. Draw binomial φ_{w} from a Beta prior γ;
3. Draw multinomial σ_{w} from a Dirichlet prior δ;
4. For each document d, draw a multinomial $\theta^{(\mathrm{d})}$from a Dirichlet prior α; then for each word $\mathrm{w_i}^{(\mathrm{d})}$ in document d:

   a. Draw $\mathrm{x_i}^{(\mathrm{d})}$ from binomial ${}^{\varphi}\mathrm{w_{i-1}}^{(\mathrm{d})}$;
   b. Draw $\mathrm{z_i}^{(\mathrm{d})}$ from multinomial $\theta^{(\mathrm{d})}$;
   c. Draw $\mathrm{wi}^{(\mathrm{d})}$ from multinomial $\sigma_{\mathrm{wi}}{}^{(\mathrm{d})}{}_{-1}$ if $\mathrm{x}^{(\mathrm{d})}{}_{\mathrm{i}} = 1$; else draw $\mathrm{w_i}^{(\mathrm{d})}$ from multinomial $\phi_{\mathrm{zi}}{}^{(\mathrm{d})}$.

Table 1. Displays topics from a 20-topic run on the dataset alongside a comparison to the equivalent nearest LDA topics as anecdotal evidence.

	Topic 1	Topic 2	Topic 3	Topic 4
Unigram + LDA	Intelligent, artificial, computer, vision, image	Learning, model, machine, accuracy, network	Cloud, computing, service, resource, model	Power, energy, grid, distribution, control, load
Bigram + LDA	Artificial Intelligence, Computer vision	Neural network, machine learning, Deep learning	Cloud computing, Server model, Cloud service	Power system, Load distribution, Voltage control

Experimental Results. The research publication dataset made up of 10 years' worth of IEEE articles abstracts from various domains, we apply the Bigram LDA model. The dataset includes 30988 unique terms and 16000 research paper abstracts. In simple LDA model unigrams are considered as basic unit of processing while in bigram model combination of unigram and bigram terms are considered as basic units. Following table indicates keyword representing topics which clearly indicates bigram model gives more contextual data as compared to unigram model (Table 1).

As per above table topic 1 provides summary of Artificial Intelligence domain research articles. Simple LDA Model generated word like "artificial", "computer", "vision" that are highly probable occurring words. But Bigram Model find these generic words associates with each other such as "Artificial Intelligence", "computer Vision". It indicates simple LDA generated word makes topic less understandable than bigram model. It is very meaningful Phrase "Artificial Intelligence" as compared to single word "artificial", "intelligence".

Topic model is evaluated by Topic Coherence value. Topic coherence is score measures semantic similarity between words in the topics derived. Topic coherences are high when semantically interpretable topics are generated. Many measures based on the word co-occurrence score of the most important terms for each individual topic have been employed in the topic modeling literature. The best technique to determine how interpretable a topic is to evaluate the coherence of the topic. The gold standard for assessing coherence is human topic ranking, but they are pricy. In this experiment, we adopt a particular case of topic coherence called "mean PMI" (mean point wise mutual information). Here topic modeling is applied for topics 6, 9, 12, 15, 18 Coherence score for simple LDA is observed near to .49 and it is nearly same for topics 9 to 15. Coherence score ranges from 0.64 to 0.67 for bigram model. A result indicates more interpreting information retrieval using bigram topic model. Motivating factor in this work is high values of coherence obtained from excellent model with more interpreting coherent topics. Results indicate topic model using bigrams increases contextual data and thus coherence of model. Research articles are classified with more prominent areas (Figs. 2 and 3).

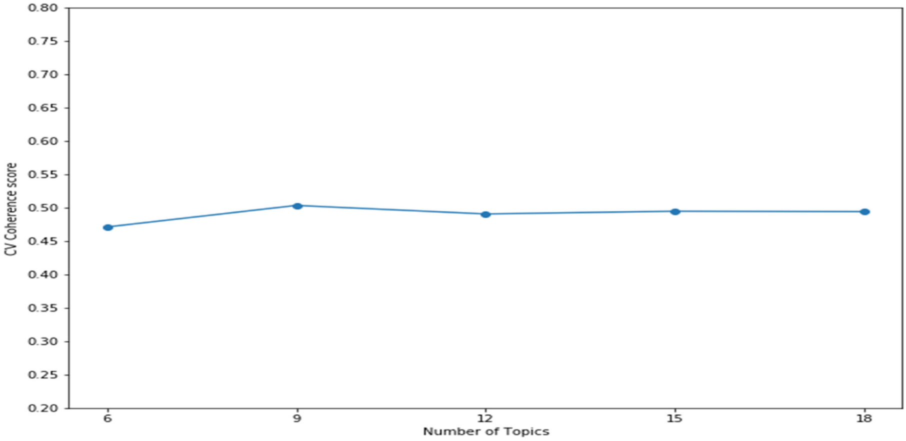

Fig. 2. Topic Coherence values using simple LDA.

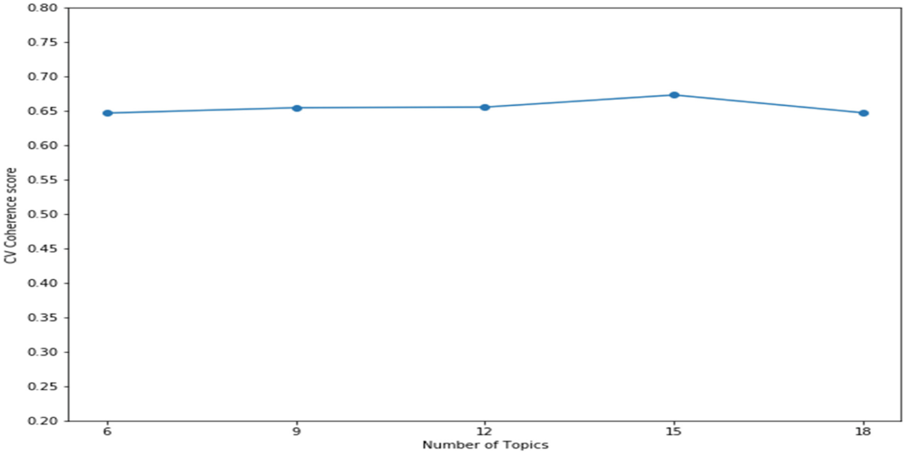

Fig. 3. Topic Coherence values for bigram Model LDA.

4 Conclusion

In this paper we present extended LDA using bigram language modeling, this model has shown very promising results as compared to simple LDA. In this proposed work, we have presented the bigram LDA model. By considering the topic coherence values Bigram LDA model outperforms on Simple LDA model. When LDA is used to classify research paper articles, bigram terms generated gives better information retrieval classification of research articles. Results are more interpreting. Here we successfully integrated bigram model using language modeling for contextual information retrieval.

References

1. Y. Park, M. H. Alam, W. -J. Ryu and S. Lee, “BL-LDA: Bringing Bigram to Supervised Topic Model,” 2015 International Conference on Computational Science and Computational Intelligence (CSCI), 2015, pp. 83-88, doi: https://doi.org/10.1109/CSCI.2015.146.
2. X. Cheng, X. Yan, Y. Lan and J. Guo, “BTM: topic modeling over short texts”, TKDE, vol. 26, no. 12, pp. 2928-2941, 2014.
3. X. Wang, A. McCallum and X. Wei, “Topical n-grams: phrase and topic discover with an application to information retrieval”, ICDM, pp. 697–702, 2007.
4. A. El-Kishky, Y. Song, C. Wang, C. R. Voss and J. Han, “Scalable topical phrase mining from text corpora”, VLDB, pp. 305–316, 2015.
5. H. M. Wallach, “Topic modeling: beyond bag-of-words”, ICML, pp. 977–984, 2006.
6. S Likhitha, B S Harish and Keerthi H M Kumar. A Detailed Survey on Topic Modeling for Document and Short Text Data. International Journal of Computer Applications 178(39):1–9, August 2019.
7. A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han, \Scalable topical phrase mining from text corpora,” Proc. VLDB Endow., vol. 8, no. 3, p. 305{316, Nov. 2014. [Online]. Available: https://doi.org/10.14778/2735508.2735519
8. N. Kawamae, \Supervised n-gram topic model,” in Proceedings of the 7th ACM International Conference on Web Search and Data Mining, ser. WSDM ‘14.New York, NY, USA: Association for Computing Machinery, 2014, p. 473{482. [Online]. Available:https://doi.org/10.1145/2556195.2559895
9. X.Wang, A. McCallum, and X.Wei, \Topical n-grams: Phrase and topic discovery, with an application to information retrieval,” in Seventh IEEE International Conference on Data Mining (ICDM 2007), 2007, p. 697{702.
10. J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han, \Automated phrase mining from massive text corpora,” IEEE Transactions on Knowledge and DataEngineering, vol. 30, no. 10, p. 1825{1837, 2018.
11. K.-h. Chen and H.-H. Chen, \Extracting noun phrases from large-scale texts: A hybrid approach and its automatic evaluation,” in Proceedings of the 32nd annualmeeting on Association for Computational Linguistics. Association for Computational Linguistics, 1994, pp.234{241.
12. J. Liu, J. Shang, C. Wang, X. Ren, and J. Han,”Mining quality phrases from massive text corpora,” in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 2015, pp. 1729–1744.
13. A. El-Kishky, Y. Song, C. Wang, C. Voss, and J. Han, “Scalable topical phrase mining from text corpora,” arXiv preprint arXiv:1406.6312, 2014.
14. F. Boudin, “Unsupervised keyphrase extraction with multipartite graphs,” arXiv preprint arXiv:1803.08721, 2018.
15. C. Florescu and C. Caragea, \Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents,” in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1105–1115.
16. B. Li, X. Yang, B. Wang, and W. Cui, “Efficiently mining high quality phrases from texts,” in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, p. 3474–3481.
17. Sanandres Campis, Eliana & Llanos, Raimundo &Madariaga Orozco, Camilo. (2018). Topic Modeling of Twitter Conversations.

18. Chakkarwar Vrishali, Tamane Sharvari. (2020). Quick Insight of Research Literature Using Topic Modeling. https://doi.org/10.1007/978-981-15-0077-0_20.
19. V. Chakkarwar and S. Tamane, "Social Media Analytics during Pandemic for Covid19 using Topic Modeling," 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), 2020, pp. 279–282, doi: https://doi.org/10.1109/ICSIDEMPC49020.2020.9299617.