Topic Modelling with Bag-of-concepts Document Representation

Metwally Rashad
Computer Science Department
Faculty of Computers & Artificial Intelligence, Benha University, Egypt
Artificial Intelligence, Delta university for science and technology, Gamasa, Egypt
metwally.rashad@fci.bu.edu.eg

Ibrahim Reyad
Information Systems Department
Faculty of Computers & Artificial Intelligence, Benha University, Egypt
ibrahim.elsayed@fci.bu.edu.eg

Mohamed Abdelfatah
Information Systems Department
Faculty of Computers & Artificial Intelligence, Benha University,Egypt
mohamed.abdo@fci.bu.edu.eg

***Abstract*—Traditionally, text mining tasks have been implemented by applying topic models like Latent Dirichlet Allocation (LDA). These topic models occasionally produce noisy words in illogical topics with a high probability. The problem is that topic model-based approaches are sparse, have binary weighting for terms, and lack semantic data. The topic model technique is combined with a document representation technique called Bag-of-Concepts to solve these problems. The bag-of-concepts approach groups word vectors from word2vec to create concepts, which are subsequently represented in document vectors by these concept cluster occurrences. The performance of document proximity preservation is taken into account by Bag-of-concepts when using the suitable weighting formula concept frequency-inverse document frequency. Latent Dirichlet Allocation is adjusted for use in document clustering and quality tasks for topics. The results are compared with different LDA frameworks on text documents, as well as the bag-of-concepts representation of documents. LDA with Bag-of-concepts representation generates more cohesive themes in comparison to the other techniques.**

***Index Terms*—Bag-of-concepts, Latent Dirichlet Allocation, Document Representation**

I. Introduction

Using topic models [1] for analysis and discovery of significant statistical patterns in vast collections of text documents has been demonstrated to be effective. Topic Models, which are probabilistic models used to find out the semantic structure that is latent in the massive amounts of text content. Text mining apps for Topic Models are numerous, including machine translation [2], sentiment analysis [3, 4], opinion mining [5], social network analysis [6], and multi-document summarization [7, 8, 9].

The assumption made by topic modeling is that the text documents are a combination of various latent topics, and that each topic is a dispersion of vocabulary words. Each document's topic dispersion and each topic's word distribution are inferred using a probabilistic framework[1]. Topic models frequently create topics low quality and with large probabilities of noisy or unrelated words. There are three causes for this. First, words with high frequency, some of which are stop-words with specialized domains are given higher weight. Additionally, a low frequency of an informational term may cause it to be given less weight. Second, there are extremely few correlations between document terms. Only a tiny number of vocabulary terms are used in a document. This issue is worse for short texts like tweets and snippets. Third, semantic associations between words are not taken into account by topic models.

In the bag-of-concepts method, concepts are generated by grouping word vectors generated by trained word2vec, and document vectors are then represented by the occurrences of these concept clusters. Bag-of-concepts make use of an appropriate weighting formula called concept frequency-inverse document frequency, and consider how semantically comparable words affect the effectiveness of document proximity preservation. For clustering of document and quality tasks for topics, Latent Dirichlet Allocation is adapted for usage Bag-of-concepts representation of documents is compared, as are the results with various LDA frameworks on text documents. Compared to the other methods, LDA with Bag-of-concepts representation produces more coherent topics.

This paper is structured as follows: The literature review is found in the second section. The third section describes the proposed method. In forth Section, the suggested model's experimental results, as well as all comparing approaches, are provided in three standard datasets.

II. Literature review

Regular topic models assign equal weights to each word in the document. subsequently, uninformative words that appear regularly throughout the corpus are included. There have been many term weighting schemes proposed [10, 11, 12], that enhance the performance quality of text mining tasks such as topic modelling. According to the authors of [11], high-occurrence words contribute bit to document interpretation for discrimination. These words with a high frequency of occurrence are either ordinary stop-words or stop-words of domain-specific. In [11] LDA with log weighted proposed which uses information theory to reduce the weights of similar words by allocating log weights $\log(\mathrm{p(w)})$ to every word w, where $\mathrm{p(w)}$ is the likelihood of the word w appearing in the dataset.

The LDA with PMI weight [12] method allocates various weights to various words in various documents. The weight assigned to a word w in a document is $\log(\mathrm{p(w/d)/p(w)})$. In [11] an approach which is supervised known as the BDC term weighting scheme to penalize domain-specific stop words. An entropy-based term weighting suggested in [11], also the researchers in [10] combined an entropy-based term weighting scheme with log weight and BDC weight to create CEW, which stands for mixed entropy-based term weighting scheme.

On several text mining tasks, the term weighting schemes outperformed non-weighted topic models using LDA and its variants.

III. PROPOSED METHOD

In the proposed method, we will explain proposed model steps as shown in figure 1 including, how Bag-of-Concepts of words is used to represent documents,and how Latent Dirichlet Allocation is adjusted to take into account the bag-of-concepts document representation.

A. Preprocessing step

Text preprocessing is required to convert the text into a readable format for use with machine learning techniques. If we didn't preprocess the text data, the algorithm built on top of it would be useless. It has no commercial value. The output of the algorithm, like its raw data, would be garbage (useless). And as a result, The preprocessing steps are applied to datasets to eliminate erroneously and stop words. as some algorithms use exact word in matching process, LDA also performs tokenization [13], stopword removal [14] using N-grams [15], and stemming [16, 17].

B. Bag-of-concepts Representation of Documents

In word2vec representation model word embeddings are produced using a group of related models. These models include 2-layer neural networks that have been taught to construct word context from linguistic information. The word2vec model, in particular, takes a sizable corpus of texts as input and creates a vector space with hundreds of dimensions. Additionally, a matching vector in the space is assigned to every distinct word in the dataset. Words that have common context in the corpus are placed next to one another in the space because word vectors are located there [18]. This word2vec model does not provide an intuitive explanation of the produced document vectors. as using a neural network every document vector is trained, each vector figure means only the robustness of the link among the input and hidden nodes. As a result, it is difficult to understand what every feature of document vector representation denotes in respect of the document's actual contents. As a result, if a text mining model, such as a document classifier, is trained using these word2vec document vectors, it does not succeed to add any information to the model's operating logic. The most goal of text mining is not to create a good document representation. To have a meaningful impact and implication in the real business environment, these representation methods must be capable of providing a clear explanation after the representation and its subsequently made text mining model. Bag-of-concept representation [19], which offers intensive document representation and takes word semantics into account, may be utilized to address the issues with word2vec representation.

In Bag-of-concepts representation as shown in figure 2, word vectors from documents are first trained using word2vec's skip-gram model. word vectors trained in a non-linear semantic vector space, in which semantically relations among words are effectively retained, using a neural network of context words to estimate the next word in a document. As a result, word vectors in the nearby semantic vector space are embedded with similar contextual information. When embedding vectors of word in a semantic vector space, vectors of words in adjoining semantic spaces are grouped into a common cluster. Because word2vec increases the cross-product of the embedding vectors and the context vectors, cosine distance was selected as an suitable metric for calculating semantic space distances among vectors of word and grouping close word vectors into a popular concept. Then, the spherical k-means algorithm is used to cluster vectors of word. Spherical k-means clustering for a fixed value of k. Spherical k-means, like any other clustering algorithm, indicate that grouping is susceptible to an initialization problem. The clustering results can differ between trials depending on the initial points used. To solve this problem, spherical k means were used with a large number of random starting points. Among the various clustered results, a clustering result with the lowest average within cluster distance was chosen. Each of the clusters produced by this process is then designated as a concept. As a result, words within a cluster will be labeled with a popular concept.

Document vectors are generated from these generated concepts. Because each word is now connected with a concept, the number of words in a document will correspond to the number of concepts. The numbers of these concepts are then used to represent a document vector. Words that appear frequently across many documents are not considered to be good markers for representing and grouping documents. Then, we implement concept frequency-inverse document frequency (CF-IDF) to vectors of document to highlight their discriminative concepts while removing the effect of consequently taking place but non-discriminative groups. CF-IDF, like the term frequency-inverse document frequency (TFIDF), minimize the effect of frequently occurring concepts using Equation 1.

$$CF - \mathrm{IDF}\left(c_i, d_j, D\right) = CF\left(c_i, d_j\right) \times \log \frac{|D|}{|d \in D; c_i \in d|} \tag{1}$$

where,

$$(c_i, d_j, D) = (\text{ Concept }_i,\ \text{Document }_j,\ \text{Corpus }) \\ |D| = \text{ Documents number in dataset} \\ |d \in D; t_i \in d| = \text{Documents number in dataset with Concept } t_i$$

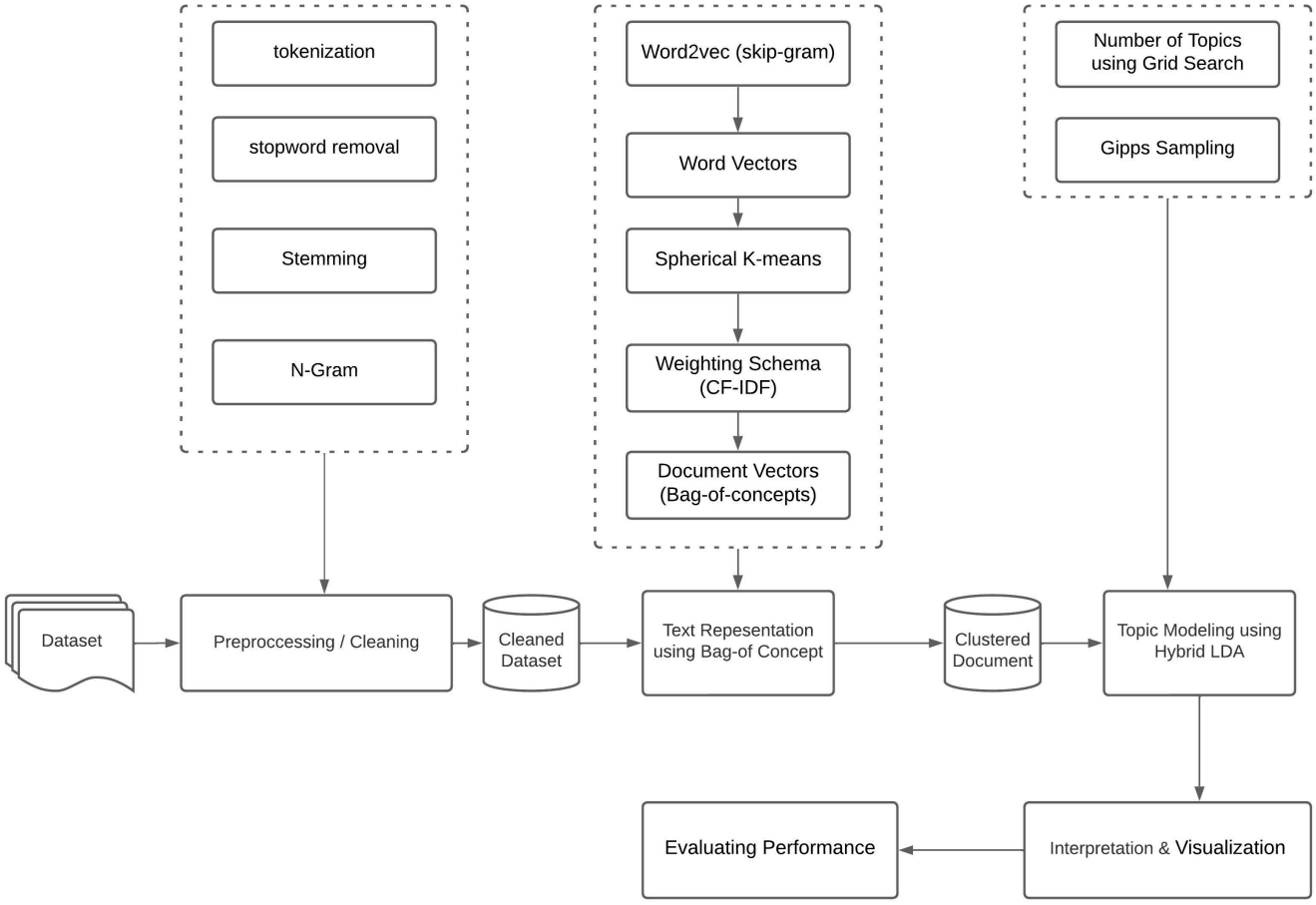

Fig. 1. The Flowchart of The Proposed Method

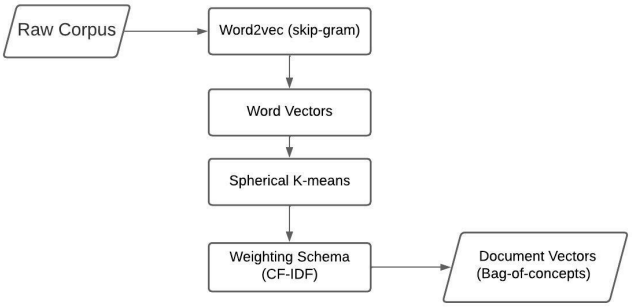

Fig. 2. The Flowchart of Bag-of-Concepts

C. Latent Dirichlet Allocation for Bag-of-concepts Representation

For topic modelling of text documents, latent Dirichlet allocation has already been commonly used. Every document in the dataset is assumed to be a mix of numerous topics in LDA. Each word in the document should be from a different topic. Every document has a fixed count of topics with a fixed number of vocabulary words, and every topic has a fixed number number of topics with a fixed number of vocabulary words. To produce a word in a document, the LDA generative process first produces a topic based on the document's likelihood distribution over topics, and then produces a word based on the likelihood, distribution of the chosen topic over vocabulary words.

The LDA's performance is hugely affected by document priors, which are the Dirichlet priors used to sample the topic distribution for each document. Because of this information, The model's performance can be improved by learning more relevant document priors. As a result, Gibbs sampling [20, 21] and Markov Chain Monte Carlo (MCMC) [22] can be used to approximate the intractable posterior of LDA. The topic for each word in each document is drawn from the conditional distribution of topics, which is given by the Equation 2.

$$p\left(z_{di}=k \mid W_{-i}, Z_{-i}\right)= \frac{n_{k,-i}^{w}+\beta}{\sum_{w=1}^{W}\left(n_{k,-i}^{w}+\beta\right)} \times \frac{n_{d,-i}^{k}+\alpha}{\sum_{k=1}^{K}\left(n_{d,-i}^{k}+\alpha\right)} \tag{2}$$

Where z_{di} is the assignment of topic for i^{th} document word

$\mathrm{d}, n^{w}_{k,-i}$ is the connt of occurrence word w show with topic $\mathrm{k}, n^{k}_{d,-i}$ is count of occurrence a word in document d is allocated topic k. ' $-$i ' points that the current word is not comprise in the counts. K is the topics and W is the vocabulary size. α and β are dirichlet hyper-parameter for word and topic allocations respectively.

LDA and Bag-of-concepts Inference Each word in LDA Interpretation is detected using binary weights. Words that are not existent in the document are not discovered (i.e., have a weight of zero) and words that are existent in the document have a weight of one. In Bag-of-concept document representation, all basic terms are delegated to all documents which have associateship degree. By the same way, LDA Interpretation method used for Bag-of-concept document representation with concept frequency-inverse document frequency (CF-IDF) weights. When a word w is discovered in document d, it actually adds to the common concept/cluster after representation using trained word2vec, where each concept contains similar words, and document vectors are generated based on these concepts.

IV. Experiments

A. Datasets

Measurements are carried out on Three datasets. The first one is a part of Reuters-21578 [23]. Only documents in Reuters-21578 with a non-empty topic feature are used. Some documents cover several topics. Only the first referenced topic is took into account for evaluation in such documents. The dataset used in the tests consist of documents and topics whose count is 9094 and 82.

Experiments on short text documents are also done to prove the quality of the bag-of-concepts representation. The dataset Snippets [24] is used, which contains 12340 web snippets with an average document length of 17.5. The dataset contains documents in eight different categories.

Another widely used publicly available dataset is the 20 Newsgroup dataset [25]. As shown, it contains 18,821 documents from 20 different classes.

TABLE I
Datasets Overview

Dataset	Train	Test	classes
Reuters-21578	7275	1819	90
20-Newsgroups	14400	3600	20
Snippets	10064	2276	8

B. Evaluation Criterion

1) Document Clustering: The first task is measuring document clustering quality. The proposed method's performance is measured by how well it clusters text documents. Cosine similarity between the vector representation of the document and cluster signatures is calculated to allocate a document to a cluster. Each document is assigned to the cluster with the highest grade of similarity. Purity and Normalized Mutual Information (NMI) [24] are used to get away directly mapping the calculated clusters with the ground truth clusters, as required in Precision and Recall. If the values of Purity and NMI are large, the clustering quality is good.

2) Quality of Topic: The second task is to assess topic quality, which is measured using the PMI score [26], which corresponds to human-judged topic coherence. The PMI score is calculated using point-by-point mutual information from an external knowledge base. In our test, we used an English Wikipedia with 1.3 million articles to calculate the PMI score. For each topic, the top 20 words are utilized to measure the PMI.

V. Results and Discussion

Table II displays the Purity, Normalized Mutual Information (NMI), and Pointwise Mutual Information (PMI) outcomes. these outcomes show that LDA topic models combined with representation of document from bag-of-concepts produce better topic allocation on short documents than topic models with vector representation based on binary and term weighted
.

TABLE II
results from Purity, NMI and PMI clustering evaluation

Method	Purity	NMI	PMI
Snippets dataset			
LDA	0.3826	0.1694	0.488
logLDA	0.4570	0.2577	0.476
FBLDA	0.5474	0.3749	0.600
Proposed Mothod	0.6531	0.4806	0.7057
Snippets dataset			
LDA	0.5353	0.3403	4.085
logLDA	0.5408	0.5408	3.93
FBLDA	0.6484	0.4366	4.643
Proposed Mothod	0.7541	0.5423	4.7487

References

[1] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[2] Y. Hu, K. Zhai, V. Eidelman, and J. Boyd-Graber, "Polylingual tree-based topic models for translation domain adaptation," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 1166–1176.

[3] N. Akhtar, N. Zubair, A. Kumar, and T. Ahmad, "Aspect based sentiment oriented summarization of hotel reviews," *Procedia computer science*, vol. 115, pp. 563–571, 2017.

[4] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 375–384.

[5] J. J. Palop, L. Mucke, and E. D. Roberson, "Quantifying biomarkers of cognitive dysfunction and neuronal network hyperexcitability in mouse models of alzheimer's disease: depletion of calcium-dependent proteins and inhibitory hippocampal remodeling," in *Alzheimer's Disease and Frontotemporal Dementia*. Springer, 2010, pp. 245–262.

[6] M. Wolf, A. Semm, and C. Erfurth, "Digital transformation in companies–challenges and success factors," in *International Conference on Innovations for Community Services*. Springer, 2018, pp. 178–193.

[7] D. Wang, S. Zhu, T. Li, and Y. Gong, "Multi-document summarization using sentence-based topic models," in *Proceedings of the ACL-IJCNLP 2009 conference short papers*, 2009, pp. 297–300.

[8] G. Yang, D. Wen, N.-S. Chen, E. Sutinen *et al.*, "A novel contextual topic model for multi-document summarization," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1340–1352, 2015.

[9] E. Amer and K. M. Fouad, "Keyphrase extraction methodology from short abstracts of medical documents," in *2016 8th Cairo International Biomedical Engineering Conference (CIBEC)*. IEEE, 2016, pp. 23–26.

[10] X. Li, A. Zhang, C. Li, J. Ouyang, and Y. Cai, "Exploring coherent topics by topic modeling with term weighting," *Information Processing & Management*, vol. 54, no. 6, pp. 1345–1358, 2018.

[11] K. Yang, Y. Cai, Z. Chen, H.-f. Leung, and R. Lau, "Exploring topic discriminating power of words in latent dirichlet allocation," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 2238–2247.

[12] T. Wang, Y. Cai, H.-f. Leung, Z. Cai, and H. Min, "Entropy-based term weighting schemes for text categorization in vsm," in *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2015, pp. 325–332.

[13] D. Yogish, T. Manjunath, and R. S. Hegadi, "Review on natural language processing trends and techniques using nltk," in *International Conference on Recent Trends in Image Processing and Pattern Recognition*. Springer, 2018, pp. 589–606.

[14] S. Sarica and J. Luo, "Stopwords in technical language processing," *Plos one*, vol. 16, no. 8, p. e0254937, 2021.

[15] S. Avasthi, R. Chauhan, and D. P. Acharjya, "Processing large text corpus using n-gram language modeling and smoothing," in *Proceedings of the Second International Conference on Information Management and Machine Intelligence*. Springer, 2021, pp. 21–32.

[16] A. A. Youssif, A. Z. Ghalwash, and E. Amer, "Kpe: an automatic keyphrase extraction algorithm," in *IEEE proceeding of international conference on information systems and computational intelligence (ICISCI 2011)*, 2011, pp. 103–107.

[17] E. Amer, "Enhancing efficiency of web search engines through ontology learning from unstructured information sources," in *2015 IEEE international conference on information reuse and integration*. IEEE, 2015, pp. 542–549.

[18] M. Xue, "A text retrieval algorithm based on the hybrid lda and word2vec model," in *2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*. IEEE, 2019, pp. 373–376.

[19] H. K. Kim, H. Kim, and S. Cho, "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation," *Neurocomputing*, vol. 266, pp. 336–352, 2017.

[20] D. Zhao, J. He, and J. Liu, "An improved lda algorithm for text classification," in *2014 International Conference on Information Science, Electronics and Electrical Engineering*, vol. 1. IEEE, 2014, pp. 217–221.

[21] T. Griffiths, "Gibbs sampling in the generative model of latent dirichlet allocation," 2002.

[22] D. Turek, P. de Valpine, and C. J. Paciorek, "Efficient markov chain monte carlo sampling for hierarchical hidden markov models," *Environmental and ecological statistics*, vol. 23, no. 4, pp. 549–564, 2016.

[23] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 42–49.

[24] C. D. Manning, *Introduction to information retrieval*. Syngress Publishing,, 2008.

[25] empty, "20 newsgroups dataset," empty. [Online]. Available: http://people.csail.mit.edu/jrennie/20Newsgroups/

[26] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 2010, pp. 100–108.