 doi:10.20944/preprints202204.0026.v1

SATLabel: A Framework for Sentiment and Aspect Terms Based Automatic Topic Labeling

Khandaker Tayef Shahriar[1,*], Mohammad Ali Moni[2], Mohammed Moshiul Hoque[1], Muhammad Nazrul Islam[3], Iqbal H. Sarker[1,*]

[1]Department of Computer Science and Engineering, Chittagong University of Engineering & Technology, Chittagong-4349, Bangladesh.
[2]Artificial Intelligence & Digital Health Data Science, School of Health and Rehabilitation Sciences, Faculty of Health and Behavioural Sciences, The University of Queensland St Lucia, QLD 4072, Australia.
[3]Department of Computer Science and Engineering, Military Institute of Science and Technology, Dhaka-1216, Bangladesh.
*Correspondence: u17mcse008p@student.cuet.ac.bd, iqbal@cuet.ac.bd

Abstract. In this paper, we present a framework that automatically labels Latent Dirichlet Allocation (LDA) generated topics using sentiment and aspect terms from COVID-19 tweets to help the end-users by minimizing the cognitive overhead of identifying key topics labels. Social media platforms especially Twitter are considered as one of the most influential sources of information for providing public opinion related to a critical situation like the COVID-19 pandemic. LDA is a popular topic modelling algorithm that extracts hidden themes of documents without assigning a specific label. Thus automatic labelling of LDA-generated topics from COVID-19 tweets is a great challenge instead of following the manual labelling approach to get an overview of wider public opinion. To overcome this problem, in this paper, we propose a framework named `SATLabel` that effectively identifies significant topic labels using *top unigrams features of sentiment terms and aspect terms clusters from LDA generated topics* of COVID-19 related tweets to uncover various issues related to the COVID-19 pandemic. The experimental results show that our methodology is more effective, simpler, and traces better topic labels compare to the manual topic labelling approach.

Keywords: Data-driven Framework · LDA · Sentiment Terms · Aspect Terms · Unigrams · Soft Cosine Similarity · Topic · Automatic labeling

1 Introduction

Twitter nowadays is considered as one of the most important social media platforms to explain the characteristics and predict the status of the pandemic [9]. In Wuhan, at the end of 2019, a novel coronavirus disease that causes COVID-19 was reported by the World Health Organisation (WHO). The declaration of COVID-19 as an international concern of public health emergency by WHO

 doi:10.20944/preprints202204.0026.v1

was reported on January 30, 2020 [1]. During the pandemic, the use of Twitter increases immensely and plays a critical role by reflecting real-time public panic and providing rich information to raise public awareness through posts and comments. However, text mining and analysis of data from social media platforms such as Twitter have become a burning issue to extract necessary information. Moreover, it is a great challenge to extract meaningful topic labels by machines instead of following diverse human interpretations of the manual labelling approach [**?**]. Hence, in this paper, we propose `SATLabel`, a framework that effectively identifies key topic labels of tweets automatically from the huge volume of the Twitter dataset to reduce the human effort of cumbersome topic labelling tasks.

A large number of labelled datasets is required for traditional supervised methods. Obtaining such a labelled dataset for topic labelling purposes is very difficult and expensive. In this paper, we use LDA [3], which is an unsupervised probabilistic algorithm for text documents. Thus `SATLabel` does not need any labelled dataset for topics. A set of topics available in the documents is discovered by LDA. Sentiment terms express emotions from tweets and Aspect terms describe features of an entity [19]. We create sentiment terms cluster and aspect terms cluster for each LDA generated topic. However, Unigram is a probabilistic language model that is extensively used in natural language processing tasks and text mining to exhibit the context of texts. `SATLabel` uses the top Unigrams features from sentiment terms cluster and aspect terms cluster respectively and create attribute tags concatenating the two top Unigrams features (first sentiment term and then aspect term). We select the attribute tag which has the highest soft cosine similarity value with respect to the tweets of the same topic to assign a meaningful label for that LDA-generated topic. Our experimental results show that the label generated by `SATLabel` has a high soft cosine similarity value with the tweets of the same topic than the manual labelling approach. The main contributions of this paper can be summarized as follows:

- We effectively utilize sentiment terms and aspect terms of tweets to produce significant topic labels.
- We propose a new framework named `SATLabel` that is useful to extract topics from COVID-19 tweets and labels them automatically instead of following the manual method.
- `SATLabel` effectively reduces the human effort for difficult topic labelling tasks of tweets.
- We have shown the effectiveness of `SATLabel` comparing with the manual labelling approach by conducting a range of experiments.

The organization of the rest of the paper is as follows. Related works are reviewed in section 2. In section 3, we present the methodology of the proposed framework. In section 4, we assess the evaluation results of our framework by conducting experiments on the Twitter dataset. Next, we present the discussion, and finally, we conclude this paper and highlight the direction of future work.

2 Related Work

COVID-19 tweets can be helpful for identifying meaningful topic labels to highlight user conversation and understand ideas of people's needs and interests. Many researchers used the LDA algorithm to extract hidden themes of documents. Patil et. al. [12] proposed a paper using the frequency-based technique to extract topics from people's reviews without mentioning the proper labelling techniques for describing the topics. Hingmire et. al. [7] proposed a paper to construct LDA based topic model but the expert association is required to assign the topic to the class labels. Hourani et. al. [8] proposed a paper to classify articles according to their topics for which labelled dataset is required. Asmussen et. al. [2] proposed a topic modelling method for researchers but topic labelling depends on the researcher's view without having any automatic method. Wang et. al. [18] proposed a paper that minimizes the problem of data sparsity without labelling key topics specifically. Zhu et. al. [20] presented the change of the number of texts on topics with respect to time by following the manual topic labelling approach. Satu et. al. [15] proposed a framework that extracts topics from the best cluster of sentiment classification having a manual explanation of topic labels tends to misinterpretation. Kee et. al. [10] used LDA to extract higher-order arbitrary topics but only 61.3% clear collective themes were evaluated. Maier et. al. [11] presented accessibility and applicability of communication researchers using LDA based topic labelling approach which depends manually on broader context knowledge. In our previous work, we only considered the top unigram feature of aspect terms cluster to identify the key topics with labels by implementing LDA [17]. Elgesem et. al. [5] presented an analysis about the discussion of the Snowden affair using a manual topic labelling approach. Guo et. al. [6] compared dictionary-based analysis and LDA analysis using a manual topic labelling approach.

The summary of the above works describes that most of the works considered a manual topic labelling approach to categorize documents and get an overview which is expensive, time-consuming, and requires cumbersome human interpretations. Hence, an automatic and effective topic labelling approach would be helpful to reduce human effort and save time. Thus in this paper, we consider the development of a framework named `SATLabel` to generate significant topic labels automatically to highlight users' conversations on Twitter.

3 Methodology

In this section, we present `SATLabel` that is a framework to label LDA generated topics automatically as shown in Fig. 1. For analyzing and mining textual data like tweets, text preprocessing is one of the most essential steps to advance in further processing steps. The working principle and overall steps to generate automatic topic labels from the Twitter dataset are shown in Algorithm 1. After preprocessing of highly unstructured and non-grammatical tweets, several processing steps are followed to produce the expected output.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 April 2022 doi:10.20944/preprints202204.0026.v1

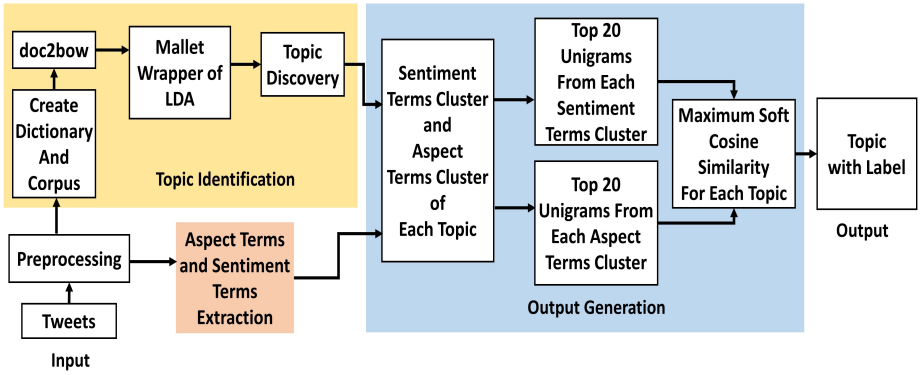

Fig. 1. `SATLabel`: Proposed framework for automatic topic labeling

Algorithm 1: Automatic Topic Labeling

Input: T: number of Tweets in dataset
Output: Topic Label (T_{Label}).

for ***each*** $t \in \{1,2,...,T\}$ **do**
  $T_p \leftarrow Preprocess(t)$;
for ***each*** $t_p \in \{1,2,...,T_p\}$ **do**
  `// Corpus Development`
  $C \leftarrow Create\_Corpus(t_p)$;
  `// Sentiment Terms and Aspect Terms Extraction`
  $S_{T_p} \leftarrow Sentiment\_Terms(t_p)$;
  $A_{T_p} \leftarrow Aspect\_Terms(t_p)$;
`// Topic Discovery`
$K \sim Mallet(LDA(Doc2bow(C)))$;
for ***each*** $k \in \{1,2,...,K\}$ **do**
  for ***each*** $t_p \in \{1,2,...,T_p\}$ **do**
    $k_{dominant,t_p} \sim dominant\_topic(t_p, k)$;
    `// Create Clusters from Topic`
    $C_S \sim Cluster(S_{T_p} \rightarrow k_{dominant,t_p})$;
    $C_A \sim Cluster(A_{T_p} \rightarrow k_{dominant,t_p})$;
for ***each*** $k \in \{1,2,...,K\}$ **do**
  $U_S \sim max\_count(Top\_Unigrams(C_S \rightarrow k))$;
  $U_A \sim max\_count(Top\_Unigrams(C_A \rightarrow k))$;
  $T_{Label} \leftarrow max\_soft\_cosine\_similarity(U_S + U_A, k)$

3.1 Sentiment and Aspect Terms Extraction

Sentiment terms carry the tone or opinion of the text. Usually, adjectives and verbs of sentences are considered as sentiment terms that indicate expressed

opinion of the text. Noun and noun phrases are considered as aspects terms of text. Objects of verbs are often regarded as aspect terms that describe the features of an entity, product, or event [19]. We follow precise parts of speech tagging which is an efficient approach to extract sentiment terms and aspect terms from texts. Examples of sentiment terms and aspect terms of sample tweets are shown in Table 1.

Table 1. Example of Sentiment Terms and Aspect Terms

Sample Tweet	**Sentiment Terms**	**Aspect Terms**
Please read the thread.	read	thread
To enjoy and relax for your dinner it is a great place.	enjoy, relax, great	dinner, place
Links with info on communicating with children regarding COVID-19.	communicate, covid	links, info, children
The retail store owners right now	retail, right	owners, store

3.2 Topic Identification Using LDA

LDA is a popular topic modeling algorithm to discover hidden topics available in the corpus from unlabelled dataset [14]. But the challenge is how to assign significant labels to LDA-generated topics. The steps for topic discovery that we follow in `SATLabel` framework are discussed below:

1) Creating Dictionary and Corpus: A systematic way of creating a number of lexicons of a language is supported by a dictionary and a corpus generally refers to an arbitrary sample of that language. A document corpus is built with words or phrases. In Natural Language Processing (NLP) paradigm, the corpus of a language plays a vital role in developing a knowledge-based system and mining texts. In the proposed framework, we create a dictionary and develop a corpus from the preprocessed text.

2) Creating a BoW Corpus: Corpus contains the word id and its frequency in every document. Documents are converted into Bag of Words (BoW) format by applying Doc2bow embedding. Each word is assumed as a normalized and tokenized string.

3) Topic Discovery: BoW corpus is transferred to the mallet wrapper of LDA. The presence of a set of topics in the corpus is discovered by LDA. Mallet wrapper of LDA runs faster and provides precise division of topics using Gibbs Sampling technique [4]. LDA generates the most prominent words in a topic. Thus by using the word probabilities one can manually find dominant themes in the documents. To overcome the complex manual labeling approach our framework `SATLabel` generates automatic topic labels using

sentiment terms and aspect terms of documents without any human interpretation. Based on the topic coherence score, we choose a model that discovers 20 optimal number of topics itself. Then we enumerate the dominant topic for each tweet to understand the distribution of topics across the tweets in the dataset.

3.3 Output Generation

The steps to generate significant topic labels automatically as output from the topics extracted by LDA are discussed below:

1) *Generation of Sentiment Terms and Aspect Terms Cluster:* We create clusters of sentiment terms and aspect terms independently from tweets corresponding to each LDA-generated topic. Thus, we get 20 sentiment terms cluster and 20 aspect terms cluster from the discovered topics by LDA.
2) *Labeling Topic using Top Unigrams:* A Unigram is a one-word sequence of n-gram. The use of unigrams can be observed in NLP, cryptography, and mathematical analysis. Soft cosine similarity considers the similarity of features in vector space model [16]. We extract the top 20 unigrams from sentiment terms cluster and aspect terms cluster respectively for each topic. Then we concatenate all the possible combinations of top unigrams of sentiment terms and top unigrams of aspect terms of topic. We select a combination that has the highest soft cosine similarity value with respect to the tweets of that topic to assign with a significant topic label. We use a sentiment and aspect term tag to label each topic because that feature tag presents an attribute to describe that topic of tweets.

In the section of the methodology of this paper, we present a framework called `SATLabel` to detect key topic labels from the tweets automatically as shown in Table 2. We compare the quality of topic labels generated by `SATLabel` with the manually assigned topic labels in the experiment section. To categorize a tweet with a specific topic label from test data, we search the topic number that has a greater impact of percentage on that tweet.

4 Experiments

4.1 Dataset

We collect the Twitter dataset from the website at `https://www.kaggle.com/datatattle/covid-19-nlp-text-classification`. There are two csv files in the dataset. One is Corona_NLP_train.csv and another is Corona_NLP_test.csv. Tweets available in the dataset are highly unstructured and non-grammatical in syntax. There are 41,157 and 3,798 COVID-19 related tweets are available in Corona_NLP_train.csv and Corona_NLP_test.csv files respectively. We apply a series of preprocessing functions to get the normalized form of noisy tweets for further processing.

 doi:10.20944/preprints202204.0026.v1

4.2 Data Preprocessing

Handling ill-formatted, noisy and unstructured twitter data is one of the most important tasks for us. We preprocess the Twitter dataset to get the normalized form using the functions of transforming words into lowercase, replacing hyperlinks, mentions, and hashtags with empty string, dealing with contractions, replacing punctuation with space, striping space from words, removing words less than two characters, removing stop words, handling Unicode and non-English words.

4.3 Finding the optimal number of topics for LDA

We create a function to return several LDA models with multiple values of a number of topics (k) to find the optimal number of topics. The interpretable topics can be found by selecting a 'k' that identifies the end of a quick rise of topic coherence score. Sometimes we get more granular sub-topics by choosing a higher value of topic coherence score. We pick the model giving the highest

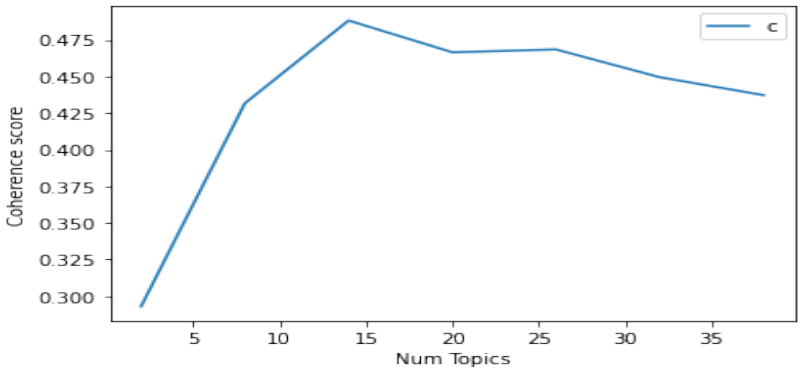

Fig. 2. Selection of the optimal number of LDA topics

coherence value before flattening out considering better sense while the coherence score seems to keep growing as shown in Fig. 2. For the next steps, we choose the model having 20 topics itself.

4.4 Selection of Top Unigrams Features from Clusters

We create sentiment terms cluster and aspect terms cluster of tweets for each topic. We find the top counted 20 unigrams from each cluster. Fig. 3 and Fig. 4 show the top 20 unigrams from sentiment and aspect terms clusters of topic no. 12 respectively. Then we detect the topic label depending on the highest soft cosine similarity value of sentiment and aspect term tag with respect to the tweets of that topic.

 doi:10.20944/preprints202204.0026.v1

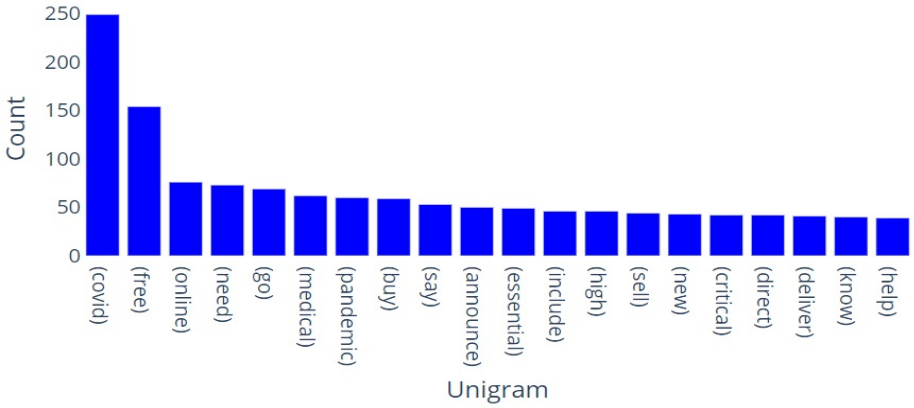

Fig. 3. Top 20 unigrams from sentiment terms cluster of topic no. 12

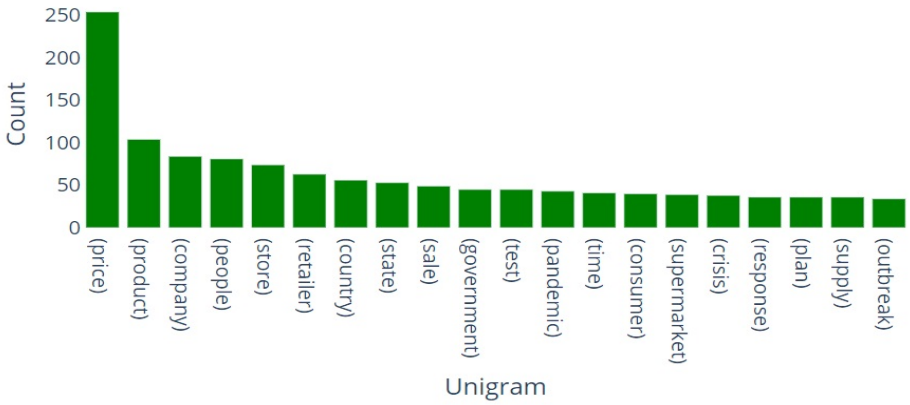

Fig. 4. Top 20 unigrams from aspect terms cluster of topic no. 12

Table 2. Example of Topics Detected on Tweets

Sample Tweet	Topic No.	Detected Topic Label (SATLabel)
Due to the COVID-19 virus and the global health pandemic, we will be closed at our retail location until further notice.	17	Shut Location
Dubai Becomes Cheaper To Live In.	9	Drop Cost
covid-19 is already affecting the online shopping, ok somebody slap meee plsss ?	16	Online Shopping
You guys still can buy food during lockdown then why need to do panic buying?	0	Hoard Food
I'm going to try patenting my world-famous vegetable phall as a killer of covid-19.	14	Learn Scam
It's Not Covid 19. It's due fall in global oil prices Oil cost 30 barrel...	15	Drop Barrel
Here's a buying guide our community set up for the neighborhood supermarket. Feel free to use it as a template.	12	Covid Product
The Consumer Financial Protection Bureau today announced that it is postponing some data collection from the financial industry.	3	Learn Insight
Food demand in poorer countries is more linked to income...	6	Covid Food

4.5 Qualitative Evaluation of Topic Labels

An expert annotator assigns the topic labels manually using the word probabilities in LDA-generated topics to a randomly selected set of tweets. In Table 2, we present a portion of set of tweets assigned by the `SATLabel` generated topic labels. Table 2 shows that `SATLabel` generated topic labels are well-aligned and closely coherent with the descriptions of tweets. We can extract useful information related to a topic, simply by categorizing the tweets using the key label generated by `SATLabel` of that topic.

4.6 Effectiveness Analysis

In this experiment section, we calculate the Soft Cosine Similarity (SCS) values of detected topic labels by `SATLabel` and manual approach for LDA-generated 20 topics. SCS is used to detect the semantic text similarities between two documents. A high SCS value provides a high similarity index and similarity is smaller for unrelated documents. We train the word2vec embedding model to use SCS. We show the comparison of `SATLabel` and manual labeling approach for all LDA-generated topics in terms of SCS value in Fig. 5. We get SCS values generated by proposed `SATLabel` for topic no. 4, 8, 10, 14 are 0.77, 0.61, 0.53, 0.64 respectively while manual approach generates 0.09, 0.06, 0.02, 0.07 SCS

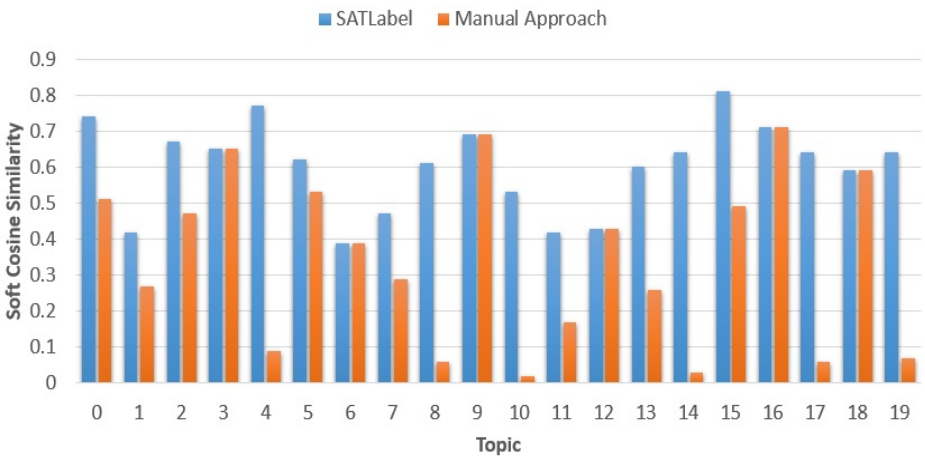

Fig. 5. Comparison of `SATLabel` with manual approach

scores for those topics which are very low. Diverse human interpretation of topics is the possible reason for the high difference of SCS scores between proposed `SATLabel` and manual approach. For topics no. 3, 6, 9, 12, 16, 18 we get the same scs values for SATLabel and manual approach because of identical topic labels generated by both approaches. From Fig. 5, we can observe that the topic labels generated by the proposed framework `SATLabel` produce high SCS values for a maximum number of topics compared with the manual labeling approach. Hence, our proposed framework is more effective and traces better topic labels from unlabelled datasets to reduce the cumbersome task of the human manual labeling approach.

5 Discussion

Automatic labeling of LDA-generated topics of the tweets of social media platforms like Twitter is helpful to understand people's ideas and feelings by going through meaningful insights rather than following traditional strategies like the manual labeling approach. In this paper, we use LDA, a popular probabilistic topic modeling algorithm to extract hidden topics from tweets. We then effectively use sentiment terms and aspect terms of tweets to create clusters. After that, we select top unigrams from the clusters to produce significant topic labels using the maximum soft cosine similarity values. Our proposed framework `SATLabel` helps to produce semantically similar topic labels of tweets to highlight the user's conversations and notice several COVID-19 related issues.

Overall, `SATLabel` is a data-driven framework for topic labeling purposes for mining texts to provide helpful information from the dataset of Twitter related to COVID-19. We firmly believe that `SATLabel` can be effectively used

in other domains of applications like agriculture, healthcare, education, business, cyber-security, etc, and also can be used to generate target class from unlabeled datasets to train deep learning models [13]. These types of contributions allow the researchers and experts in relevant departments to take necessary actions in critical situations like the COVID-19 pandemic by efficiently utilizing social media platforms.

6 Conclustion and Future Work

In this paper, we propose a new framework named `SATLabel` that effectively and automatically identifies key topic labels from COVID-19 tweets. Our framework saves time and reduces the human effort to minimize the overhead of difficult topic labeling tasks from the huge volumes of data to get an overview of broader public opinions on social media platforms like Twitter. We believe that `SATLabel` will help the reformists to discover various COVID-19 related issues by analyzing automatically extracted topic labels.

In the future, we want to increase our scope of experiments by integrating the proposed framework with sentiment classification tasks using hybridization of deep learning methods. We will also implement our proposed framework to other social media platforms on different events to generate significant topic labels to handle the overload of ever-increasing data volume.

References

1. Adhikari, S.P., Meng, S., Wu, Y.J., Mao, Y.P., Ye, R.X., Wang, Q.Z., Sun, C., Sylvia, S., Rozelle, S., Raat, H., et al.: Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (covid-19) during the early outbreak period: a scoping review. Infectious diseases of poverty **9**(1), 1–12 (2020)
2. Asmussen, C.B., Møller, C.: Smart literature review: a practical topic modelling approach to exploratory literature review. Journal of Big Data **6**(1), 1–18 (2019)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research **3**, 993–1022 (2003)
4. Boussaadi, S., Aliane, H., Abdeldjalil, P.O.: The researchers profile with topic modeling. In: 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS). pp. 1–6. IEEE (2020)
5. Elgesem, D., Feinerer, I., Steskal, L.: Bloggers' responses to the snowden affair: Combining automated and manual methods in the analysis of news blogging. Computer Supported Cooperative Work (CSCW) **25**(2-3), 167–191 (2016)
6. Guo, L., Vargo, C.J., Pan, Z., Ding, W., Ishwar, P.: Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. Journalism & Mass Communication Quarterly **93**(2), 332–359 (2016)
7. Hingmire, S., Chougule, S., Palshikar, G.K., Chakraborti, S.: Document classification by topic labeling. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. pp. 877–880 (2013)

8. Hourani, A.S.: Arabic topic labeling using naïve bayes (nb). In: 2021 12th International Conference on Information and Communication Systems (ICICS). pp. 478–479. IEEE (2021)
9. Jahanbin, K., Rahmanian, V., et al.: Using twitter and web news mining to predict covid-19 outbreak. Asian Pacific Journal of Tropical Medicine **13**(8), 378 (2020)
10. Kee, Y.H., Li, C., Kong, L.C., Tang, C.J., Chuang, K.L.: Scoping review of mindfulness research: A topic modelling approach. Mindfulness **10**(8), 1474–1488 (2019)
11. Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., et al.: Applying lda topic modeling in communication research: Toward a valid and reliable methodology. Communication Methods and Measures **12**(2-3), 93–118 (2018)
12. Patil, P.P., Phansalkar, S., Kryssanov, V.V.: Topic modelling for aspect-level sentiment analysis. In: Proceedings of the 2nd International Conference on Data Engineering and Communication Technology. pp. 221–229. Springer (2019)
13. Sarker, I.H.: Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. SN Computer Science **2**(6), 1–20 (2021)
14. Sarker, I.H.: Machine learning: Algorithms, real-world applications and research directions. SN Computer Science **2**(3), 1–21 (2021)
15. Satu, M.S., Khan, M.I., Mahmud, M., Uddin, S., Summers, M.A., Quinn, J.M., Moni, M.A.: Tclustvid: a novel machine learning classification model to investigate topics and sentiment in covid-19 tweets. Knowledge-Based Systems **226**, 107126 (2021)
16. Sidorov, G., Gelbukh, A., Gómez-Adorno, H., Pinto, D.: Soft similarity and soft cosine measure: Similarity of features in vector space model. Computación y Sistemas **18**(3), 491–504 (2014)
17. Tayef Shahriar, K., Sarker, I.H., Nazrul Islam, M., Moni, M.A.: A dynamic topic identification and labeling approach of covid-19 tweets. In: International Conference on Big Data, IoT and Machine Learning (BIM 2021). Taylor and Francis (2021)
18. Wang, B., Liakata, M., Zubiaga, A., Procter, R.: A hierarchical topic modelling approach for tweet clustering. In: International Conference on Social Informatics. pp. 378–390. Springer (2017)
19. Wang, W., Pan, S.J., Dahlmeier, D., Xiao, X.: Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
20. Zhu, B., Zheng, X., Liu, H., Li, J., Wang, P.: Analysis of spatiotemporal characteristics of big data on social media sentiment with covid-19 epidemic topics. Chaos, Solitons & Fractals **140**, 110123 (2020)