Joining LDA and Word Embeddings for Covid-19 Topic Modeling on English and Arabic Data

Amara Amina[a], Mohamed Ali Hadj Taieb[b] and Mohamed Ben Aouicha[c]
Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax, Tunisia

Keywords: Topic Modeling, Latent Dirichlet Allocation, Word Representation Learning, Covid-19, Multilingual.

Abstract: The value of user-generated content on social media platforms has been well established and acknowledged since their rich and subjective information allows for favorable computational analysis. Nevertheless, social data are often text-heavy and unstructured, thereby complicating the process of data analysis. Topic models act as a bridge between social science and unstructured social data analysis to provide new perspectives for interpreting social phenomena. Latent Dirichlet Allocation (LDA) is one of the most used topic modeling techniques. However, the LDA-based topic models alone do not always provide promising results and do not consider the recent advancement in the natural language processing field by leveraging word embeddings when learning latent topics to capture more word-level semantic and syntactic regularities. In this work, we extend the LDA model by mixing the Skip-gram model with Dirichlet-optimized sparse topic mixtures to learn dense word embeddings jointly with the Dirichlet distributed latent document-level mixtures of topic vectors. The embeddings produced through the proposed model were submitted to experimental evaluation using a Covid-19 based multilingual dataset extracted from the Facebook social network. Experimental results show that the proposed model outperforms all compared baselines in terms of both topic quality and predictive performance.

1 INTRODUCTION

The richness of social data has opened a new avenue for social science research to gain insights into human behaviors and experiences. However, the text-heavy, sparse, short, and unstructured nature of social content often leads to methodological challenges in both data collection and analysis (Amara et al., 2021). Consequently, there is a need for more efficient methods and tools that can help in detecting and analyzing content of such social platforms, particularly for those using user-generated content (UGC) as a source of data. In fact, a large number of online social websites produce a huge amount of data during emergency events including disease outbreaks, natural and human-induced disasters. This fact makes it very hard to understand and exploit the knowledge hidden in their UGC and as a result, there is a need to develop optimized techniques to automatically analyze such data. In this regard, topic models (Wang and Zhang, 2023) are used to bridge the gap between social science and unstructured analysis when dealing with the complex nature of such data. Basically, a topic model is a key technique that has been used to extract knowledge and latent topics from a set of documents, such as scientific articles. The core idea is that documents consist of multiple topics and each topic appears in different proportions in each document. In this new world, social media posts are a new type of documents and understanding the main topics of those online conversations requires the use of topic models. In particular, topic modelling was used to harness real-time communications through social platforms to monitor and respond to crisis and disaster management communication (Amara et al., 2021). The most popular and highly researched topic models are generative models such as LDA (Blei et al., 2003) and matrix factorization based models such as Non-negative Matrix Factorization (NMF) (Egger, 2022). Generative models are derived from the probability distribution of words and topics across the corpus of documents. Matrix factorization-based models decompose a high-dimensional data matrix into

[a] https://orcid.org/0000-0003-4257-6916
[b] https://orcid.org/0000-0002-2786-8913
[c] https://orcid.org/0000-0002-2277-5814

Amara, A., Hadj Taieb, M. and Ben Aouicha, M.
Joining LDA and Word Embeddings for Covid-19 Topic Modeling on English and Arabic Data.
DOI: 10.5220/0012320900003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 275-282
ISBN: 978-989-758-680-4; ISSN: 2184-433X

two lower-dimensional matrices with non-negative elements. However, these models tend to incur a high computational cost due to the calculation of either the posterior joint probability distribution or the matrix factorization. Due to the limitations of the above-mentioned models, the focus of research community was shifted towards a modification of those traditional models by incorporating more sophisticated Natural Language Processing (NLP) techniques into topic models to improve their predictive power. Along this direction, the present work aims to take advantage of distributed word representations to build document-level distributed representations by joining LDA with Skip-gram model (Mikolov et al., 2013a), the well-known NLP model for learning distributed word representations, in order to capture more coherent Covid-19 topics on English and Arabic posts extracted from Facebook.
The paper is organized as follows. We review the related work in Section 2. Section 3 introduces the proposed approach used for Covid-19 multilingual topic modeling based on Facebook posts. An in-depth exploration of the used dataset, baseline methods, along with the process of data analysis and experimental results presentation are detailed in Section 4. Conclusions and future work are made in Section 5.

2 RELATED WORKS

Numerous studies have been conducted on social media to analyze and mine helpful information about the most discussed topics during the Covid-19 pandemic. In this context, the following two sub-sections briefly discuss research on two main branches: topic modeling approaches and Covid-19 trends analysis on social media.

2.1 Topic Modeling Approaches

Understanding the core topics associated with a corpus of documents is a fundamental task in today's big data era. Topic models are a class of unsupervised machine learning techniques designed for this task (Laureate et al., 2023). They focus on the extraction of hidden topics from a corpus of documents. Labeling the documents of the corpus with these topics allows users to understand the importance of the topic in each document and the collection as a whole. Conventional topic modeling approaches such as NMF (Egger, 2022), Latent Semantic Analysis (LSA) (Deerwester et al., 1990), Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999), and LDA (Blei et al., 2003) have soared in popularity across various domains in the past years to discover hidden semantic topics from large corpora of documents. Both NMF and LSA are non-probabilistic approaches based on algebraic models and more precisely matrix factorization. The latter uses to convert the high-dimensional corpus (term-document matrix) into two lower-dimensional matrices: the first matrix is for the documents grouped by topics (document-topic distribution) while the second one is for words grouped by topics (word-topic distribution). Previous mentioned models only considered the count of terms and ignored word co-occurrence, which encodes more semantic meaning for inferred topics. Probabilistic models were created to solve these problems and improve non-probabilistic approaches by introducing the probability sense through generative models (Blei et al., 2003). In this context, PLSA is a latent variable model that uses a term-document matrix of co-occurrence to detect semantic co-occurrence of terms in a set of documents. To fill the gaps in PLSA's topic probability distribution, LDA utilizes the Bayesian approach to build a probability model at the document level. However, the documents that are inputted into topic models have changed much more drastically than the topic models themselves. Historically, topics models have been used to extract topics from scientific articles, books, and newspaper articles, but they have recently been used to identify hidden topics in social media posts and other short texts. Research on topic modeling has evolved to address the challenges of inferring topics on these new types of data by incorporating more sophisticated NLP techniques, such as word embedding vectors (e.g., Word2vec) and deep neural networks (Laureate et al., 2023), into topic models. For instance, Thompson and Mimno (Thompson and Mimno, 2020) proposed using the language model BERT to produce topics. Indeed, BERT is a language model, originally proposed in 2018, that uses the Transformer architecture to generate high-quality representations of words. Such models rely on a new vector representations learned from an encoder/decoder network.

2.2 Covid-19 Trends Analysis on Social Media

Social media such as Facebook and Twitter have proven to be a useful resource in understanding public opinion towards real world events. Millions of discussions or posts on social media are generated every day due to the COVID-19 pandemic outbreak

Table 1: Topic modeling studies based on Covid-19 data.

Related work	Source	Collected data	Time	Language	Used Model
(Xue et al., 2020)	Twitter	1,963,285 tweets	From January 23 to March 7, 2020	English	LDA
(Garcia and Berton, 2021)	Twitter	6,487,842 tweets	From April 17, 2020 to August 08, 2020	English and Portuguese	LDA
(Amara et al., 2021)	Facebook	46,091 posts	From January 1st, 2020 to May 15th, 2020	English, Arabic, Spanish, Italian, German, French and Japanese	LDA
(Cinelli et al., 2020)	Twitter, Instagram, YouTube, Reddit and Gab	1,342,103 posts and 7,465,721 comments	From the 1st of January to the 14th of February	English	Partitioning Around Medoids algorithm with Skip-gram model
Our study	Facebook	39,211 posts	From January 1st, 2020 to May 15th, 2020	English and Arabic	LDA with Skip-gram model

(Amara et al., 2021). Such a large number of user-generated posts provide a valuable source of data and thus receive great attention from researchers (Chen et al., 2023). Topic modeling techniques have been employed by researchers not only to understand aspects of the Covid-19 pandemic but also to understand public attitudes and behaviors during this crisis. This is considered as a new way to support crisis communication and health promotion messaging. Most of the Covid-19 related works (Xue et al., 2020; Cinelli et al., 2020) explored only English data where most of them are retrieved from Twitter social network as shown by Table 1. For instance, Xue et al. (Xue et al., 2020) analyzed 1.9 million English tweets related to coronavirus collected from January 23, 2020 to March 7, 2020. They used the LDA model to identify a total of 11 topics which are then categorized into ten themes. In addition, the Covid-19 situation inspired researchers to conduct some other studies either on heterogeneous datasets considering multiple languages (Amara et al., 2021; Garcia and Berton, 2021) or on other social media platforms such as Reddit (Basile et al., 2021), Facebook (Amara et al., 2021), and even on Covid-19 related data shared between multiple social platforms. In fact, Garcia and Berton (Garcia and Berton, 2021) examined Covid-19 related Twitter data collected from April to August 2020 in two different languages, English and Portuguese primarily from the USA and Brazil. They identified ten topics for both languages and then examined the relationship between the identified topics and the feelings thus created. Amara et al. (Amara et al., 2021) leveraged Facebook posts in 7 different languages for tracking the evolution of Covid-19 related trends during the period spanned from January 1st, 2020 to May 15, 2020. They also exploit the Latent Dirichlet Allocation for extracting the most discussed topics in this period.

Most of the above cited works typically utilize the LDA topic model and do not take advantage of word embedding-based approaches to capture more semantic information in produced topics. Such approaches introduce a new perspective to topic modeling, taking a different direction from the previous traditional models. Therefore, joining the LDA topic modeling approach with the Skip-gram model can help capture additional semantics by learning latent representations for words, documents, and even topics.

3 MULTILINGUAL TOPIC MODELING

This section presents the proposed model for topic modeling on Covid-19 data. The first part describes the LDA model and the second part details the combination of the LDA model with the Skip-gram, a variant of Word2vec (Mikolov et al., 2013a).

3.1 LDA Topic Model

LDA is a generative model commonly used in topic modeling to extract high-level semantic latent topics from a corpus of unstructured texts. The term "latent" in LDA is related to the discovery of hidden semantics from a huge collection of text documents. The only assumption by LDA is that each document can be represented as a probabilistic distribution over topics and the topic distribution in all documents shares a common Dirichlet prior. A latent topic in LDA is also defined by a probabilistic distribution of words from the vocabulary and these word-topic distributions share a common Dirichlet prior as well. More specifically, the Dirichlet distribution controls the mixture proportions of the topics in the documents and the words in each topic. In this study, we focus on public Covid-19 posts collected from Facebook. Indeed, each post represents a text document and we exploit LDA to discover the hidden topics discussed in these public posts. LDA considers a document as a finite mixture of latent topics where each topic is defined by a distribution of words from the vocabulary. In figure 1, given a corpus D composed of $\mathcal{M}$ documents and each document consists of $\mathcal{N}$ words. The multinomial distribution θ represents the per-document topic distribution having an hyper-parameter α which follows the Dirichlet distribution. Similarly, the multinomial distribution ϕ refers to the per-topic word distribution characterized by the hyper-parameter β following the Dirichlet distribution for the fixed number of topics denoted as K topics.

Despite the popular use of LDA, it presents a major drawback. It learns a document vector that predicts words within the document without considering their structure or how these words interact on a local level. In other words, LDA model treats a document as a bag of words and it does not use the "*context*" that capture how words are ordered and grouped together.

3.2 Joining LDA with Skip-Gram Model

The aforementioned topic modeling approach, i.e., LDA, does not take advantage of the recent advancement in the NLP field by leveraging distributed representations of words when extracting latent topics from a given textual corpus. The aim is to embed both words and document representations into the same latent space and train both representations simultaneously by mixing Skip-gram model with Dirichlet-optimized sparse topic mixtures. In fact, document vectors are generated from a mixture of topic vectors in order to represent the topic tendency of a given document. Formally, a document vector $\overrightarrow{d}_j$ is calculated as a weighted sum of topic vectors:

$$\overrightarrow{d}_j = \sum_{k=1}^{K} p_{jk} \overrightarrow{t}_k \quad (1)$$

where K denotes the number of topics, p_{jk} represents the weight of topic k on document j, $0 \leq p_{jk} \leq 1$, and $\overrightarrow{t}_k$ denotes the k-th topic vector.

During the training process, document vector was updated by the p_{jk} weights that are normalized to ensure $\sum_k p_{jk} = 1$. These weights are optimized using a Dirichlet likelihood with a low concentration parameter α:

$$\mathcal{L}^d = \gamma \sum_{jk} (\alpha - 1) p_{jk} \quad (2)$$

where $\mathcal{L}^d$ measures the likelihood of document j in topic k summed over all the documents and γ denotes the strength of $\mathcal{L}^d$ in the training process.

In Skip-gram model, a pivot word vector is leveraged to predict a set of context words which are selected using a sliding window moving behind and after the pivot word. This word vector is exploited with the aforementioned document vector to create a context vector forming a semantically meaningful combination of both word and document in order to capture the long- and short-term themes. The context vector $\overrightarrow{c_j}$ is explicitly defined as the sum of the pivot word vector $\overrightarrow{w_j}$ and document vector $\overrightarrow{d_j}$ as follows:

$$\overrightarrow{c_j} = \overrightarrow{w_j} + \overrightarrow{d_j} \quad (3)$$

To jointly train the context vector and topic-enhanced word vectors, we attempt to alter the Negative Sampling approach used in the Skip-gram model (Mikolov et al., 2013b) which tries to differentiate the target words from the negative ones picked randomly from a noise distribution. To separate target words from negative ones, the loss function $\mathcal{L}_{ij}^{neg}$ is minimized by the following equation:

$$\mathcal{L}_{ij}^{neg} = log\sigma(\overrightarrow{c}_j \cdot \overrightarrow{w}_i) + \sum_{l=0}^{n} log\sigma(-\overrightarrow{c}_j \cdot \overrightarrow{w}_l) \quad (4)$$

where $\overrightarrow{c}_j$ denotes the context vector, $\overrightarrow{w}_i$ denotes the target word vector, and $\overrightarrow{w}_l$ represents the latent vector of a given negative word randomly picked from n negative samples.

The total loss term of the model is defined by the sum of the Skip-gram Negative Sampling loss function $\mathcal{L}_{ij}^{neg}$ and the Dirichlet likelihood $\mathcal{L}^d$ defined above:

$$\mathcal{L} = \mathcal{L}^d + \sum_{ij} \mathcal{L}_{ij}^{neg} \quad (5)$$

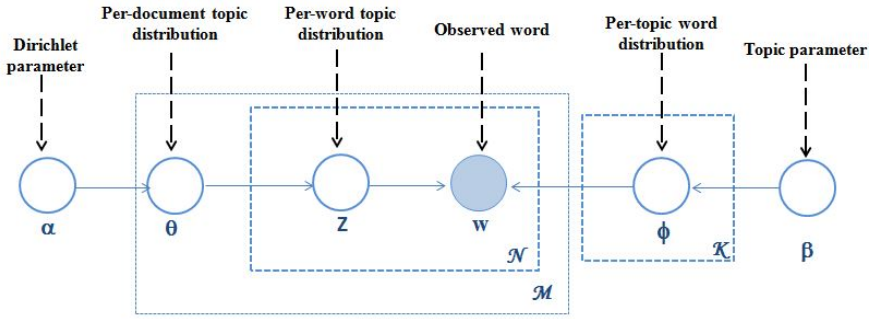


Figure 1: LDA graphical model.

In this study, the corpus consists of public posts collected from Facebook and written in two different languages: Arabic and English languages. Thus, each public post represents a text document in the corpus. Thereafter, the proposed model is applied to the English and Arabic corpora to discover discussed topics. Then, we compare the output of both models to demonstrate the importance of word representation learning in capturing more hidden semantics and extracting more coherent topics from a given corpus.

4 EXPERIMENTS

This section includes a detailed description of the used dataset and its preprocessing steps for the experimental study. Subsequently, a set of baseline methods is presented along with evaluation metrics. At last, the experimental results are provided in detail.

4.1 Data Preprocessing and Statistics

Data collection was conducted from January 1st, 2020 to May 15th, 2020 using a customized web crawler to collect public Covid-19 posts, based on a set of predefined keywords, written in Arabic and English languages from Facebook. Thereafter, the dataset was preprocessed to clean up the unstructured data and then convert the extracted information into a structured format to analyze the patterns. Corpus preparation and cleaning were done using a series of packages running on top of Python using the Natural Language Toolkit (NLTK)[1] which supports multiple languages. The procedure includes the following steps for Arabic and English languages: stop-words, URLs, emojis and punctuation removal, tokenization, lemmatizing, stemming, identifying n-gram procedures, and lowercase transformation for English posts. As shown by Fig. 2, the distribution of posts length is right-skewed with a high concentration of posts containing a number of words between 30 and 120 words for English and Arabic languages. We removed additional posts with words' length less than 10 words, since they usually cannot provide reasonable semantics. At the end, we got a dataset of 26,320 arabic posts and the second dataset contains 12,891 english posts. Moreover, the word cloud visualization is used to give the bird's eye view of what the major terms used in the two datasets. This means, we could see what are the most significant words which are frequently discussed during the outbreak of Covid-19 virus as shown by Fig. 3.

[1]https://www.nltk.org/

4.2 Baseline Models

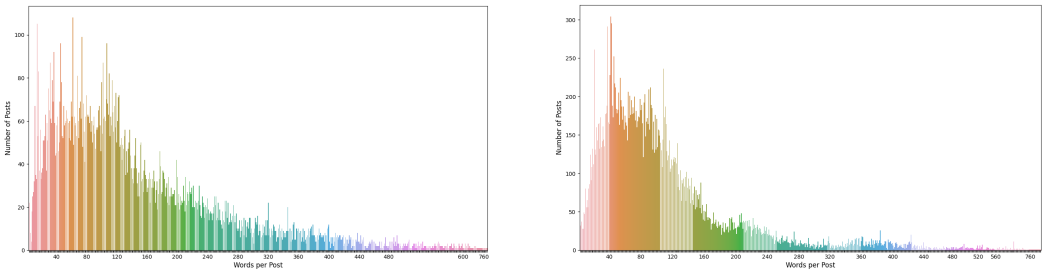
Three baseline topic models are used in this study:

- LDA (Blei et al., 2003): a generative probabilistic model widely used to extract topics from a corpus of documents, where each document is represented as a finite mixture over a set of topics, and each topic is represented as an infinite mixture over a collection of topic probabilities.
- NMF (Egger, 2022): is non-probabilistic algorithm using matrix factorization and works on TF-IDF transformed data by breaking down a matrix into two lower ranking matrices.
- BERTopic (Thompson and Mimno, 2020): provides continuous rather than discrete topic modeling using a sentence-transformers model for more than 50 languages.

4.3 Evaluation Metrics

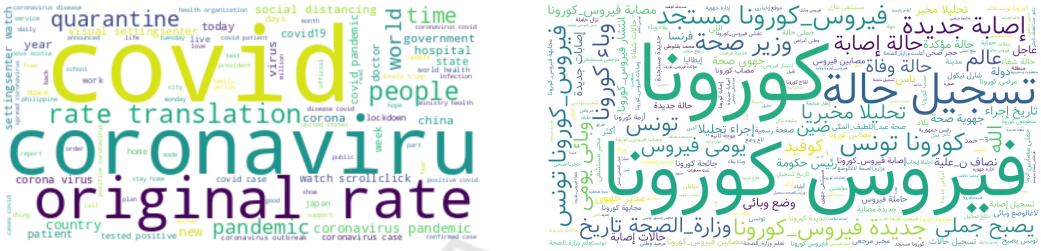
To evaluate the quality of produced topics, we use a set of topic coherence metrics to score the degree of semantic similarity between high scoring words in a topic.

- C_v is based on four parts: (i) segmentation of the data into word pairs, i.e., S_i a segmented



(a) English dataset (b) Arabic dataset

Figure 2: Post length distribution.



(a) English dataset (b) Arabic dataset

Figure 3: WordCloud of the most prevalent words used in the two datasets.

pair of each word $W' \in W$ paired with all other words $W^* \in W$, (ii) calculation of word pair probabilities, (iii) calculation of a confirmation measure that quantifies how strongly a word set supports another word set, i.e., W' and W^* are represented as context vectors to capture the semantic support of all the words in W, and this confirmation measure is quantified by computing the cosine vector similarity of all context vectors $\phi_{S_i}(\vec{u}, \vec{w})$ within S_i, with $\vec{v}(W') \in \vec{u}$ and $\vec{v}(W^*) \in \vec{w}$ as expressed in 8 and finally (iv) aggregation of individual confirmation measures into an overall coherence score.

$$\vec{v}(W') = \{ \sum_{w_i \in W'} NPMI(w_i, w_j)^{\gamma} \}_{j=1..N} \quad (6)$$

$$NPMI(w_i, w_j) = \frac{log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-logP(w_i, w_j) + \varepsilon} \quad (7)$$

$$\phi_{S_i}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{N} u_i \cdot w_i}{||\vec{u}||_2 \cdot ||\vec{w}||_2} \quad (8)$$

where $P(w_i)$ denotes the probability of seeing a word w_i in a random document, $P(w_i, w_j)$ represents the probability of co-occurrence of w_i and w_j in a random document, and N is the number of selected top probable words, ε is used to account for the logarithm of zero and γ to give more weight on higher NPMI values.

- C_{UCI} computes the Pointwise Mutual Information (PMI) for every pair of words. The arithmetic mean of all computed values of PMI is the final result.

$$C_{UCI} = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI(w_i, w_j) \quad (9)$$

$$PMI(w_i, w_j) = log \frac{P(w_i, w_j) + \varepsilon}{P(w_i)P(w_j)} \quad (10)$$

where $P(w_i)$ represents the probability of seeing w_i in a random document, $P(w_i, w_j)$ represents the probability of co-occurrence of w_i and w_j in a random document from an external reference corpus, e.g., Wikipedia, N is the number of selected top probable words, and ε is added to avoid logarithm of zero.

- C_{UMass} defines the score to be based on document co-occurrence. The main idea behind this measure is that the occurrence of every top word should be supported by every preceding top word.

$$C_{UMass} = \frac{2}{N \cdot (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} log \frac{D(w_i, w_j) + \varepsilon}{D(w_j)} \quad (11)$$

where $D(w_i, w_j)$ counts the number of documents containing words w_i and w_j and $D(w_i)$ counts the number of documents containing w_j, N is the number of selected top probable words, and ε is added to avoid logarithm of zero.

4.4 Experimental Results and Analysis

To pinpoint optimal values for the three hyperparameters required for the proposed method, a grid search was performed for the number of topics (K) as well as for beta and alpha. The higher the beta, the more words the topics consist of; likewise, the higher the alpha, the more diverse the topics are.

The number of epochs for the proposed model is fixed to 650 epochs and word vectors are initialized to the pretrained values used by Mikolov et al. (Mikolov et al., 2013b). The search for an optimal number of topics started with a range from 5 to 35, with a step of five. In the first step of the learning process, K was pre-defined, and the search for beta and alpha was applied accordingly. During the process, only one hyperparameter varied, and the other remained unchanged until reaching the highest coherence score. The grid search then yielded the values of 0.1 and 0.02 for alpha and beta parameters, respectively. As shown by Fig. 4, the coherence score is optimal for K=25 topics with an optimal value of 0.61 for English corpus and 0.71 for the Arabic dataset.

We evaluate the topic quality and performance of the proposed model against three baseline models. For this reason, as shown by Table 2, the evaluation metrics for these models were computed to assess their performance. We observe that during the evaluation, the results of all the methods performed similarly. Negative values presented in the table indicate coherent topics (Tijare and Rani, 2020). Briefly, by comparing the outcomes of the coherence metrics, the proposed model produced the highest coherence value for the three used metrics; NMF and BERTopic provided similar performance; and LDA is the worst compared to other methods. Indeed, as LDA extracts independent topics from word distributions, topics that are deemed dissimilar in the document may not be identified separately.

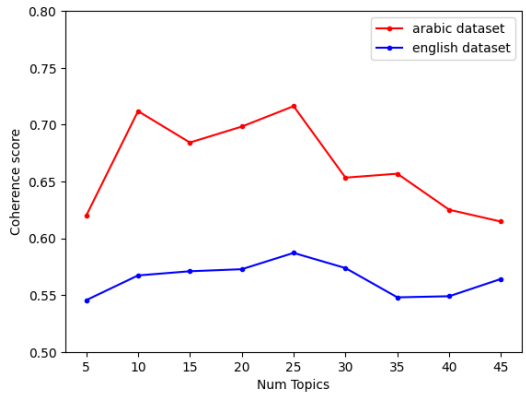


Figure 4: Coherence scores corresponding to the different number of topics.

Moreover, the proposed model and NMF produce higher quality topics and more coherent topics than the other examined methods for both Arabic and English datasets. But the proposed model was more flexible and provided more meaningful and logical extracted topics that match our final aim of defining a topic model that can understand and efficiently analyze the online UGC. A major drawback of NMF revolves around its low capability to identify embedded meanings within a collection of texts.

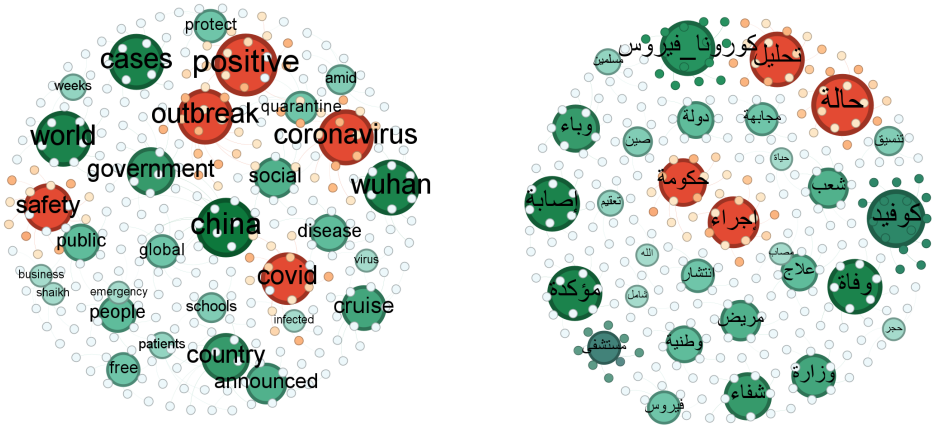
Table 2: Evaluation metrics on topic models.

Model	English dataset			Arabic dataset		
	C_v	C_{UMASS}	C_{UCI}	C_v	C_{UMASS}	C_{UCI}
LDA	0.40	-12.42	-8.08	0.41	-7.11	-4.52
NMF	0.56	-5.14	-1.22	0.67	**-3.28**	-0.19
BERTopic	0.53	-8.87	-2.29	0.52	-9.45	-3.24
Our model	**0.61**	**-2.88**	**0.33**	**0.71**	-3.30	**-0.17**

Regarding the BERTopic model, it might actually generate more outliers than expected. Meanwhile, another flaw involves topic distributions: they cannot be generated within a single document because each document is assigned to a single topic. Moreover, baselines do not allow for an in-depth understanding of both English and Arabic Covid-19 datasets and the proposed model was able to generate novel insights using its embedding approach. The obtained results of the proposed model are visualized in graph structure using the Gephi, as open graph visualization platform, according to the importance degree of the extracted topics linked to the posts having mostly participated to form the given topic. Fig. 5 shows the extracted topics involved in the two Facebook datasets in relation with English and Arabic languages, using the proposed model. It is worth mentioning that the extracted topics for both languages reflects more interesting insights explored based on the proposed model in contrast to the world clouds generated from the two datasets as shown by Fig. 3.

5 CONCLUSION

The present work demonstrates an analysis of an extended version of the popular LDA topic model by combining the LDA with the popular NLP Skip-gram model to take advantage from the recent advances in the field of word representation learning. To capture more semantic between discovered topics, the standard LDA topic model extends the Skip-gram model to leverage distributed representations of words along with document representations while



(a) English COVID-19 related topics (b) Arabic COVID-19 related topics

Figure 5: Extracted topics using the proposed model.

preserving semantic regularities between the learned word vectors when extracting latent topics from a collection of social data. The proposed model has been evaluated against a set of baseline models using English and Arabic Covid-19 datasets collected from Facebook. From experimental results, it is clear that the proposed model outperforms all baselines. Topics discovered using the proposed model yield high scores for the three used evaluation metrics and these results showed that topics discovered by this model are more coherent and human interpretable. Finally, this model can be used for system recommendations on social networks, text mining, and other statistical analysis on a corpus of unstructured texts.

REFERENCES

Amara, A., Hadj Taieb, M. A., and Ben Aouicha, M. (2021). Multilingual topic modeling for tracking covid-19 trends based on facebook data analysis. *Applied Intelligence*, 51:3052–3073.

Basile, V., Cauteruccio, F., and Terracina, G. (2021). How dramatic events can affect emotionality in social posting: The impact of covid-19 on reddit. *Future Internet*, 13(2):29.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Chen, Y., Sherren, K., Smit, M., and Lee, K. Y. (2023). Using social media images as data in social science research. *New Media & Society*, 25(4):849–871.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., and Scala, A. (2020). The covid-19 social media infodemic. *Scientific reports*, 10(1):1–10.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Egger, R. (2022). Topic modelling: Modelling hidden semantic structures in textual data. In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, pages 375–403. Springer.

Garcia, K. and Berton, L. (2021). Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa. *Applied soft computing*, 101:107057.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Laureate, C. D. P., Buntine, W., and Linger, H. (2023). A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review*, pages 1–33.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Thompson, L. and Mimno, D. (2020). Topic modeling with contextualized word representation clusters. *arXiv preprint arXiv:2010.12626*.

Tijare, P. and Rani, P. J. (2020). Exploring popular topic models. In *Journal of Physics: Conference Series*, volume 1706, page 012171. IOP Publishing.

Wang, J. and Zhang, X.-L. (2023). Deep nmf topic modeling. *Neurocomputing*, 515:157–173.

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., and Zhu, T. (2020). Public discourse and sentiment during the covid 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter. *PloS one*, 15(9):e0239441.