Social Network Analysis and Mining (2023) 13:119
https://doi.org/10.1007/s13278-023-01125-5

ORIGINAL ARTICLE

Reactions to science communication: discovering social network topics using word embeddings and semantic knowledge

Bernardo Cerqueira de Lima[1] · Renata Maria Abrantes Baracho[1] · Thomas Mandl[2] · Patricia Baracho Porto[3]

Received: 22 June 2023 / Revised: 30 August 2023 / Accepted: 31 August 2023 / Published online: 22 September 2023

Abstract

Social media platforms that disseminate scientific information to the public during the COVID-19 pandemic highlighted the importance of the topic of scientific communication. Content creators in the field, as well as researchers who study the impact of scientific information online, are interested in how people react to these information resources. This study aims to devise a framework that can sift through large social media datasets and find specific feedback to content delivery, enabling scientific content creators to gain insights into how the public perceives scientific information, and how their behavior toward science communication (e.g., through videos or texts) is related to their information-seeking behavior. To collect public reactions to scientific information, the study focused on Twitter users who are doctors, researchers, science communicators, or representatives of research institutes, and processed their replies for two years from the start of the pandemic. The study aimed in developing a solution powered by topic modeling enhanced by manual validation and other machine learning techniques, such as word embeddings, that is capable of filtering massive social media datasets in search of documents related to reactions to scientific communication. The architecture developed in this paper can be replicated for finding any documents related to niche topics in social media data.

Keywords Pandemic · Topic modeling · Machine learning · Communication

1 Introduction

The COVID-19 pandemic has altered the information-seeking behavior of people, much like other crises. This is evident from the shift in web popularity rankings, which indicates that individuals have specific information needs during times of crisis communication (Dreisiebner et al. 2022). As a result, they face challenges in determining the reliability and trustworthiness of the information they receive (Barnwal et al. 2019). Social media channels have become important sources of diverse scientific communication content aimed at meeting these needs. Given the significance of scientific understanding during a crisis, it is crucial to comprehend the patterns of information-seeking behavior (Montesi 2021). In particular, media creators must understand the quality criteria that users utilize in selecting their resources and the factors that influence their preferences. While information resources in general and multimodal science communication, in particular, differ in their portrayal of scientific information, research has not delved into the detailed examination of the most effective ways of disseminating scientific information.

Renata Maria Abrantes Baracho, Thomas Mandl and Patricia Baracho Porto have contributed equally to this work.

✉ Thomas Mandl
mandl@uni-hildesheim.de

Bernardo Cerqueira de Lima
bernardolima95@gmail.com

Renata Maria Abrantes Baracho
renatabaracho@arq.ufmg.br

Patricia Baracho Porto
pattybarachoporto@gmail.com

[1] Federal University of Minas Gerais, UFMG, Belo Horizonte 31270–901, MG, Brazil

[2] Information Science, University of Hildesheim, 31141 Hildesheim, Germany

[3] Pontifical Catholic University of Minas Gerais, Belo Horizonte 30535-901, MG, Brazil

With the aim of examining the online conversation surrounding the Corona crisis, specifically in the context of science communication, the goal of this study is to develop a solution that can help creators filter out feedback in the middle of millions of comments made in relation to their work. Our focus is on identifying and analyzing a specific subset of comments that users post in science communication channels as responses to the content presented. For the goals of this study, we collected 1.12 million tweets formed by the comments in a network of Brazilian scientists, governmental bodies, doctors and scientific communicators. It is worth noting that the majority of these comments are posted at scientific channels are not necessarily related to the content or format but rather pertain to the broader discourse around the Corona crisis. They often include political viewpoints and general comments on the crisis.

First attempts into filtering out this data were performed with topic modeling through traditional algorithms such as Latent Dirichlet (Blei et al. 2003), which proved to not be good enough in encountering niche topics when dealing with short documents in large data collections (Lima 2023; Mandl et al. 2023). After manual validation of our topic model, an ensemble method of document filtering was created through the creation of a word dictionary made out of the most relevant words in topics relevant to scientific communication and their top-n closest neighbors according to the cosine similarity of their word embeddings. A filtering heuristic that uses this dictionary was put in place, and managed to lead to severe improvements in our ability to filter documents related to scientific communication. With such techniques, this paper introduces an ensembling framework that combines topic modeling and word embeddings as steps that retroactively feeds data processing, generating a very effective dataset for machine learning tasks that need to identify niche topics in large databases, as seen in Fig. 1. The data and solution obtained from this study allows for an examination of tweets that were positively received in relation to science communication strategies during the COVID-19 pandemic. This information can be utilized to enhance the dissemination of scientific information during future crises. The solution devised is also theme-agnostic, and could be applied to other instances of filtering out niche topics in social media data.

2 Related work

The research of popular science communication channels indicates comparable communicative techniques taking into account aspects such as scientific insecure communication, factuality, complexity, emotionalization, and expert presentation. Because social media communication is multimodal, visual information, location, and body language are all important. Understanding how epidemiological information is transmitted and understood by non-expert audiences requires extracting meaningful comments from internet audiences (Jaki 2021). These could be helpful for improving communication strategies and for adapting scientific information to online audiences.

Previous research on public scientific material relating to the COVID-19 pandemic, on the other hand, has primarily concentrated on qualitative analysis, ignoring internet audience reactions (Bucher et al. 2021). While datasets available for social media communication about Corona are available, it is a considerable challenge to filter this volume of information, being shown that social media information propagation during the pandemic has largely shone a focus around general trends, political attitudes, and the spread of misinformation, which surpasses true information in spread (Vosoughi et al. 2018).

In such, there is a need for comprehensive studies on the quality of science communication, including topic modeling and classification, specifically exploring reactions to science communication with the public. Several studies have been already conducted in exploring topic modeling in regards to social media data during the COVID-19 pandemic (Boon-Itt and Skunkan 2020; Melo and Figueiredo 2021; Yin et al. 2022), but few tackled the topic of scientific communication. Much rather, such studies take a special perspective. Often machine learning is used to improve a sentiment analysis, e.g., for the attitude toward vaccines in Turkey (Küçük and Arıcı 2022). Similarly, LDA has been used to detect topics in the discourse on vaccines and the documents were further processed with deep learning methods to extract to determine the sentiment in a large collection (Zulfiker et al. 2022).

In a previous attempt to focus solely on tweets reacting to science communication, LDA modeling found only few comments (Lima 2023).

When analyzing short textual data, topic modeling techniques are helpful for clustering similar documents within a corpus and for finding common themes they may share. While Latent Dirichlet Allocation (LDA) (Kalepalli et al. 2020) exhibits a strong performance in the diversity of their generated topics and in its accurate categorization, more modern solutions based on transformer architectures such as BERTopic and Top2Vec, which are based on ensembled techniques and word vectorization, achieve better results in topic coherence (Egger and Yu 2022a) and have an array of different modeling strategies that tend to generalize toward more interpretable topics. Although there are comparative evaluations for LDA, BERTopic and other models, they do not give advice for filtering small amounts of tweets with high accuracy (Egger and Yu 2022b).

When looking for such niche topics, Guided BERTopic's strategy of nudging topics together by their word similarity

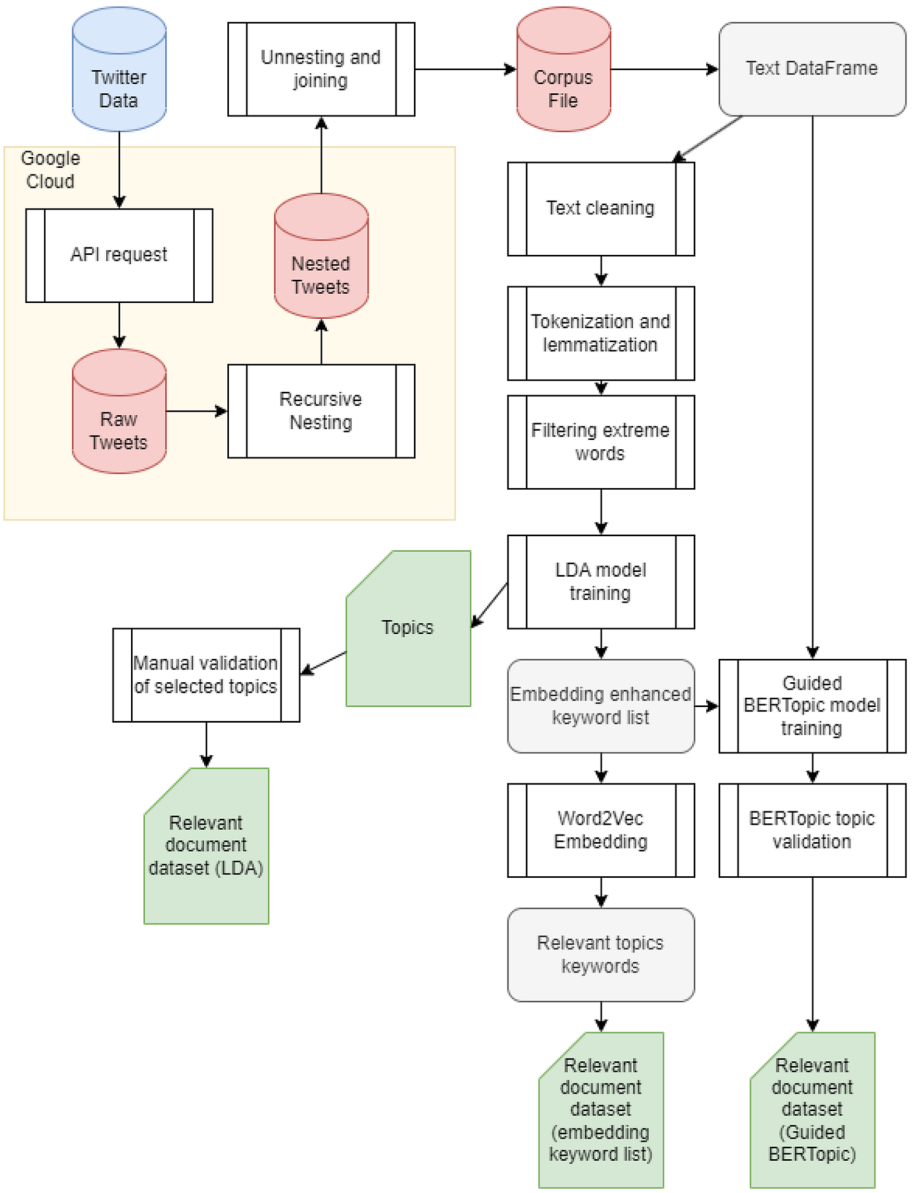

Fig. 1 The complete architecture and methodology applied in this study

produces more interesting results when looking for a specific theme inside a corpus (Grootendorst 2022).

Futhermore, most of this work is done for English (Ng et al. 2022), so research for other languages on the COVID-19 discourse is necessary.

Although machine learning techniques and topic modeling has come a long way, the means of measuring topics by their human interpretability are still unclear on which are the best practices, with the often used coherence metric not being able to accurately depict actual human interpretability (Ramírez et al. 2012) and a comparison to human judgment is difficult to quantify (Chang et al. 2009).

Interestingly, metrics based on the utilization of word embeddings, such as measuring for cosine similarity, reflects into more accurate measurements of interpretability (Doogan and Buntine 2021). The construction of word dictionaries used to aid machine learning techniques in better handling of a specific problem set has been proven to be effective (Reveilhac and Morselli 2022), and is also used in several strategies of BERTopic. This approach is specifically helpful when dealing with word embeddings, with the utilization of Google's Word2Vec model in conjunction of keyword lexicon lists proving that the mixture of techniques tends to prove successful, enhancing the capability of word embeddings inside a model, specially when dealing with filtering specific information which form a minority class inside a larger dataset (Hu et al. 2017; Koufakou and Scott 2020; Jin et al. 2018).

3 Methodology

This study went through an iterative process between techniques: from beginning with an LDA topic model, we repurposed its results into a validated dataset that was used to train a Word2Vec model, as well as nudging BERTopic into better generalizing our niche topic.

In the first stage, relevant science communication channels on Twitter were identified. Then, data from these sources containing content created during the COVID-19 pandemic were collected and underwent several textual data processing techniques, resulting in a final corpus that was prepared for natural language processing. Lastly, a topic model was created to organize the massive text data into certain themes. The goal was to identify topics that grouped terms related to reactions to scientific communication. The effectiveness of the topic model was evaluated using the NPMI metric.

The results of these topic models were then manually analyzed by our team and classified into certain categories that pertained to science communication and if they were relevant or not. Relevant documents were those which commented on the design or content of the scientific postings.

The most relevant words for the topics with the largest concentration of relevant documents were sorted into lists by their categories. From these words, a Word2Vec model was trained on a lemma database formed from the corpus. Then, embeddings were created for the sorted word lists, and their top-5 nearest neighbors were added to the lists. With these list of words, we compared them to the corpus, using the trained Word2Vec model, and measured their proximity by cosine similarity and also by overall word count. Using this mixture of techniques, we managed to filter a considerably larger amount of relevant documents than by relying solely on topic modeling. The results were analyzed and validated, as well as compared against a BERTopic model trained on the database.

3.1 Data collection and processing

3.1.1 Data collection

We manually selected 46 sources for the Brazilian market based on their relevance to COVID-19 discussion, mostly comprising doctors and research institutes.[1] Their relevance was measured by their follower count and amount of reactions (comments and likes) to their posts, as well as handpicking for official sources (such as official governmental bodies and health professionals with ties to the Ministry of Health). Additionally, we included popular science communicators and news aggregators during the pandemic. To collect data, we made requests to the Twitter API to retrieve tweets, retweets, and replies from these sources between March 1st, 2020 and March 1st, 2022, resulting in 1.3 million tweets. We organized the data into nested JSON files that reflected the website's complex structure, only collecting tweets flagged with the Portuguese language tag for this study. We believe that this collection reflects a considerable sample of the most popular and relevant scientific communicators in Brazilian social media.

Due to the massive amount of data collected and Twitter's API and computing power limitations, we conducted this step on a Google Cloud cluster with three virtual machines. This decision allowed us to maintain continuous usage of computing resources during the collection and nesting process and ensured fault-tolerance.

3.1.2 Text processing

The initial steps involved consolidating all the tweets into a single dataset and filtering out any tweets made by the original source, only retaining replies. The text content of each tweet was then processed to remove URLs, special characters, emojis, and mentions, leaving only the actual text. To reduce noise in the text data, the next step was to remove stopwords. We utilized a combination of four stopword lists: a custom list developed for the study, the Spacy (Honnibal and Montani 2017) Portuguese News stopwords list, the Gensim (Rehurek and Sojka 2011) NLP Python library, and the Wordcloud (Oesper et al. 2011) stopwords list.

Once the stopwords were eliminated, the text was tokenized using Python Spacy's library rule-based function. Another filter was then applied to ensure that only words longer than three characters were considered valid tokens. These tokens were then lemmatized using Spacy's library, which uses both rule and lookup-based methods to reduce words to their lemmas. To further reduce overfitting and noise, words that appeared in less than two documents and those that appeared in more than 99% of the documents were filtered out.

3.2 Topic modeling

Topic modeling is a technology within Text Mining (Mandl 2015) which tries to identify the topic structure of large text collections.

3.2.1 LDA

LDA is a statistical method used for topic modeling, with the aim of identifying recurring patterns in a collection of documents (Kalepalli et al. 2020). In this study, the objective was

[1] Examples are https://twitter.com/luizacaires3, https://twitter.com/ocienciaetal.

to develop a topic model that would uncover topics related to reactions to scientific communication. LDA was chosen as the algorithm to be used because it estimates the probability distribution of each word belonging to a particular topic, given a fixed number of topics.

$$P(w|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)(\prod_{d=1}^{N_d} \sum_{Z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta))\mathrm{d}\theta_d \quad (1)$$

Equation 1 represents the three levels in LDA, with variables α and β representing document-topic density and topic-word density, respectively. These parameters are defined for the whole corpus, with Θ being the topic distribution of the document, z a set of N topics, and w a set of N words.

In order to determine the optimal number of topics, the study used the normalized pointwise mutual information metric. This metric evaluates the topic model by representing the top-n words of each topic as a vector in a semantic space, calculating their probability of co-occurrence with each other and weighting these vectors by the NPMI of each term. It is important to experiment with a higher number of topics to identify a specific and narrow topic while still producing interpretable results. In this study, we utilized the Python Gensim (Rehurek and Sojka 2011) libraries for LDA and its performance metrics, such as coherence and NPMI.

3.2.2 NPMI

The NPMI metric is a useful way to evaluate the quality of the topic model because it takes into account the co-occurrence of words, which can reveal underlying patterns and relationships between topics. By representing the top-n words of each topic as vectors in a semantic space, it is possible to calculate their probability of co-occurrence and weight them by the NPMI of each term (Aletras and Stevenson 2013).

$$\mathrm{PMI}(w_i, w_j) = \frac{p(wi, wj)}{p(wi)p(wj)} \quad (2)$$

The PMI metric is used to calculate the probability of two words occurring together, taking into account the probability of each word occurring individually.

$$\mathrm{NPMI}(w_i, w_j) = \frac{\mathrm{PMI}(wi, wj)}{-\log(p(wi, wj))} \quad (3)$$

By normalizing PMI with −log(p(wi, wj)), the NPMI metric is able to reduce the impact of rare co-occurrences and increase the weight of more common ones.

By experimenting with different numbers of topics and evaluating the resulting topic models using the NPMI metric, it is possible to find the optimal number of topics for the specific dataset and research question at hand. This approach can help researchers identify meaningful and interpretable topics that capture the underlying themes in the data.

3.2.3 BERTopic

While techniques such as LDA perceive documents as a bag-of-words, BERTopic is a topic model that utilizes clustering techniques and class-based TF-IDF to model topics in such a way that the semantic relationship between words within a single document is taken into account (Grootendorst 2022).

BERTopic is based on BERT (Devlin et al. 2019) and utilizes its capacity for generating vector representations of words and sentences with semantic properties. BERTopic works by leveraging a pre-trained language model to create document embeddings, which go through dimensionality reduction and clustering through HDBSCAN. The most relevant words of each cluster are classified through a class-based variation of TF-IDF:

$$W_{t,c} = tf_{t,c} * \log(1 + \frac{A}{tf_t}) \quad (4)$$

In the equation above, $\mathrm{tf}_{x,c}$ represents the frequency of word x in class c, tf_t represents its frequency in all classes, and A is the average word per class. The resulting value represents the importance of a word in a cluster, allowing for the generation of topic-word distributions for each cluster.

BERTopic capabilities of maintaining the semantical property of documents inside topics results in more diverse and coherent topics when compared to LDA, and it is generally more robust in use, enabling for more options in fine-tuning and less dependent on preprocessing.

3.3 Word2Vec

Word2Vec is a model architecture that computes continuous vector representation with words, achieving impressive results in word similarity tasks on very large datasets, at a low computational cost (Mikolov et al. 2013). Its word embeddings are widely used to represent words as vectors.

It utilizes two-layer neural networks trained to reconstruct the linguistic context of words to output a vector space that represents an entire corpus, with each word being assigned a different vector in this space. Word2Vec has two architectures: *CBoW*, in which it uses the surrounding words to predict the word located in the center of an n-gram, and *Skip-gram*, in which it uses the central word in an n-gram to predict the surrounding words.

Word2Vec has proven to achieve interesting results in preparing results for both intrinsic and extrinsic tasks, beating several other word embedding techniques (Schnabel et al. 2015), and forming the theoretical base of state-of-the-art word embeddings (Wang et al. 2019). For the purposes

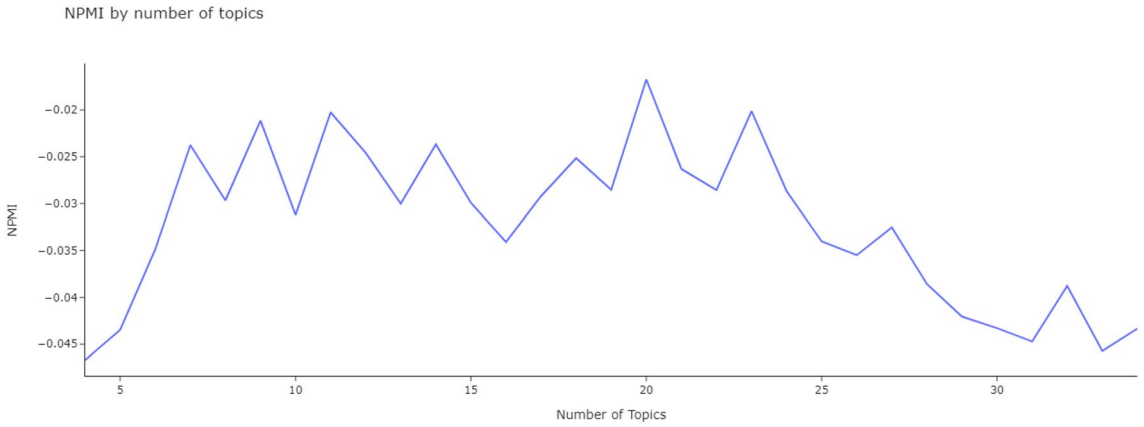

Fig. 2 NPMI coherence search by number of topics

of this study, the Word2Vec implementation utilized was made available in Python by the natural language processing library Gensim (Rehurek and Sojka 2011).

3.4 List based filtering

In this study, we experimented with a mixture of techniques that combined manual validation of topics and the properties of their most relevant words.

Once we manually identified topics that, in their top-30 words, ranked by their relevance (Sievert and Shirley 2014) and saliency (Chuang et al. 2012), contained words that were relevant to the topic of feedback to scientific communication, we set up samples of 1500 tweets from each of these topics.

The research team, composed of three people, with the help of field experts, manually classified them into six categories, in relation to scientific communication:

- Questions, comments, corrections or suggestions about the content or theme at hand, directed toward the author
- Discussion about scientific communication between users
- Praise or criticism toward the content
- Praise or criticism toward the author
- Political commentary in relation to scientific communication
- General questions about the epidemic

The process of manual classification in these categories followed a consensus-based approach, where disagreements were discussed as a group, and a document would be classified when the group reached a unanimous decision. After finishing a batch of documents, a field expert would manually verify each document in accordance with the team's classification. The final decision on classification was up to the expert's discretion.

Since the objective of our study was to identify reactions to scientific communication, these were the categories in which we decided to segregate our documents, we only considered the first four categories as relevant for properly identifying reactions to scientific communication that might prove useful to scientific content distributors.

After identifying the tweets, we selected words from the top-50 words in the respective topics that were present in the relevant tweets and added them to a word list. This list was comprised of 44 words. Then, with the trained Word2Vec model, we generated embedding vectors for each of these words, and selected, from our corpus, the 5 nearest words when measured by their cosine similarity, which has shown promising results when evaluating the semantic similarity of word embeddings (Lahitani et al. 2016). Each of the top-5 nearest words were added to our word list, reaching a count of 264 total relevant words.

4 Results

4.1 LDA topic modeling

At the start of the experiments, several LDA models were executed and scored by their NPMI metric, looking for the optimal number of topics that allowed for maximum coherence and interpretability, as seen in Fig. 2.

In our experiments, we observed that using 20 topics (represented by their intertopic distance in Fig. 3 resulted in the highest NPMI value, with the metric steadily decreasing beyond that number. While fewer topics also yielded decent NPMI results, they failed to adequately capture distinct niches within the dataset, which was necessary for our specific research goal. Given that the topic of interest was not a prevalent theme in the tweets, employing a larger number of topics allowed for more specific and targeted discussions,

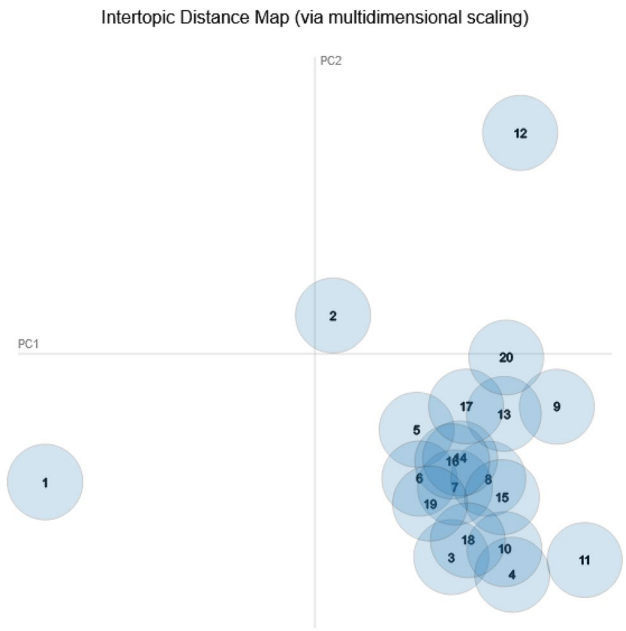

Fig. 3 Intertopic distance mapping between topics found with LDA

facilitating the filtering of tweets unrelated to the theme of reactions to scientific communication.

After careful manual validation of the topics most relevant and salient terms, we decided on exploring two specific topics, manually validating 750 tweets of each one of them, sorted by their contribution toward the topic. In Table 1, we can see translated examples of the topic's keywords and example sentences.

This approach through LDA resulted in very few actual relevant documents encountered within a sample of 1500 documents. While this was expected, given Twitter's propensity for usage for news publishing and discussion rather than educational content, we believed that more relevant avenues for filtering niche documents could be found either in combining embedding models with the results of our manual validation, or in exploring state-of-the-art topic modeling.

4.2 List based semantic filtering with Word2Vec

As discussed, this technique employed the creation of a dictionary of words relevant to the topic of scientific discussion and enhancing this list with the aid of Word2Vec word embeddings. With this list of relevant words at hand, we decided on two metrics for filtering documents: the relevant word count in each document, and their cosine similarity in relation to the entire list. We decided on a cutoff value that would leave us with a sample of similar size of in relation to our LDA experiment: each document contained at least 2 relevant words and had an average cosine similarity larger than 0.6.

This sample was also manually validated with the same criteria as the previous experiments, leading to fantastic results: many more documents were classified as relevant, and these documents were also very in line with the content that we wanted to find. Many more comments were in fact discussing the quality and characteristics of the content presented, as well as providing questions, corrections and additions.

4.3 Guided BERTopic

This experiment's Guided BERTopic model was trained and fine-tuned with a custom KMeans model, a practice that in our experiments with BERTopic led to more varied and coherent topics, and also helped in reducting the amount of themes splintered into several small topics. The model was trained with the list created by the previous experiment as its seed topic list, which nudged BERTopic in creating more topics related to the reaction toward scientific communication.

This resulted in the documents being split into 347 topics of roughly similar size. Going by the relevant tweets that we classified in the previous two experiments, we search for the topics that had the largest amount of those tweets as well as the best ratios of relevant/irrelevant documents. This led us to 10 topics. Refer to Fig. 4 as well as Table 1 for a visualization of the topics keywords and examples of the found documents. A sample of the 150 most relevant documents of each topic was taken to create a third dataset of the same size as the previous two, to be manually validated by our team and our field experts.

4.4 Comparison between techniques

With three same-sized datasets, we can then compare our different techniques in their capability of filtering a large dataset in search of documents of a niche theme (Fig. 5).

We can see that while the topic modeling approach led to much improvement when compared to LDA, the ensembling of word embeddings and list-based semantic filtering found the largest amount of relevant documents. When judging by their general relevance toward the theme, the documents found by the semantic filtering were also in general more related to direct feedback toward the content creators.

Following a classifier algorithms methodology, we can assume that each sample consisting of 1500 documents was completely classified as relevant. As such, we can measure the precision of each technique, as seen in Table 2.

4.5 Summary and contributions

Overall, this paper describes and implements a method to selecting micro-blog entries for topics with a low volume. In

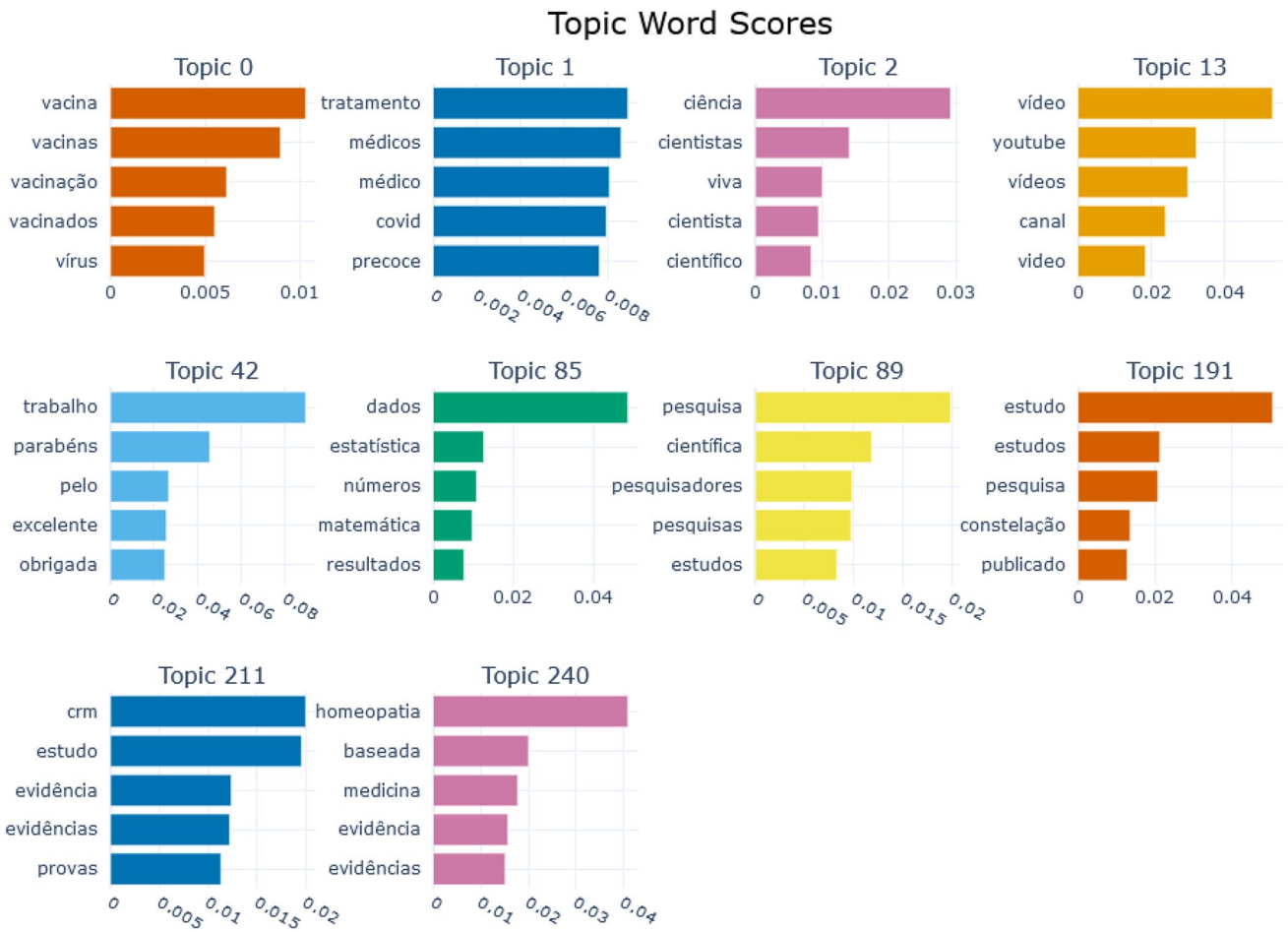

Fig. 4 The 10 chosen topics based on their relevance toward the theme

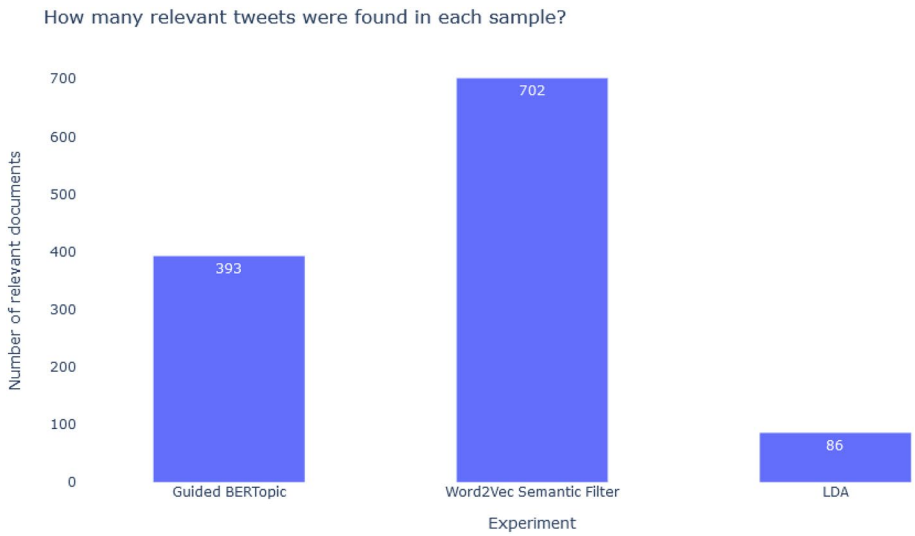

Fig. 5 Comparison between each technique in the amount of relevant tweets found by each one in a similarly-sized dataset

Big Data analysis, such topics might easily get overlooked. The methodology combines topic modeling and manual validation into a common procedure. A framework based on topic modeling, word embeddings as data processing steps was conceived to increase the performance of topic modeling and semantic filtering when searching for niche topics.

Table 1 Keywords and examples sentences of the most relevant topics founds by the topic modeling techniques tested in this study

Selection	Keywords	Sentences found
LDA (Selected topic #1)	truth, essential, working, completely, live, calm, importance, development, version, quality	"This 7th paragraph is not there for applying exceptions for Brazilians." "very good news" "NO ONE went to the hospital." "This young guy did not talk about being cured, but only that he witnessed that"
LDA (Selected topic #2)	vaccinate, coronavirus, anxious, priority, reading, context, absolutely, layman, immune, entity	"Congratulations for this material. Excellent!" "Excellent news for those who believe in the Brazilian scientists." "And I have not seen that! I do not believe it!!"
BERTopic (Selected topics)	vaccine, treatment, scientific, video, work, data, research, study, evidence, medicine	"Great idea! Could this data also be used as an indicator of what is happening until the website is back online?" "If the results are indeed promising, I don't understand why they didn't publish the partial results in the accorded date." "This article doesn't have any scientific value." "Thanks for your work! I thought that the video was really helpful in clearing things up."

Table 2 Metrics of each technique when evaluated by a 1500 document test sample

Technique	Relevant tweets	Irrelevant tweets	Precision
LDA	86	1414	0.06
Guided BERTopic	393	1107	0.26
Word2Vec list filter	702	798	0.47

The methodology could be readily be implemented into a tool for tracking the reactions to a channel. This tool would be able to select and segment reactions which go beyond the frequent political and personal communication that is generally abundant in social media.

Our methodology was applied to the reaction to science communication for the Brazilian market on Twitter. This paper provides a framework that can be used for selecting and analyzing content on social media, and employing it on Brazilian twitter data was the first application of this framework.

Obviously, the selection of channels and the collection includes a subjective element. Furthermore, the download of data from Twitter necessarily leads to problems with repeatability because of restrictions implemented in the Terms of Service and API usability during the year 2023.

5 Conclusion

While this study focused on finding reactions to scientific communication in social media, these same techniques can also be applied when searching for any niche topic in collections of documents similar in size to tweets. We found that even state-of-the-art models such a BERTopic or Word2Vec needed a certain degree of heuristical validation to be able to generalize into niche topics. The approach that required the most amount of manual validation and combination with heuristics, which was the semantic filtering technique with Word2Vec embeddings, was also the most successful in filtering through our dataset and finding our desired topic.

The framework defined in this paper can be applied to help scientific content creators to better locate useful feedback in very large datasets, as well as point out, after a brief analysis, which subjects are the viewers most interested at, and what kind of content delivery are the most effective for good retention and positive feedback. With this, content creators are able to better understand how to broadcast their knowledge to different audiences, and through different avenues.

One big hurdle that this study had to deal with it was the nature of Twitter's data: scientists and content creators do not tend to post educational content on Twitter, and prefer to use it to share important news and foster discussion—this characteristic of the social network led to difficulties in finding feedback toward scientific communication, and, when coupled with Twitter's closing of their academic access API in March of this year, leads our future studies in the direction of gathering documents from other sources such as Youtube.

With a large dataset of thousands of tweets classified by their relevance toward our theme, this study also contributed with the data labellng necessary for an interesting possible next avenue: the training of classifier to identify if a document is relevant to scientific communication or not. This classifier can be seen as the end goal of our study: a tool for scientific content creators to better understand what their audience seek for in their content. With such a tool, scientific communication can be optimized to reach larger audiences.

Acknowledgments This work was enabled and financed by the Volkswagen Foundation in Germany with the Grant A133902 (Project Information Behavior and Media Discourse during the Corona Crisis: An interdisciplinary Analysis - InDisCo). Further financial support was provided by the Coordination for the Improvement of Higher Education Personnel (CAPES) from Brazil.

Author contributions All authors contributed equally.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Competing interests The authors declare no competing interests.

References

Aletras N, Stevenson M (2013) Evaluating topic coherence using distributional semantics. In: International conference on computational semantics (IWCS), pp 13–22. ACl, Potsdam, Germany. https://aclanthology.org/W13-0102

Barnwal D, Ghelani S, Krishna R, Basu M, Ghosh S (2019) Identifying fact-checkable microblogs during disasters: a classification-ranking approach. In: International conference on distributed computing and networking, ICDCN, Bangalore, January 4–7, pp 389–392. ACM, New York. https://doi.org/10.1145/3288599.3295587

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Boon-Itt S, Skunkan Y et al (2020) Public perception of the COVID-19 pandemic on twitter: sentiment analysis and topic modeling study. JMIR Public Health Surveillance 6(4):21978. https://doi.org/10.2196/21978

Bucher H-J, Boy B, Christ K (2021) Audiovisuelle Wissenschaftskommunikation Auf YouTube: Eine Rezeptionsstudie zur Vermittlungsleistung Von Wissenschaftsvideos. Springer, Cham et al. https://doi.org/10.1007/978-3-658-35618-7

Chang J, Gerrish S, Wang C, Boyd-graber J, Blei D (2009) Reading tea leaves: How humans interpret topic models. In: Advances in neural information processing systems, vol 22. Curran Associates, Inc., Red Hook, New York. https://proceedings.neurips.cc/paper_files/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf

Chuang J, Manning C, Heer J (2012) Termite: visualization techniques for assessing textual topic models. https://doi.org/10.1145/2254556.2254572

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding

Doogan C, Buntine W (2021) Topic model or topic twaddle? re-evaluating semantic interpretability measures. In: Proceedings conference of the North American chapter of the association for computational linguistics: human language technologies, pp 3824–3848. ACL, Online. https://doi.org/10.18653/v1/2021.naacl-main.300

Dreisiebner S, März S, Mandl T (2022) Information behavior during the Covid-19 crisis in German-speaking countries. J Document 78(7):160–175. https://doi.org/10.1108/JD-12-2020-0217

Egger R, Yu J (2022) A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Front Sociol 7. https://doi.org/10.3389/fsoc.2022.886498

Egger R, Yu J (2022) A topic modeling comparison between LDA, NMF, Top2vec, and BERTopic to demystify Twitter posts. Front Sociol 7. https://doi.org/10.3389/fsoc.2022.886498

Grootendorst M (2022) BERTopic: neural topic modeling with a class-based TF-IDF procedure. https://arxiv.org/abs/2203.05794

Honnibal M, Montani I (2017) spaCy 2: Natural language understanding with Bloom embeddings. Convolutional neural networks and incremental parsing

Hu K, Wu H, Qi K, Yu J, Yang S, Yu T, Zheng J, Liu B (2017) A domain keyword analysis approach extending term frequency-keyword active index with Google Word2Vec model. Scientometrics, pp 1–38 https://doi.org/10.1007/s11192-017-2574-9

Jaki S (2021) This is simplified to the point of banality.: Social-Media-Kommentare zu Gestaltungsweisen von TV-Dokus. Journal für Medienlinguistik 4(1):54–87. https://doi.org/10.21248/jfml.2021.36

Jin X, Zhang S, Liu J (2018) Word semantic similarity calculation based on Word2vec, pp 12–16. https://doi.org/10.1109/ICCAIS.2018.8570612

Kalepalli Y, Tasneem S, Teja PDP, Manne S (2020) Effective Comparison of LDA with LSA for Topic Modelling. In: International conference on intelligent computing and control systems (ICICCS), pp 1245–1250. IEEE

Koufakou A, Scott J (2020) Lexicon-enhancement of embedding-based approaches towards the detection of abusive language. In: Proceedings of the second workshop on trolling, aggression and cyberbullying, pp 150–157. European Language Resources Association (ELRA), Marseille, France. https://aclanthology.org/2020.trac-1.24

Küçük D, Arıcı N (2022) Sentiment analysis and stance detection in turkish tweets about covid-19 vaccination. In: Handbook of research on opinion mining and text analytics on literary works and social media, pp 371–387. IGI Global, Hershey, PA, USA. https://doi.org/10.4018/978-1-7998-9594-7.ch015

Lahitani AR, Permanasari AE, Setiawan NA (2016) Cosine similarity to determine similarity measure: Study case in online essay assessment. In: 2016 4th International conference on cyber and IT service management, pp 1–6. https://doi.org/10.1109/CITSM.2016.7577578

Lima B (2023) Abrantes Baracho, R.M., Mandl, T.: Optimizing topic modelling for comments on social networks: Reactions to science communication on covid. WorldCist'23—11th world conference on information systems and technologies. Italy. April. Springer, Cham et al, pp 4–6

Mandl T (2015) Text mining. In: Encyclopedia of information science and technology, Third Edition, pp 1923–1930. IGI Global, Hershey, PA, USA. https://doi.org/10.4018/978-1-4666-5888-2.ch185

Mandl T, Jaki S, Mitera H, Schmidt F (2023) Interdisciplinary analysis of science communication on social media during the covid-19 crisis. Knowledge 3(1):97–112. https://doi.org/10.3390/knowledge3010008

Melo T, Figueiredo CM et al (2021) Comparing news articles and tweets about COVID-19 in Brazil: sentiment analysis and

topic modeling approach. JMIR Public Health Surveillance 7(2):24585
Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space
Montesi M (2021) Human information behavior during the COVID-19 health crisis: a literature review. Library Inf Sci Res 43(4):101122. https://doi.org/10.1016/j.lisr.2021.101122
Ng QX, Lim SR, Yau CE, Liew TM (2022) Examining the prevailing negative sentiments related to covid-19 vaccination: unsupervised deep learning of twitter posts over a 16 month period. Vaccines 10(9):1457. https://doi.org/10.3390/vaccines10091457
Oesper L, Merico D, Isserlin R, Bader GD (2011) Wordcloud: a cytoscape plugin to create a visual semantic summary of networks. Source Code Biol Med 6(1):7
Ramírez E, Brena R, Magatti D, Stella F (2012) Topic model validation. Neurocomputing 76:125–133. https://doi.org/10.1016/j.neucom.2011.04.032
Rehurek R, Sojka P (2011) Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2)
Reveilhac M, Morselli D (2022) Dictionary-based and machine learning classification approaches: a comparison for tonality and frame detection on Twitter data. Polit Res Exchange 4(1):2029217. https://doi.org/10.1080/2474736X.2022.2029217
Schnabel T, Labutov I, Mimno D, Joachims T (2015) Evaluation methods for unsupervised word embeddings. In: Proceedings conference on empirical methods in natural language processing, pp 298–307. ACL, Lisbon, Portugal. https://doi.org/10.18653/v1/D15-1036
Sievert C, Shirley K (2014) Ldavis: a method for visualizing and interpreting topics. https://doi.org/10.13140/2.1.1394.3043
Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151. https://doi.org/10.1126/science.aap9559
Wang B, Wang A, Chen F, Wang Y, Kuo C-CJ (2019) Evaluating word embedding models: methods and experimental results. APSIPA Trans Signal Inf Process 8(1). https://doi.org/10.1017/atsip.2019.12
Yin H, Song X, Yang S, Li J (2022) Sentiment analysis and topic modeling for COVID-19 vaccine discussions. World Wide Web 25(3):1067–1083. https://doi.org/10.1007/s11280-022-01029-y
Zulfiker MS, Kabir N, Biswas AA, Zulfiker S, Uddin MS (2022) Analyzing the public sentiment on covid-19 vaccination in social media: Bangladesh context. Array 15, 100204 https://doi.org/10.1016/j.array.2022.100204