Socio-epidemiological Insights From a Yearlong COVID-19 Twitter Stream

Ana-Arina Raileanu, Ioan-Florin-Cătălin Nițu, Batuhan Faik Derinbay
{ana-arina.raileanu, ioan.nitu, batuhan.derinbay}@epfl.ch
School of Computer and Communication Sciences, EPFL, Switzerland

***Abstract*—We analyze Twitter data from the last two years of the COVID-19 pandemic. We use two different approaches to enrich the current topic modelling approach, which is based on clustering hashtags based on their Word2Vec representation, by applying Latent Dirichlet Allocation and BERTopic on the whole content of a tweet. In this process, we compare the topic trends with various epidemiological factors to gather insights into people's interests. We collect our findings into an interactive website: https://topics.derinbay.com/.**

I. Introduction

During this project, we aimed to enhance the paper Socio-epidemiological insights from a yearlong COVID-19 Twitter stream, a study conducted over a year of data gathered from Twitter that focuses on the ”info-demic” related to the COVID-19 epi-demic [1]. We analyzed tweets and hashtags using BERTopic and LDA, cleaned data, and applied NLP techniques, like stop world removal or tokenization, to process the data. Moreover, we provided a set of visualizations (including word clouds and bubble charts of the most relevant topics) to gain insights from the data and better understand its behavior.

II. Models and Methods

A. From Hashtags to Tweets

The initial approach at analyzing tweets was based on using Word2Vec to cluster tweet hashtags. However, not all of the tweets in the dataset contained hashtags. Furthermore, the content of a tweet is often more descriptive than its hashtags alone, as it generally contains more information. Therefore, we provide two alternative methods for analyzing tweets, using BERTopic and LDA to model the topic of a tweet based on its textual content concatenated with its hashtags.

B. Evaluation Dataset

Given the large size of the Twitter dataset (over 660 millions tweets collected throughout 2 years), we have decided to evaluate the models on a limited selection of tweets. In turn, this practice allows us to iterate faster through the development steps. The limited dataset is created by applying reservoir sampling on the full dataset, with a target of 100.000 tweets. Reservoir sampling is a method that allows extracting a random sample of data points without replacement from a dataset with n number of data samples that is too large to fit into memory at once [2]. The algorithm is used when n is either really large or unknown and assumed to be really large.

C. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [3] is a probabilistic generative model for collections of discrete data, such as textual data. It is a hierarchical Bayesian model with three levels in which each item in a collection is described as a finite mixture over an underlying set of topics. Each word in a document is assigned to a separate subject in LDA, which assigns documents to a list of topics.

In order to model our topics using LDA, we first need to pre-process the data into a corpus. There are several steps taken to process *semi-clean* 100.000 raw tweet texts. These pre-processing steps are cleaning, stop word removal, tokenization, stemming and lemmatization.

1) Data Cleaning: The first step of the data pre-processing is always data exploration and cleaning. Since we are dealing with such an enormous number of tweets, hyperlinks (URLs) and mentions (usernames) are replaced with placeholders and the collected raw data is saved as *semi-clean* data. After sampling from *semi-clean* tweets, emojis are removed and the encoding is unified.

2) Stop Word Removal: One of the most challenging parts of the pre-processing was the stop word removal due to our decision of preserving the multilingual structure of the Twitter data. We have used several different `python` libraries including but not limited to `spaCy` [4], `gensim` [5], `wordcloud` [6], and many more to get a wider range of multilingual stop word support. For shorter and uncommon stop words, we manually added a custom stop word set to our global stop word collection.

3) Tokenization: Similar to stop words, we used `spaCy`'s the most comprehensive tokenizers, which are English large and multilingual small, to tokenize our data. Due to the majority of the tweets being English and coverage of the larger English tokenizer, we decided to use the English tokenizer after observing both tokenizers' results.

4) Stemming and Lemmatization: As the final step, we resorted using `spaCy`'s English stemmer and lemmatizer due to the aforementioned reasons. Just before the lemmatization step, any punctuations and pronouns are also removed that are tagged by the tokenizer.

When it comes to interpreting and evaluating the results of LDA, perplexity and coherence scores are commonly used.

During our studies we examined perplexity, C_V coherence and UMass coherence scores. Once satisfied with our results, we conducted a hyper-parameter search and trained over 300 models. Evaluation of the hyper-parameter search used C_V coherence, also known as the boolean sliding windows calculation, as the decisive metric where the coherence value c is $0 \leq c \leq 1$ and higher the better. We searched the best parameters for number of topics (n_t), number of passes (n_p), alpha (α) and eta (β) where n_p is the number of passes through the corpus during training, α and β are priori believes on topic-word distributions.

Finally, using the `pyLDAvis` [7], we visualized the results of our best performing model and published it on our website for everyone to see.

D. BERTopic

BERTopic [8] is a library that provides BERT-based models which can be used for the task of topic modelling. We have fitted two models on the evaluation dataset, the standard model provided by BERTopic and the Multilingual model, which can handle over 50 languages [1]. Afterwards, we have transformed 10% of the tweets gathered throughout the years, by randomly sampling 10% each day, to create trends related to the content of a tweet. For our analyses, we used the English model, as the majority of the tweets in our dataset come from the United States (53.4%) [1], and the English model obtained a higher coherence score when compared to the Multilingual one. No additional preprocessing was necessary [2], as all the parts of a tweet are relevant for its meaning. During fitting, the number of desired clusters was not specified, so we would have more flexibility if, at a later date, we would decide to change this number. However, as the model generated a high number of topics (451), we reduced it to 100 (out of which one is undefined), given the hyperparameter tuning results of LDA.

E. Topic Trends

Topic trends are computed for both the hashtag clustering model and the multilingual BERTopic model, by grouping the tweets per day based on the assigned cluster.

When analyzing the Twitter data, we have come across the following issue: the number of tweets scraped per day varies significantly. While for some days we have thousands of tweets, for other days we have only tens. Therefore, in order to make a fair comparison between the number of tweets per topic in two different days, we normalize the daily count per topic by dividing it to the number of all tweets of that day. When doing so, we asssume that the scraped tweets are representative of the distribution of topics of all tweets. Furthermore, data was missing for some months, such as August 2020. In this case, it remains to collect the missing data and recompute the trends for that period.

	Perplexity	CV Coherence	UMass Coherence
LDA	-9.6966	0.5210	0.5210

Table I: LDA scores on the evaluation dataset.

	CV Coherence	UMass Coherence
English BERTopic	0.6865	-0.8139
Multilingual BERTopic	0.5822	-0.9738

Table II: BERTopic scores on the evaluation dataset.

III. Results and Comparisons

A. Model Comparison

We have opted to compare our models using the CV coherence and the UMass coherence scores [9]. For this comparison, all models were ran with 100 topics on the evaluation dataset. Latent Dirichlet Allocation (LDA) obtains a lower CV coherence (table I) than the BERTopic models (table II), even after hyperparameter tuning.

In both our LDA and English BERT approaches, we have noticed a similar pattern: some clusters only contain tweets written in a certain language. Therefore, even though they may be discussing a different topic, they are all grouped together into a cluster that does not differentiate between them. In order to counteract this problem, we have considered using the Multilingual version of BERTopic. Figure 1 and figure 2 show the 6 most common topics of each model, together with their representative words. It can be observed that while Topic 2 of the English model is defined by Spanish terms (including stop-words), Topic 0 of the multilingual model contains a translation of the word ”mask” along its other terms. Furthermore, the top two words for Topic 0 of the English model seem to be unrelated with the rest of the terms in this topic. Table II shows the comparison between the two models. Since the English model obtained both a better CV coherence and a better UMass coherence score, we performed our trends analysis using it.

B. Hashtag Topic and Tweet Topic

The initial approach at clustering tweets by topic was to analyze only the hashtags of a tweet. We use the English BERTopic model to compare the most common topics with the ones obtained by extracting topics from hashtags during 3 key events: Pfizer receives authorization in the United States, the number of COVID-19 cases peak, and the first major lockdown. Figure **??** presents the most prevalent topics during these events, based on the two approaches. It can be observed that the two approaches complement each other, each being able to portray the effect of the event on the social media website. Detecting topics based on hashtags, when available, could be used to enhance the analysis of the tweet content.

[1] https://maartengr.github.io/BERTopic/index.html

[2] https://maartengr.github.io/BERTopic/faq.html

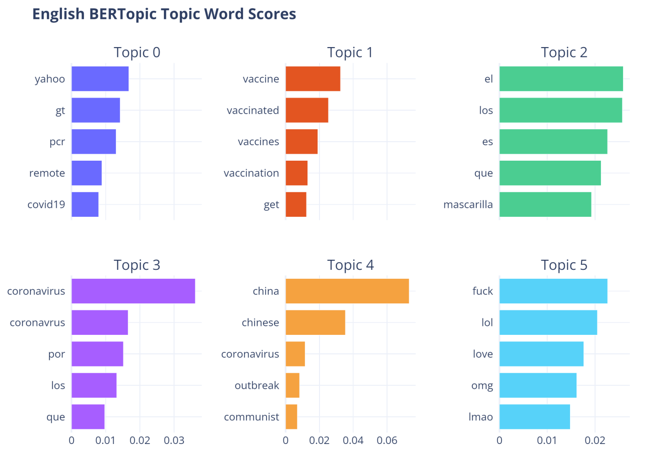

Figure 1: The 6 most common topics of English BERTopic.

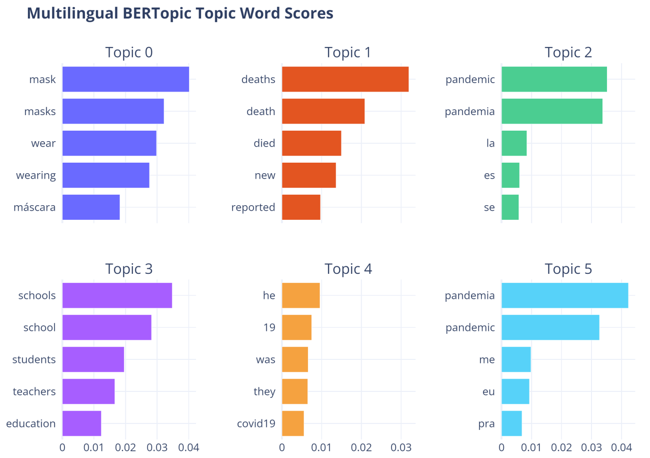

Figure 2: The 6 most common topics of Multilingual BERTopic.

C. Tweet Topic Trends

We have analyzed the epidemiological data gathered in the last two years in comparison to the trends of the COVID-19 topics, as obtained by the English BERTopic model. One such analysis compared the number of tweets related to vaccination in relation to the number of new vaccinations, as presented in figure 4. There seems to be a clear relation between an increasing number of topics discussing vaccination as soon as vaccines became available to more people in the United States. In particular, a spike can be observed in December 2020 when the Pfizer vaccine received emergency use authorization, suggesting the interest of people at that time. Figure 5 presents the relation between tweets related to fear and hope and the number of COVID-19 detected in the United States. It can be observed that when cases initially started to rise, the number of tweets related to fear and hope also increased, suggesting the fear of a pandemic. When the number of COVID-19 cases peaked in the United States, there were fewer fearful/hopeful tweets than when the pandemic began, suggesting that the vaccine may have contributed to people's feeling of safety.

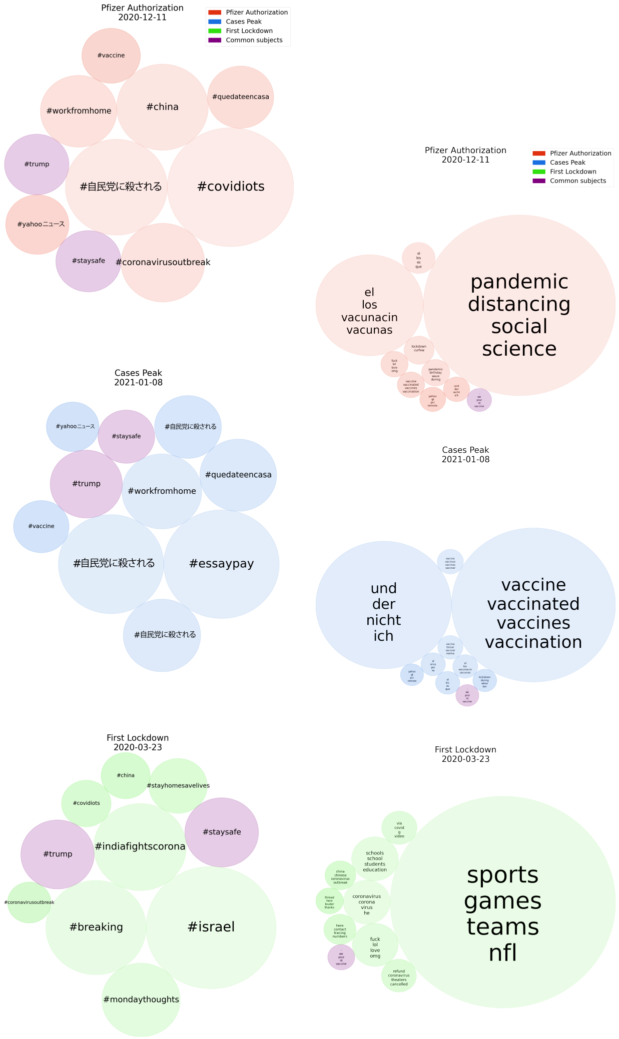

Figure 3: Main topics based on hashtag topics (left) and tweet topics (right) clustering during three crucial events.

IV. Conclusions

We have analyzed multiple approaches for topic modelling of tweets in the COVID-19 period. Starting from topic detection based on hashtags alone, we have enriched the project with two alternative approaches, using Latent Dirichlet Allocation and BERTopic. In turn, this has allowed us to take a closer look at the topics people discuss in tweets, and analyze their relation with epidemiological trends. Furthermore, we have collected our findings and presented them in an interactive on our website: https://topics.derinbay.com/.

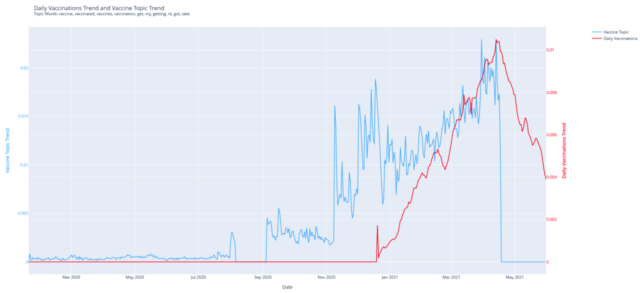

Figure 4: The topic of vaccination in relation to the daily number of vaccinations in the United States.

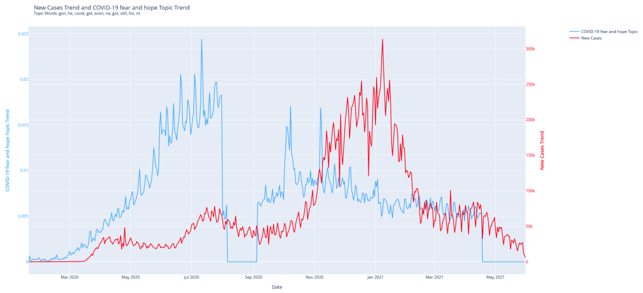

Figure 5: The topic of COVID-19 fear/hope in relation to the daily number of new cases in the United States.

Acknowledgements

We thank Mary-Anne Hartley, Prakhar Gupta, Lucas Massemin, and the whole iGH: intelligent Global Health laboratory for their support, logistics, materials, and know-how during the development of this project.

References

[1] "Socio-epidemiological insights from a yearlong COVID-19 Twitter stream," *Unpublished*, 2021.

[2] J. S. Vitter, "Random sampling with a reservoir," *ACM Trans. Math. Softw.*, vol. 11, no. 1, p. 37–57, mar 1985. [Online]. Available: https://doi.org/10.1145/3147.3165

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 993–1022, mar 2003.

[4] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

[5] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.

[6] L. Oesper, D. Merico, R. Isserlin, and G. D. Bader, "Wordcloud: a cytoscape plugin to create a visual semantic summary of networks," *Source code for biology and medicine*, vol. 6, no. 1, p. 7, 2011.

[7] B. Mabey, M. Susol, and Y. Tay, "pyldavis," https://github.com/bmabey/pyLDAvis, 2015.

[8] M. Grootendorst, "Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics." 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4381785

[9] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ser. WSDM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 399–408. [Online]. Available: https://doi.org/10.1145/2684822.2685324