Using neural topic models to track context shifts of words: a case study of COVID-related terms before and after the lockdown in April 2020

Olga Kellert and **Md Mahmud Uz Zaman**
University of Göttingen, Germany
olga.kellert@phil.uni-goettingen.de and
mail.mahmuduzzaman@gmail.com

Abstract

This paper explores lexical meaning changes in a new dataset, which includes tweets from before and after the COVID-related lockdown in April 2020. We use this dataset to evaluate traditional and more recent unsupervised approaches to lexical semantic change that make use of contextualized word representations based on the BERT neural language model to obtain representations of word usages. We argue that previous models that encode local representations of words cannot capture global context shifts such as the context shift of *face masks* since the pandemic outbreak. We experiment with neural topic models to track context shifts of words. We show that this approach can reveal textual associations of words that go beyond their lexical meaning representation. We discuss future work and how to proceed capturing the pragmatic aspect of meaning change as opposed to lexical semantic change.

1 Introduction

Various approaches have been suggested in previous research to analyze semantic change such as semantic narrowing or broadening or the appearance of new words. Traditional quantitative approaches to semantic change identify meaning change by relative-frequency-based methods (Gulordava and Baroni, 2011). Another traditional method is the n-gram approach, that identifies the change in likelihood of words co-occurrence across time (Gulordava and Baroni, 2011; Butler and Simon-Vandenbergen, 2021; Luo, 2021). While these approaches can detect and track lexical changes that rely on local lexical and syntactic differences of words in time, they cannot capture lexical changes that rely on more distant relations between words (Giulianelli et al., 2020).

More recent computational approaches to semantic change, have exploited a pre-trained neural language model BERT (Devlin et al., 2019) to obtain contextualized representations for every occurrence of a word of interest and measure changes of semantic clusters in time (Giulianelli et al., 2020). Contextualized embeddings can help to detect lexical changes in case linguistic differences are encoded in non-local word relations such as in the whole construction like in *here comes your coach, Cinderella* or in *you can always go, coach* (Giulianelli et al., 2020). These constructions capture lexical changes by being associated with different uses of the word *coach*, which linguistically represent notions like (in)animacy.

Our goal in this paper is to capture more global relations between words than word relations in constructions. We want to capture context shifts by global representations of words that share the same paragraph or document. We use neural contextualized embeddings (Devlin et al., 2019) to generate better topics, which we consider as proxies for various word contexts, similar to constructions as proxies for word senses in previous approaches (Giulianelli et al., 2020). By studying thematic relations between words in time, we find out changes in these relations. This approach is not new and has been already applied by (Sagi et al., 2013). The authors used classical topic modeling to find out new associations of words. Words associations have been described in cognitive linguistics by the notion of semantic frames (Lakoff, 2008; Fillmore et al., 2002). According to (Lakoff, 2008), every word evokes a certain frame, i.e. a conceptual structure used in human communication. Words evoke certain images, feelings and (personal) experiences that can be expressed by other words used in the same context. In order to capture new associations of words in time, Sagi et al. (2013) used topic models to track new thematic relations of words like *war* after September 11, 2001. The underlined assumption of their approach is that context shifts or shifts of semantic frames are closely related to topic shifts (Sagi et al., 2013).

We apply this idea to our dataset to track changes

Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 131 - 139

of word associations by topic modeling. Our new contribution is the use of an improved version of topic models, namely neural topic models (Bianchi et al., 2021; Grootendorst, 2022). Neural topic models exploit the advantages of transformer based pre-trained language models and considerably improve the coherence of topics(Bianchi et al., 2021; Grootendorst, 2022). A set representing the topic about fruits is considered more coherent if it contains words that represent fruits such as "apple, pear, lemon, banana, kiwi", not if it contains elements that represent other objects as well such as "apple, knife, lemon, banana, spoon." (Bianchi et al., 2021). Previous work has shown that adding contextual information to neural topic models provides a significant increase in topic coherence, which is missing in Bag-of-Words representations (Bianchi et al., 2021; Grootendorst, 2022). Incorporation of contextualized representations can thus improve a topic model's performance. By using improved topic models, we hope to improve analyses of context shifts of words uses.

In this work, we exploit a particular version of neural topic models, namely BERTopic (Grootendorst, 2022), which includes three ingredients: (1) a specific version of pre-trained neural language model BERT (Devlin et al. (2019)) to obtain contextualised representations, (2) additional semantic clustering of these representations, and (3) calculation of topic words on the basis of c-TF-IDF, which we define in §2.5 and in the Appendix B. Despite other existent neural Topic Models such as combined TM and Top2Vec(Bianchi et al., 2021; Angelov, 2020), we have chosen BERTopic, because it is an appropriate method for modeling changes in a corpus containing short messages as it contains c-TF-IDF method (Ghosh et al., 2017; Wang and Deng, 2017). This method is particularly useful for analyzing corpora comprised of short documents such as tweets.

We make the following contributions:

1. We present an approach of using neural topic models to measure context shifts of words that make use of state-of-the-art contextualised word representations.

2. We use this approach on short documents, namely tweets, from before and after COVID-19 related lockdown in April 2020.

3. We provide a quantitative and a qualitative analysis of context shifts by comparing the use of COVID-related words per topic.

Overall, our study demonstrates the potential of using neural topic models for analysing context shifts of words that have preserved their lexical meaning and are thus difficult to capture by a local analysis.

The paper is organized as follows. We first start with the traditional frequency and n-gram approaches that show which words have increased in relative frequency and which words have obtained new linguistic neighbors. We then apply unsupervised approaches to lexical semantic change that make use of Word Embeddings. Finally, we apply unsupervised topic analysis to capture thematic relations between words and context shifts. The paper finishes with a discussion and evaluation of these approaches.

2 Methodology

2.1 Data collection

We have used English tweets from the Social Media platform Twitter divided into two periods: before and after the lockdown in April 2020. The four countries with the highest number of tweets in our data set are: 1. United States, 2.Canada, 3. UK and 4. Australia. We have a similar number of tweets per country in the two periods. Before the lockdown, data is distributed in 3 years: from 2017 Oct we have around 9k, from 2018 January around 12k and the rest 59 k from 2019 June. The data after the lockdown was collected from just after lockdown 2020 April, which includes around 76k. The number of tweets covered in both datasets were around 80K. After doing all the necessary pre-processing steps, the number of unique words were around 90k in both datasets (94k before lockdown and 87k after lockdown).

2.2 Relative frequency analysis

We filtered out the most frequent words (Top words) that appeared before and after lockdown and calculated the relative frequency as well as the difference in relative frequency of these words. The details of this approach can be found in the table A. The theoretical prediction of the frequency method is that if a word gets an additional meaning over time, its relative frequency will also rise.

The list in table 1 contains most frequent words from the two periods and their relative frequency. We highlighted the words that do not appear before the lockdown.

	Word	rf before	rf after	rf diff
0	home	0.012	0.03	0.025
1	**quarantine**	0.0	0.02	0.0196
2	easter	0.00004	0.02	0.0193
3	**covid**	0.0	0.03	0.015
.	.	.	.	.
8	stay	0.0024	0.0126	0.010
.	.	.	.	.
15	**coronavirus**	0.0	0.008	0.008
.	.	.	.	.
18	safe	0.0009	0.0075	0.0066
19	**stayhome**	0.0	0.0065	0.0065
.	.	.	.	.
24	social	0.0015	0.0075	0.006
.	.	.	.	.
33	**lockdown**	0.0	0.0054	0.0054
.	.	.	.	.
42	**distancing**	0.0	0.0049	0.0049
.	.	.	.	.

Table 1: Top 50 partial list of words with relative frequency differences. Words that do not appear before the lockdown are in bold

In table 1, we see that the increase in relative frequency of words before and after lockdown is in many cases pandemic related as in lines 1,3,15,33 and 42. However, contextual information is missing to evaluate the increase of relative frequency of other words and to detect a potential lexical change.

2.3 N-gram analysis

The n-gram analysis can capture lexical meaning changes by differences in collocation neighbors and differences in the likelihood of the bigram and trigram (Manning and Schutze, 1999). We demonstrate this point by using Bigram and Trigram Collocation finder packages from nltk library (Bird et al., 2009). We merged all the documents as a list of words (around 1 Million words both in before and after lockdown datasets) and reported the top 5 collocations of the words *distance* and *mask* based on likelihood ratio. We see a change in collocations and likelihood ratio in Table 2 in two periods of these words (Butler and Simon-Vandenbergen, 2021; Luo, 2021). However, a collocation analysis does not capture the global context of word meanings. Take the word *mask*, for instance. Table 2 shows that this word was more often used as 'face mask' after the lockdown. However, what also changed after the lockdown is that the word *mask* is now used in a different pragmatic context than before the lockdown, namely in the everyday life practices around the world including western countries, where wearing masks was not part of everyday life practice before the lockdown. To capture this effect, a different approach is needed that includes a more global contextual information of word senses.

2.4 Word Embeddings

The representation of word meanings by Embeddings is nowadays a very standard approach (Giulianelli et al., 2020; Devlin et al., 2019; Mikolov et al., 2013). We follow this line of approach to use word meaning representations by Word Embeddings to capture semantic changes in times on our dataset. For creating Word Embeddings, we have used Gensim word2vec package[1]. We also used Google's Word Embeddings built before the lockdown as a second model for testing.

The analysis in table 3 shows that Word Embeddings for the word *distance* change over time as the top words associated with this word are not the same before and after the lockdown. Compare *hiking* and *film* before the lockdown in column left with *distancing* and *practicing* after the lockdown in column right. The meaning of the textually related words after the lockdown are clearly more pandemic related as evidenced by the word *practicing* (e.g. *practicing social distancing*). Note that the size of the dataset as evidenced by Google's dataset from before the lockdown in the Middle Column does not change the fact that the word *distance* has different word representations from before and after the lockdown. However, Word Embeddings do not capture the global context of word meanings either, if we think about *face masks* and their use in various contexts of our everyday life. To capture contextual meaning shift, we use Topic Modeling as an approximation to track pragmatic meanings or semantic frames of new word senses (Sagi et al., 2013).

2.5 Topic modeling approach

We applied BERTopic (Grootendorst, 2022) for creating topics from the set of sentences. We used our dataset without stop-words for this purpose. BERTopic is a topic modeling technique that leverages transformers and c-TF-IDF to create dense clusters allowing for easily interpretable topics

[1] https://radimrehurek.com/gensim/auto_examples/tutorials/run_word2vec.html

Word	Before (top 5)	After (top 5)
distance	**(walking, distance), 36.75** (distance, summerofyes), 22.08 (mincing, distance), 22.08 (twxn__, distance, 22.08) (long, distance), 18.45	**(social, distance), 752.43** (safe, distance), 71.17 (distance, learning), 69.44 (distance, runner), 22.64 (distance, cruise), 21.41
mask	**(eye,mask), 26.53** (blackface, mask), 22.19 (firespitter, mask), 22.19 (mask, colorsofbeauty, 22.19) (mask, mermay), 22.19	**(face, mask), 496.4** (yashicamm, mask), 244.43 (mask, covid), 107.83 (a, mask), 99.56 (covid, mask), 35.02

Table 2: N-gram analysis of 2 words: *distance and mask*

Before lockdown	Google	After lockdown
hiking, 0.92	distances, 0.75	distancing, 0.94
film, 0.91	Distance, 0.55	practicing, 0.91
race, 0.91	withing_striking, 0.541	holi, 0.84
competition, 0.91	SMA##_remained_##.##, 0.53	self, 0.81
views, 0.91	Distances, 0.51	practice, 0.81
sweat, 0.91	visiting_http:www.newswire.cawebcast, 0.51	distancin, 0.80
riding, 0.91	Chainsaws_hummed, 0.51	donation, 0.80

Table 3: Top most 7 word embedding comparison of word *distance* between before and after lockdown and Google Word Embedding

whilst keeping important words in the topic descriptions (Appendix B). The inverse document part of classic TF-IDF measures how much information a term provides to a document. However in c-TF-IDF, the whole cluster is considered as a document and hence the top 10 terms or topic words become representative of the cluster. The c-TF-IDF method can be used to scale better and works even when topic reductions are used (Grootendorst, 2022). The model produced 202 topics from before lockdown dataset and 220 topics after the lockdown dataset. A topic is represented by a list of 10 Topic words. We define pragmatic contexts or semantic frames as topics and investigate word distributions per topic to track word contexts. We suggest two analyses of word distributions per topic. The first analysis represents distributions of words as topic words and the second analysis represents distributions of words as tokens. By looking at distributions of words as topic words in the first analysis, we capture only frequent words and their contexts or topics. The latter analysis allows us also to capture contexts of less frequent words. The differences in distributions in time will inform us about context shifts of words.

2.5.1 Distribution of words as topic words per topic

Table 4 represents words that are used as topic words after the lockdown, but not before the lockdown. We have already seen in the frequency Table 1, which words appear only in the dataset after the lockdown. However, looking at new words as topic words provides us much more information. Table 4 not only informs us about which words became much more frequent, but also in which contexts or topics these words occur. For instance, the most frequent topic of the word *lockdown* is related to the pandemic situation as evidenced by words such as *coronavirus, covid, virus* (Topic 24) and thus gives us insights about the cause of the lockdown. Other less frequent topics with the word *lockdown* inform us about where the lockdown occurred (Topic 70) and what the consequences of the lockdown are (Topic 135). The most frequent topic with the word *virus* as a topic word is connected to the lockdown (Topic 24). Less frequent topics are associated with locations where the virus occurred and their effects on social practices.The words *quarantine* and *stay* both appear in Topic 5, which is a topic about suggestions to *stay home*, to *cook* and *chill* during the *quarantine*. The word

Word	Number of Topics		Topic IDs
	Before	After	After
mask	0	1(465)	18
quarantine	0	1(1427)	5
stay	0	4(954)	5,24,33,145
distance	0	1(353)	13
lockdown	0	3(262)	24,70,135
corona	0	3(560)	24,34,90
virus	0	3(85)	24,34,90

Table 4: Distribution of words as topic words per topic before and after lockdown dataset. The number of tweets is given in the brackets. Topic IDs refer to topics from table 7

distance appears in Topic 13 about *practicing social distancing* in *parks* and by *hiking*. The word *mask* appears in Topic 18, which represents preventive measures against virus infection such as *facemask, hand gloves*. The word *stay* appears in four different topics as a topic word, that are related to suggestions to *stay home, stay safe* and *stay healthy*.

Note that some terms in Table 4 appear in the same topics such as the word *corona, virus* and some terms share common topics such as the words *stay, distance, corona, virus* (Topic 24). This observation emphasizes the thematic relatedness of these words with the COVID-outbreak. However, the frequency of the words as topic words is not equally distributed per topic as table 4 shows. The word *mask* is more prominent in Topic 18, whereas *quarantine* and the word *stay* are more prominent in Topic 5. This observation emphasizes the specific contexts of use of these words and their specific meanings.

2.5.2 Distribution of words as tokens per topic

The second analysis represents the distribution of words as tokens and not as topic words per topic (Table 5). It shows that words like *mask, stay, distance* changed their context in time by appearing in a much wider range of topics after the lockdown than before and that these topics cover many topics of our everyday life experience. This means that these words are used in conversations about drinking beer with friends, music events, eating pizza, having a haircut and other mundane topics. For instance, the word *mask* appears in only 3 topics as a token before the lockdown, namely in topics about casino in Las Vegas, homosexual activism (LGBTQ) and commercial discount (Topics 20,23 and 59 in Table 6). Since the lockdown, the word *mask* appears in many more topics, namely in 33 different topics. The most salient topic, i.e. the topic with the highest number of tweets, is the topic about preventive measures against virus infection (Topic 18) as already shown in Table 4. In addition to this salient topic, *mask* appears in topics about everyday life activities and events such as tweets about Easter (bunny, eggs) in Topic 3, tweets about food in Topic 4, tweets about photography and selfies in Topic 22, Topic 41 about reposts of tweets (Table 7). The number of tweets containing the word *mask* in these topics representing everyday life activities is much lower than in the salient Topic 18 about preventive measures against virus infection. However, it is considerably higher than the number of tweets with the word *mask* before the lockdown. The wider range of topics of the word *mask* after the lockdown can be therefore considered as an indicator for a contextual change of this word.

The most frequent collocations in (Table 5) show that in both periods *mask* is used as *face mask* in the most salient topics. Just by looking at collocations, we do not know how *face mask* is used before and after the lockdown. This emphasizes our criticism of local analyses in §2.3. The topic descriptions of the word *mask* adds contextual information about this word, which is why a topic analysis is a better analysis. However, collocations can also change with a topic change as (Table 5) shows. For instance, the word *stay* is not only used in different topics, but also in different collocations before and after the lockdown. This said, context shifts or topic shifts of words can correlate with collocation shifts, but they do not need to. It is this important observation that motivates the use of our method.

To sum up, we have shown that a topic analysis provides information about context shifts of word uses and the change of thematic word relations in time.

3 Conclusion

We have introduced a novel dataset that contains lexical change triggered by the COVID-related outbreak. We have used this dataset to discuss different analyses capable of capturing linguistic change, namely the relative frequency analysis, the n-gram analysis and lexical change captured by Word Embeddings. We have shown that these analyses miss

Word	Number of Topics		3 Top collocations of tweets in Topics	
	Before	After	Before	After
mask	3	33	'face','mask', 16.67 'mask','look', 12.56	'face','mask', 379.94 'wear','mask',118.00 'wearing', 'mask', 98.31
quarantine	0	42		'quarantine','day', 105.40 'quarantine','quarantinelife',73 'quarantine','stayhome', 58.57
stay	56	121	'stay', 'tuned', 256.21 'stay', 'hydrated',32.65 'stay','focused', 22.22	'stay','home', 1085.74 'stay', 'safe', 1019.46 'stay', 'tuned', 371.55
distance	6	49	'walking','distance', 41.62 'mincing','distance',19.78 'twxn__', 'distance', 19.78	'social','distance', 386.603 'distance','learning',49.22 'keeping', 'distance', 47.11

Table 5: Distribution of words as tokens before and after lockdown dataset. Top 3 collocations are calculated only considering the tweets of the topics.

an important aspect of meaning change, namely the pragmatic aspect. This meaning change represents a change of cultural or everyday practices associated with words such as *mask*. We suggested tracking the pragmatic change via Topic Modeling by looking at the distribution of words per topic in two different periods. We discovered that topics capture the contextual meaning of a word by the textual association with other words. Changes of word distributions in topics can give us insights about pragmatic meaning change of words.

Exploring context shift by neural topic models falls into the family of neural models used to track and measure language change by contextualized word representations (Giulianelli et al., 2020; Del Tredici et al., 2019). In this sense, our contribution is very much related to this work as all these models have a similar architecture and capture the meaning of words. However, we use neural contextualized embedding for an improved version of topic modeling and use then topics as proxies for word contexts. This method allows us to explore more global relations between words by looking at their relations to other words in the same document or text. We admit that this is a very approximate approach of capturing the pragmatic aspect of words with an unsupervised method. One important issue of this approach is that it is very much dependent on the data size and the size of each document or tweet, which influence the quantity of topics and the topic specification. Another important point is that it does not capture the very many implicit discourse relations between Topic words such as causal relations between the lockdown and the virus in Topic 24 in Table 7. One way to approach this issue is to use unsupervised approaches of capturing discourse relations between Topic words of the same topic (Liu and Lapata, 2018), which we reserve for future work.

4 Acknowledgments

This project has received funding from the German Research Foundation (DFG) 2021 (No. KE 2048/1-1). We would like to thank Dr. Nicholas Matlis, the Language Technology Group of the university of Hamburg for providing us useful feedback as well as the anonymous ACL reviewers for their helpful comments.

References

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: contextualized document embeddings improve topic coherence. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

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

Christopher S Butler and Anne-Marie Simon-Vandenbergen. 2021. Social and physical distance/distancing: A corpus-based analysis of recent changes in usage. *Corpus Pragmatics*, 5(4):427–462.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of NAACL-HLT*, pages 2069–2075.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Charles J Fillmore, Collin F Baker, and Hiroaki Sato. 2002. The framenet database and software tools. In *LREC*.

Samujjwal Ghosh, PK Srijith, and Maunendra Sankar Desarkar. 2017. Using social media for classifying actionable insights in disaster scenario. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, 9(4):224–237.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

George Lakoff. 2008. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press.

Yang Liu and Mirella Lapata. 2018. Learning structured text representations. *Transactions of the Association for Computational Linguistics*, 6:63–75.

Huan Luo. 2021. How has the coronavirus pandemic affected our use of language? a corpus-based study of neologisms and semantic shifts in english and chinese web texts.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Eyal Sagi, Daniel Diermeier, and Stefan Kaufmann. 2013. Identifying issue frames in text. *PloS one*, 8(7):e69185.

Hao Wang and Sanhong Deng. 2017. A paper-text perspective: studies on the influence of feature granularity for chinese short-text-classification in the big data era. *The Electronic Library*.

A Creation of Top words

After some standard pre-processing, we created Top words from the datasets, by the following steps:

1. Took the list of sentences and removed stopwords

2. Create a two dimensional vector for each words and documents. For example if we have 100 sentences and 100 unique words in the whole set, the resulting dimension will be (100*100). The package was used from sklearn: CountVectorizer.

3. Extracted and saved the following parameters for each word in the dataset.

   (a) Total found (total_occur): The number of times in total the word appeared in the whole dataset.

   (b) Number of documents (number_docs): The number how many documents the word appeared.

   (c) Relative frequency (rf):
   rf = Total_found / Total Documents.

   (d) Cumulative score (cs): A scoring system to give emphasis on number of documents the word occurred by multiplying it with relative frequency.
   Cs = rf * number_docs

4. Sorted the words in the dataset based on cs scores and reported Top 100

B Topic Modeling

BERTopic (Grootendorst, 2022) is a topic modeling technique that leverages transformers and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions.

B.1 Our implementation procedure:

We have used BERTopic for creating topics from the set of sentences. We have used our dataset without stop-words for this purpose.

The process is as below

1. Created topics using the model
2. Created dictionary of topic words.
3. Filtered out the topics which matches any of the covid related words

B.2 Topic modeling Theory

Topic modeling technique LDA (Blei et al., 2003) is well known but has some limitations like bag of words, Fixed K (the number of topics is fixed and must be known ahead), non hierarchical, etc. BERTopic on the other hand does not pose these problems.

The procedure how BERTopic works can be divided in 3 steps:

1. Converting sentences into embeddings : The first step is to convert the documents into embeddings. BERT is used to create the embeddings.
2. Clustering the embeddings based HDBScan (McInnes et al., 2017) (a density based Unsupervised clustering technique). This stage comprises two parts: Dimensionality reduction and Clustering. UMAP (McInnes et al., 2018) is used for Dimensionality reduction and Hierarchical Density based clustering is used for Clustering the embeddings.
3. cTF-IDF : Finally, class based TF-IDF (cTF-IDF) is used to extract words that represent a clustering.

   $W_{t,c} = tf_{t,c} \cdot log(1 + \frac{A}{tf_t})$.

   Where the term frequency models the frequency of term t in a class c or in this instance. Here,the class c is the collection of documents concatenated into a single document for each cluster. Then, the inverse document frequency is replaced by the inverse class frequency to measure how much information a term provides to a class. It is calculated by taking the logarithm of the average number of words per class A divided by the frequency of term t across all classes. To output only positive values, we add one to the division within the logarithm (Grootendorst, 2022).

Topic ID	Topic words	Number of tweets
20	vegas, las, casino, lasvegas, vegastraffic, nv, nevada, hotel, clark, accident	280
23	pride, gay, lgbtq, pridemonth, month, lgbt, happy, gaypride, rainbow, loveislove	262
59	code, discount, discountcode, fwcom, bestprice, fyi, orders, extra, get, sexy	118

Table 6: Topics related to COVID words which appeared as a token from before lockdown dataset

Topic IDs	Topic words	Number of tweets
3	easter, birthday, happy, bunny, family, happyeaster, everyone, eggs, sunday, hope	2147
4	pizza, dinner, chicken, cake, garlic, cookies, sauce, pork, rice, fried	1394
5	quarantine, quarantinelife, quarantined, day, stayhome, life, quarantinecooking, cooking, best, quarantineandchill	1241
13	distancing, social, park, hike, walk, trail, socialdistancing, distance, practicing, hiking	718
18	mask, masks, face, skin, wear, facemask, dermatology, wearing, hand, gloves	628
22	photography, camera, photographer, selfie, portrait, photos, streetphotography, pictures, model, photooftheday	581
24	coronavirus, covid, virus, pandemic, corona, lockdown, stayhome, update, tests, outbreak	543
33	amp, safe, call, stay, got, back, keep, need, many, things	393
34	francisco, san, california, angeles, los, thoughts, coronavirus, diego, photo, posted	373
41	repost, getrepost, reposted, makerepost, talkkellyzola, makeyourselfhappy, onlinetradefair, makeyourselfproud, iamyourlovestory, thanks	285
47	run, miles, running, mile, ran, runner, marathon, ismoothrun, race, runners	206
70	lockdown, isolation, locked, portelizabeth, self, christchurch, lock, cuenca, zealand, europa	149
90	stigma, fighting, ireland, stigmabase, hong, kong, china, coronavirus, northern, health	106
135	notes, unreliable, lockeddown, testing, data, proverty, lockdown, rate, heavily, adequate	59
145	staysafe, weloveourhealthcareworkers, stayhome, stayhealthy, greenwich, stay, gratitude, village, staystrong, healthy	45

Table 7: Topic words related to COVID found in the dataset after the lockdown.