Topic Modeling Technique on Covid19 Tweets in Serbian

Adela Ljajić*, Nikola Prodanović*, Darija Medvecki*, Bojana Bašaragin*, Jelena Mitrović*, **
* The Institute for AI R&D of Serbia, Serbia
** Faculty of Computer Science and Mathematics, University of Passau, Germany
{adela.ljajic, nikola.prodanovic, darija.medvecki, bojana.basaragin, jelena.mitrovic}@ivi.ac.rs

***Abstract*—The COVID19 pandemic has brought health problems that concern individuals, the state, and the whole world. The information available on social networks, which were used more frequently and intensively during the pandemic than before, may contain hidden knowledge that can help to better address some problems and apply protective measures more adequately. Since the messages on Twitter are specific in their length, informal style, figurative speech, and frequent use of slang, this analysis requires the application of slightly different techniques than those classically applied to long, formal documents. To determine which topics appear in tweets related to vaccination, we apply state-of-the-art topic modeling techniques to determine which one is the most appropriate. This kind of research is meant to give us an insight into the opinions of the Twitter community on the phenomenon of vaccination and all related aspects. Comparing the results of the LDA with the topics obtained by manual annotation over the same set, we concluded that the LDA method provides a very good interpretation of the topics. Such data allow the analysis of sentiment, in this case pro- or anti-vaccination attitudes, and of specific groups of data and topics.**

I. Introduction

Structured data are mostly in the hands of the state authorities and are used for various statistical analyses related to the pandemic. The research community has predominantly used social media such as Twitter to collect data on the Covid-19 vaccination. Twitter provides a keyword search endpoint, and this has allowed us to collect tweets related to vaccination in the Republic of Serbia. Existing NLP methods allow us to analyze these short texts, which usually express clear opinions on various topics. In a set of 9,623 tweets, we used the Latent Dirichlet Allocation (LDA) topic modeling method [1] to gain insight into which topics are being discussed on Twitter, how we can group the tweets, and what is the topic of the groups within each cluster.

Young people usually freely express their opinions on social networks. The reasons for the hesitant attitude towards vaccination are numerous and, in most cases, not strictly anti-vaccination oriented. We will try to determine which topics are present in the tweets, hoping to help domain experts influence the public in a more informed way when it comes to vaccination. If we know why people, especially young people, are hesitant, we are better equipped for planning campaigns oriented towards vaccination.

We used 9,623 tweets containing keywords related to the vaccination process. Data collection began on January 1, 2021, and is still ongoing. The 9,623 tweets used in this analysis were collected through November 28, 2021. This collection contains tweets in Serbian that have gone through the process of normalization: case lowering, conversion from Cyrillic to Latin script, removal of stop words, tokenization, lemmatization/stemming. We trained state-of-the-art topic modeling algorithms on a normalized dataset. Both clustering and topic modeling are unsupervised methods for data analysis. Clustering uses similarity metrics based on which it divides the document into several groups (clusters). Topic modeling has the task of identifying the topic to which the document belongs, based on a group of words that frequently occur together.

The tweets are short and the existing topic detection and clustering methods do not provide satisfactory results for this type of text. The authors of the Biterm Topic Model (BTM) [2] state that the application of conventional LDA and PLSA topic modeling methods does not work well for short texts because they implicitly capture the document-level word co-occurrence pattern to reveal a topic which is not fitting for short texts due to data sparsity. BTM directly models word co-occurrence patterns at the corpus level and uses those patterns to learn topics and thus solve the problem of sparse word co-occurrence biterm patterns at document-level.

To discover latent topics, we used the most popular topic modeling algorithm developed by Andrew Ng and his colleagues [1]. LDA, the classic and universal model for topic modeling, provided us with the best model coherence and interpretability, even better than expected from BTM considering the shortness of the texts. Using the Eye balling approach for topic evaluation, we compared the topics discovered with LDA with those obtained by two annotators and concluded that LDA provides coherent and human-interpretable topics.

II. Related work

Topic modeling is one of the most efficient ways to detect latent topics and find hidden meaning in a collection of texts. It was initially conceived in the early 1980s and has since evolved into various methods. Among the most common ones are Latent Semantic Analysis (LSA) [3], Probabilistic Latent Semantic Analysis (pLSA) [4], Latent Dirichlet Allocation (LDA) [1], and Non-Negative Matrix Factorization (NMF) [5]. Historically, topic modeling methods were devised for longer texts with a sufficient amount of contextual information and word co-occurrence patterns. With the rise of social media content and the commercial need to gain quick insights from comments on Facebook, Twitter, or Reddit, topic modeling has been confronted with the challenge of unveiling topics in short texts. In addition to being less structured and typically more informal, social media comments present the sparsity issue - because they are limited in size, they suffer from lack of context and lower co-occurrence of words.

There have been several creative attempts to tackle the sparsity issue. Some authors have proposed using the LDA method on additional and closely related datasets of longer texts to jointly learn or infer the topics of short texts [6], [7]. This method is limited by the availability of such closely related datasets. Others have attempted to aggregate shorter texts into larger pseudo-documents, grouping them by a parameter such as the same user or hashtag in [8], [9] and [10]. These authors then applied LDA to the aggregated texts and obtained promising results. Similarly, the authors in [11] performed aggregation combining the texts into an aggregated model if their similarity was sufficiently high and reported a significant improvement in topic coherence.

Several novel methods have been proposed specifically for topic modeling of short texts. The Mixture of Unigrams (MoU) [12] was one of the earliest attempts at short text modeling. This method starts from the assumption that each short document covers a single topic. The authors in [13] report the competitive performance of MoU mixed with LDA, as a method much better suited for short than for long texts. Its main downside lies in its inability to account for the fact that a short text can still cover more than one topic. Another method proposed for short text modeling is the Biterm Topic Model (BTM) [2]. BTM provides a solution to data sparsity by considering the word co-occurrence patterns (biterms) throughout the entire corpus. The authors report significant improvements over LDA by being able to learn a global topic distribution, and MoU, by allowing every biterm to represent a different topic and thus cover possible multiple topics for every short text.

By testing the existing topic modeling methods on a Facebook conversation dataset, the authors in [14] show that out of five topic modeling methods (LDA, LSA, NMF, PCA, and RP), LDA and NMF perform best in terms of producing higher quality and more coherent topics. Despite being slower than NMF, they show that LDA is more flexible and consistent and that it provides more meaningful and logical topics. Even though many authors point to drawbacks of LDA when it comes to topic modeling on short texts, as in [15], [16] and [17], it nevertheless seems to be the preferred method for discovering latent topics in short texts.

From the beginning of the pandemic COVID -19, topic modeling techniques have been used to determine public attitudes toward various aspects of the pandemic, particularly vaccination. Kwok, Vadde and Wang report being the first to apply LDA to the topic of COVID-19 vaccines [17]. Using topic modeling and sentiment analysis, they were able to discover three latent topics in a Twitter corpus. Several other researchers continued along the same lines, using the combination of LDA and sentiment analysis to gauge public sentiment using either tweets or Reddit corpora. Lyu, Han and Luili managed to correlate the detected topics to the main COVID-19 vaccination events [18], and Wang and Chen tracked changes in public sentiment regarding vaccine hesitancy [19], both using an LDA algorithm on Twitter corpora. Melton et al. in [20] applied LDA to a Reddit corpus, but mentioned the need to manually inspect the automatically returned topics due to potential problems with qualitative coherence. Ma, Zeng-Treitler and Nelson [21] compared the performance of LDA to a novel model, Top2Vec, in detecting tweets that express vaccine hesitancy and topics represented by those tweets. Besides being able to extract more relevant and differentiated topics using Top2Vec (8 compared to 4 detected by LDA), they highlight a simplified data-cleaning pipeline as its important asset.

Even though substantial work has been done on sentiment analysis for Serbian [22], [23] and [24], to the best of our knowledge this is the first attempt to apply topic modeling methods to Serbian in general and to tweets in Serbian in particular. In parallel with this effort, Prodanović et al. in [25] applied BERTić, a BERT-based deep learning model adapted to Bosnian, Croatian, Montenegrin, and Serbian, to tweets related to COVID-19 vaccination hesitancy, to classify the author's sentiment.

III. Method

We investigated the COVID-19 vaccine hesitancy by analyzing Twitter posts (tweets) using topic modeling. Topic modeling is an unsupervised, probabilistic process of learning and extracting abstract (hidden) topics from a large number of documents. We used it to obtain hidden topic in negative tweet related to vaccine hesitancy in the Serbian language. Our pipeline - from tweets to topics - consists of several steps: extraction and annotation of data, data preprocessing, data transformation, creation of a topic model - LDA, model evaluation and result interpretation. The whole process is depicted in Fig.1. Each step of the pipeline is described in detail in the following subsections.

A. *Extracting and annotating data*

Tweets were collected using the Twitter API. Only tweets in Serbian language (both in Cyrillic and Latin script) were used. Also, we filtered the tweets to be only from the location of the Republic of Serbia - since we want the results to reflect the opinions and topics of people currently living in Serbia. Some of the example keywords used for tweets filtering are: ‘vakcina', 'moderna’, fajzer', 'astrazeneca', ‘sputnikV', 'irnk’, ’ирнк', ‘vektorska vakcina', etc. The choice of keywords reflects the intention to choose attitudes towards different types of vaccine manufacturers and technologies as an attitude towards the vaccination process in general.

Data were annotated by 2 annotators using three classes of sentiment (positive, negative, neutral). Annotators also provided lists of aspects/topics that predominate in positive and negative tweets separately. These lists were used to compare the manually discovered topics with topics obtained with LDA model for topic modeling. The inter-annotator agreement score for sentiment classification, obtained after the first phase of manual annotation using the Cohen Kappa score is 0.52139.

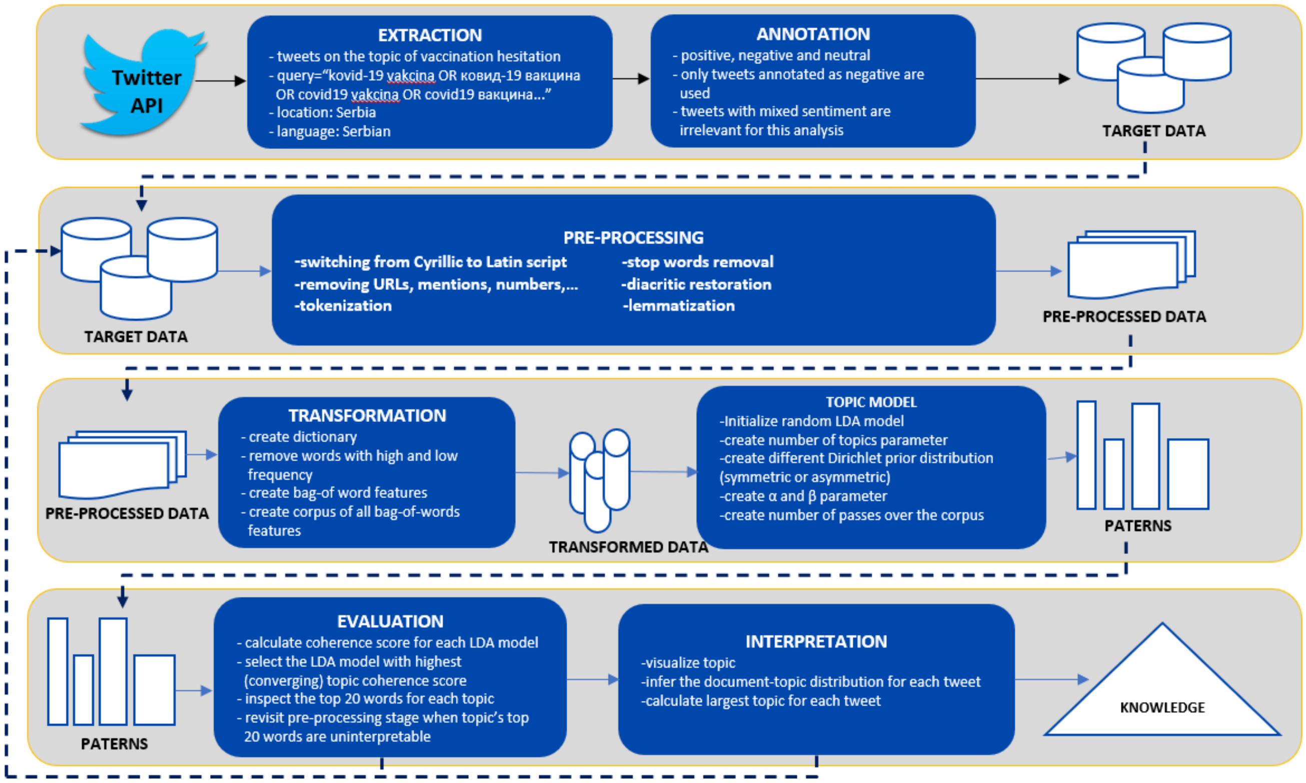

Twitter API
EXTRACTION
- tweets on the topic of vaccination hesitation
- query="kovid-19 vakcina OR ковид-19 вакцина OR covid19 vakcina OR covid19 вакцина..."
- location: Serbia
- language: Serbian
ANNOTATION
- positive, negative and neutral
- only tweets annotated as negative are used
- tweets with mixed sentiment are irrelevant for this analysis
TARGET DATA
TARGET DATA
PRE-PROCESSING
-switching from Cyrillic to Latin script
-removing URLs, mentions, numbers,...
-tokenization
-stop words removal
-diacritic restoration
-lemmatization
PRE-PROCESSED DATA
PRE-PROCESSED DATA
TRANSFORMATION
- create dictionary
- remove words with high and low frequency
- create bag-of word features
- create corpus of all bag-of-words features
TRANSFORMED DATA
TOPIC MODEL
-Initialize random LDA model
-create number of topics parameter
-create different Dirichlet prior distribution (symmetric or asymmetric)
-create α and β parameter
-create number of passes over the corpus
PATERNS
PATERNS
EVALUATION
- calculate coherence score for each LDA model
- select the LDA model with highest (converging) topic coherence score
- inspect the top 20 words for each topic
- revisit pre-processing stage when topic's top 20 words are uninterpretable
INTERPRETATION
-visualize topic
-infer the document-topic distribution for each tweet
-calculate largest topic for each tweet
KNOWLEDGE

Figure 1. Topic modeling pipeline

Out of a total of 9,623 tweets we initially had, we ended up with 1768 tweets that we used to model topics. This final set of tweets was annotated as negative by two annotators.

B. Preprocessing

To remove various types of noise and reduce the space for topic modeling, we applied the following processing steps to the input text:

- Switching from Cyrillic to Latin script;
- Tokenization;
- Removing URLs, mentions, numbers, new lines, emojis, images, special characters, ...;
- Stop words removal;
- Diacritic restoration;
- Lemmatization.

We performed the conversion to Latin script using the [26] Python library and removed URLs, mentions etc. using "re" python standard package [27]. Tokenization, diacritic restoration and lemmatization was done using a CLASSLA pipeline for Serbian [28].

C. Transformation for LDA

The transformation required to create the LDA model consists of creating dictionary/vocabulary (a list of unique words represented as integers), the pruning process of removing words with low and high frequency and representing the tweets as bag-of-words features. Creating a corpus of all tweets as bag-of-words features is also necessary at this stage.

D. Create topic model – LDA

To obtain cluster assignments, the LDA model uses two probability values: P(word|topics) and P(topics|documents). We create a topic model by initializing a random LDA model, creating random number of topics parameter, α and β parameter and the number of passes over the corpus. We trained the initial model with 10 topics and 'auto' values for α and β parameters and got the coherence Cv score of 0.27017.

This initial model is evaluated and the hyperparameters are tuned in order to give the enhanced model (with a better coherence score).

IV. Model Evaluation – Topic Coherence

Moreover, there is no gold standard list of topics to compare with the corpus we used. For the evaluation, we used an intrinsic metric that is able to capture the semantics of the model and the interpretability of the topics. We did not use an extrinsic evaluation metric, as it evaluates how good a model is at performing tasks such as classification.

Topic Coherence measures the semantic similarity between the top words in each topic generated by the LDA model. The assumptions are that words with similar meanings tend to occur together in similar contexts and that a topic would contain semantically similar words among its top scored representatives.

We used Cv, which is commonly used as a coherence score because it has performed well in scoring how interpretable topics are by human readers.

A. *Hyperparameter tuning for LDA*

LDA requires that the number of topics be determined in advance. Based on this number (we will call it K), the algorithm generates K topics that best fit the data.

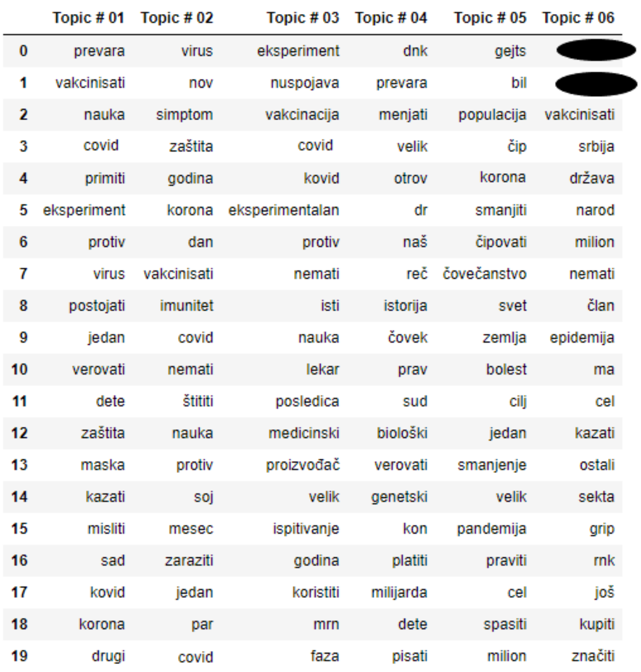
We perform hyperparameter tuning to get the parameters that match the model with the best coherence score. We perform a series of sensitivity tests to help determine the following model hyperparameters: number of topics K, dirichlet hyperparameter alpha (document-topic density), dirichlet hyperparameter beta (word-topic density).

To quantitatively evaluate topic models by the measure of topic coherence we used Gensim library implementation [29]. We found the best value of the hyperparameter at which the maximum topic coherence is achieved: alpha=asymmetric, beta=0,91. no. of topics=6. This parameter optimization yields approx. 9% improvement over the baseline score. The coherence score with the best parameters is 0.2973.

V. Results

The obtained topics (based on 20 representative words) are shown in Table I. Common reasons for vaccine hesitancy found in the dataset included concerns about experiment/scam, side effects associated with the COVID-19 vaccines, DNA change, conspiracy theory, vaccine hesitation mixed with a political attitude, and a mixture of scientific and general skepticism related to vaccine development and distribution. The experts interpreted and named the topics as shown in the second column of the Table II. The topics are very close and contain similar words – due to the very narrow field (not only COVID-19, but Covid-19 vaccine). More data is needed for more accurate topics modeling.

TABLE I.
TOPICS SCORED 20 REPRESENTATIVE WORDS



	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06
0	prevara	virus	eksperiment	dnk	gejts	[redacted]
1	vakcinisati	nov	nuspojava	prevara	bil	[redacted]
2	nauka	simptom	vakcinacija	menjati	populacija	vakcinisati
3	covid	zaštita	covid	velik	čip	srbija
4	primiti	godina	kovid	otrov	korona	država
5	eksperiment	korona	eksperimentalan	dr	smanjiti	narod
6	protiv	dan	protiv	naš	čipovati	milion
7	virus	vakcinisati	nemati	reč	čovečanstvo	nemati
8	postojati	imunitet	isti	istorija	svet	član
9	jedan	covid	nauka	čovek	zemlja	epidemija
10	verovati	nemati	lekar	prav	bolest	ma
11	dete	štititi	posledica	sud	cilj	cel
12	zaštita	nauka	medicinski	biološki	jedan	kazati
13	maska	protiv	proizvođač	verovati	smanjenje	ostali
14	kazati	soj	velik	genetski	velik	sekta
15	misliti	mesec	ispitivanje	kon	pandemija	grip
16	sad	zaraziti	godina	platiti	praviti	rnk
17	kovid	jedan	koristiti	milijarda	cel	još
18	korona	par	mrn	dete	spasiti	kupiti
19	drugi	covid	faza	pisati	milion	značiti

TABLE II.
COMPARATION OF HUMAN ANNOTATED AND LDA DISCOVERED TOPIC

Topic #	**Topic identified by LDA-named by experts**	**Topics identified by human annotators**	
		First annotator	**Second annotator**
01	Concern that vaccine is experiment/scam	- Experiments - Produced too quickly	- Untested vaccine
02	General uncertainty	- Lack of info - Inconsistency of info - Doubt about the vaccination process - Effectiveness	- Ignorance
03	Side Effects/ harmful consequences	- Safety - Ingredients of the vaccine	- Ingredients of the vaccine
04	Fear of DNA change	- Genetic modification - Planned gene therapy	- Genetic modification
05	Conspiracy theory/ population controlled by Bill Gates	- The attitudes of the others - Planned gene therapy	-
06	Concern mixed with political attitude	- Politics	-

A. *Eyeballing and LDA topic comparation*

Since our dataset is small, it is difficult to measure the exact success of the LDA model. Therefore, we decided to compare the results with the topics identified separately by two human annotators in each tweet with negative sentiment as a reason for hesitation toward vaccination if the Twitter user expressed one. The first annotator labeled the topics in more detail based on well-known reasons for hesitation to vaccinate, while the second annotator went through the annotation with a more general approach and strictly annotated the tweets with a clear attitude toward vaccines, excluding the tweets based on political attitudes and conspiracy theories. Our results showed that the topics identified by the LDA model were highly consistent with the human annotations.

The annotators identified concerns expressed by users that the vaccine is an *experiment* and that it is being developed too quickly and used on a massive scale and has not been sufficiently tested. The *general uncertainty* is consistent with both human annotators regarding the lack of information and knowledge about vaccines that Twitter users see as a problem. In particular, the first annotator noted doubts about the vaccination process and about whether there are any health effects. The third topic identified by the LDA model corresponds to the concerns identified by the annotators about the presence of *side effects* that question the safety of vaccines based on a general opinion or a more specific opinion because of some specific ingredients of vaccines. The topic of *fear of DNA change* is completely consistent with the topics noted by the two annotators. The first annotator also noted that some Twitter users believe that vaccination is a planned gene therapy, which could also be classified as a *conspiracy theory* due to the nature of these tweets. Also on this topic, the first annotator identified that the personal attitudes of various people from around the world can trigger conspiracy theories. *Concern mixed with political attitude* is not an uncommon case on Twitter identified by the LDA model and the first annotator.

VI. CONCLUSION

The goal of the research presented in this paper was to determine how much the LDA method can yield topics that are coherent and interpretable. Due to the modest size of the training data set, and since the topics to be discovered are quite semantically loose (representative words often occur in multiple topics), we had not expected such good results with the LDA method. The reasons for vaccination hesitancy almost completely coincide with the results that were manually determined by the annotators. In the future, we plan to compare these results with the results of other methods for modeling topics in short texts. We also plan to apply this model along with a system that classifies tweets based on their sentiment and discovers topics for different sentiment polarities.

ACKNOWLEDGMENT

This work was partially supported by the Government of Republic of Serbia and partially by organizations USAID and the United Nations Development Programme.

REFERENCES

[1] D. M. Blei, “Latent Dirichlet Allocation,” p. 30.

[2] X. Yan, J. Guo, Y. Lan, and X. Cheng, “A biterm topic model for short texts,” in *Proceedings of the 22nd international conference on World Wide Web*, New York, NY, USA, May 2013, pp. 1445–1456. doi: 10.1145/2488388.2488514.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[4] T. Hofmann, “Probabilistic latent semantic indexing,” in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50–57.

[5] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[6] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang, “Transferring topical knowledge from auxiliary long texts for short text clustering,” in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 775–784.

[7] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, “Learning to classify short and sparse text & web with hidden topics from large-scale data collections,” in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 91–100.

[8] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, “Improving lda topic models for microblogs via tweet pooling and automatic labeling,” in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 2013, pp. 889–892. [Online]. Available: https://www.researchgate.net/publication/260639616_Improving_LDA_Topic_Models_for_Microblogs_via_Tweet_Pooling_and_Automatic_Labeling

[9] A. Steinskog, J. Therkelsen, and B. Gambäck, “Twitter topic modeling by tweet aggregation,” in *Proceedings of the 21st nordic conference on computational linguistics*, 2017, pp. 77–86.

[10] J. Weng, E.-P. Lim, J. Jiang, and Q. He, “Twitterrank: finding topic-sensitive influential twitterers,” in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 261–270.

[11] S. Blair, Y. Bi, and M. Mulvenna, “Aggregated topic models for increasing social media topic coherence,” *Appl. Intell.*, vol. 50, Jan. 2020, doi: 10.1007/s10489-019-01438-z.

[12] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, “Text classification from labeled and unlabeled documents using EM,” *Mach. Learn.*, vol. 39, no. 2, pp. 103–134, 2000.

[13] X. Quan, C. Kit, Y. Ge, and S. J. Pan, “Short and sparse text topic modeling via self-aggregation,” 2015. [Online]. Available: https://personal.ntu.edu.sg/sinnopan/publications/[IJCAI15]Short%20and%20Sparse%20Text%20Topic%20Modeling%20via%20Self-Aggregation.pdf

[14] R. Albalawi, T. H. Yeap, and M. Benyoucef, “Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis,” *Front. Artif. Intell.*, vol. 3, 2020, Accessed: Mar. 28, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/frai.2020.00042

[15] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, and J. Boyd-Graber, “Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter,” in *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, 2015, pp. 99–107. [Online]. Available: https://aclanthology.org/W15-1212.pdf

[16] I. Vayansky and S. A. P. Kumar, “A review of topic modeling methods,” *Inf. Syst.*, vol. 94, p. 101582, Dec. 2020, doi: 10.1016/j.is.2020.101582.

[17] S. W. H. Kwok, S. K. Vadde, and G. Wang, “Tweet Topics and Sentiments Relating to COVID-19 Vaccination Among Australian Twitter Users: Machine Learning Analysis,” *J. Med. Internet Res.*, vol. 23, no. 5, p. e26953, May 2021, doi: 10.2196/26953.

[18] J. C. Lyu, E. L. Han, and G. K. Luli, “COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis,” *J. Med. Internet Res.*, vol. 23, no. 6, p. e24435, Jun. 2021, doi: 10.2196/24435.

[19] Y. Wang and Y. Chen, “Characterizing discourses about COVID-19 vaccines on Twitter: a topic modeling and sentiment analysis approach,” *J. Commun. Healthc.*, vol. 0, no. 0, pp. 1–10, Mar. 2022, doi: 10.1080/17538068.2022.2054196.

[20] C. A. Melton, O. A. Olusanya, N. Ammar, and A. Shaban-Nejad, “Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence,” *J. Infect. Public Health*, vol. 14, no. 10, pp. 1505–1512, Oct. 2021, doi: 10.1016/j.jiph.2021.08.010.

[21] P. Ma, Q. Zeng-Treitler, and S. J. Nelson, “Use of two topic modeling methods to investigate covid vaccine hesitancy,” in *14th International Conference on ICT, Society, and Human Beings, ICT 2021, 18th International Conference on Web Based Communities and Social Media, WBC 2021 and 13th International Conference on e-Health, EH 2021-Held at the 15th Multi-Conference on Computer Science and Information Systems, MCCSIS 2021*, 2021, pp. 221–226. [Online]. Available: https://www.ict-conf.org/wp-content/uploads/2021/07/04_202106C030_Ma.pdf

[22] M. Mladenović, J. Mitrović, C. Krstev, and D. Vitas, “Hybrid sentiment analysis framework for a morphologically rich language,” *J. Intell. Inf. Syst.*, vol. 46, no. 3, pp. 599–620, 2016.

[23] V. Batanović and B. Nikolić, “Sentiment classification of documents in Serbian: The effects of morphological normalization and word embeddings,” *Telfor J.*, vol. 9, no. 2, pp. 104–109, 2017.

[24] A. Ljajic and U. Marovac, “Improving sentiment analysis for twitter data by handling negation rules in the Serbian language,” *Comput. Sci. Inf. Syst.*, vol. 16, no. 1, pp. 289–311, 2019, doi: 10.2298/CSIS180122013L.

[25] N. Prodanović, A. Ljajić, D. Medvecki, J. Mitrović, and D. Ćulibrk, “Deep learning analysis of tweets regarding Covid19 Vaccination in the Serbian language,” *ICIST 2022 - 12th Int. Conf. Inf. Soc. Technol. Kopaonik Serbia*, 2022.

[26] “cyrtranslit · PyPI.” https://pypi.org/project/cyrtranslit/ (accessed May 31, 2022).

[27] “re — Regular expression operations — Python 3.10.4 documentation.” https://docs.python.org/3/library/re.html (accessed May 31, 2022).

[28] N. Ljubešić, “The CLASSLA-StanfordNLP model for lemmatisation of standard Serbian 1.1,” 2020.

[29] “gensim: topic modelling for humans,” *Doc2vec paragraph embeddings*. https://radimrehurek.com/gensim_3.8.3/models/doc2vec.html (accessed Jan. 18, 2022).

.