Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Information Systems 109 (2022) 102054

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

Mining association rules from COVID-19 related twitter data to discover word patterns, topics and inferences

Paraskevas Koukaras, Christos Tjortjis *, Dimitrios Rousidis

The Data Mining & Analytics Research Group, School of Science and Technology, International Hellenic University, 14th km Thessaloniki – N. Moudania, 57001 Thermi, Greece

ARTICLE INFO

Article history:
Received 5 March 2021
Received in revised form 21 September 2021
Accepted 20 April 2022
Available online 25 April 2022
Recommended by Ioannis Katakis

Keywords:
Social media
Topic extraction
Association rule mining
Data mining
COVID-19

ABSTRACT

This work utilizes data from Twitter to mine association rules and extract knowledge about public attitudes regarding worldwide crises. It exploits the COVID-19 pandemic as a use case, and analyzes tweets gathered between February and August 2020. The proposed methodology comprises topic extraction and visualization techniques, such as WordClouds, to form clusters or themes of opinions. It then uses Association Rule Mining (ARM) to discover frequent wordsets and generate rules that infer to user attitudes. The goal is to utilize ARM as a postprocessing technique to enhance the output of any topic extraction method. Therefore, only strong wordsets are stored after discarding trivia ones. We also employ frequent wordset identification to reduce the number of extracted topics. Our findings showcase that 50 initially retrieved topics are narrowed down to just 4, when combining Latent Dirichlet Allocation with ARM. Our methodology facilitates producing more accurate and generalizable results, whilst exposing implications regarding social media user attitudes.

1. Introduction

People have been using Social Media (SM) in an extensive global scale, exchanging messages, posting opinions, news and more. During recent years SM public usage has greatly increased. There are multiple functionalities on offer, rendering SM one of the most popular online activities [1]. In 2020, over 3.6 billion people worldwide engaged with SM, with a predicted number of around 4.41 billion users in 2025 [2]. They became a great source of data for knowledge extraction.

The COVID-19 pandemic constitutes a worldwide health crisis, which became one of the hottest discussion topics. People generate vast amounts of online data regarding a variety of issues pertaining economic, social, political or health implications. At the same time, individuals, organizations, corporations and governments use SM. They act as a medium for exchanging information or monitoring opinions and attitudes about this crisis. The analysis of content derived from SM, such as Twitter, is a challenging task, since a large amount of data needs to be accumulated, summarized or aggregated. In general, SM platforms tend to generate text data that are sparse and noisy. Great endeavor is required to analyze them for knowledge extraction.

Therefore, nowadays more than ever, there is a need for methods and techniques to handle these huge volumes of data and generate opportunities for mitigating global crises like COVID-19. Governments, companies, organizations, and other involved parties need tools for understanding the topics of discussions in such events. These opportunities may involve better decision support for policy makers, improved and useful online information retrieval and more.

Data Mining (DM) techniques, such as Clustering, Classification and Association Rule Mining (ARM) are widely used for the extraction of knowledge from SM data in various domains, such as healthcare [3]. We often utilize such techniques for aggregating or summarizing information retrieved from online content. If grouped/modeled appropriately this content can generate topics or themes that effectively represent the essence of the data [4].

This paper experiments on COVID-19 Twitter data by introducing a novel methodology for identifying topics of discussions related with it. It performs a combination of Latent Dirichlet Allocation (LDA) [5], a common topic extraction technique, and then enhances its output with ARM. The methodology aims at preprocessing tweets for a specific period and then extracting knowledge related to public opinion.

The main contribution of this work is that it mitigates issues arising during topic extraction and it identifies topics precisely. Topic extraction techniques can be generic, often generating wordsets that do not infer to topics clearly. At the same time, quite often, words can appear in multiple extracted topics, leading to redundant data that are not ACID (Atomicity, Consistency, Isolation, Durability) compliant [6]. For example, LDA tends to appoint the same words within tweets to multiple topics. This

* Corresponding author.
E-mail addresses: p.koukaras@ihu.edu.gr (P. Koukaras), c.tjortjis@ihu.edu.gr (C. Tjortjis), d.rousidis@ihu.edu.gr (D. Rousidis).

https://doi.org/10.1016/j.is.2022.102054
0306-4379/

occurs since the selection of words per topic is based on the probability of each word belonging to that topic. Finding the same words along multiple topics means that these words have high probabilities in those topics. Moreover, LDA requires a fixed number of topics to be known ahead of time. We overcome this by performing topic coherence and topic stability analysis. We tackle the non-hierarchical nature of most topic extraction approaches by allowing the sharing of unique wordsets by creating a pool of words and retrieve the strongest of them. Finally, we mitigate issues regarding static topic extraction by predefining the period for retrieving COVID-19 related topics, but also by not considering time as an investigated feature.

The methodology consists of two main processes. First, topic extraction is performed using LDA. Then ARM takes place to extract the strongest wordset rules enhancing the overall topic extraction output. The resulting wordsets infer to discussion topics about the pandemic, whilst mitigating the issue of multiple word-to-topic assignments. The strongest rules can be visualized using graphs or WordClouds outlining in a more accurate and clear manner the frequently discussed topics. Our methodology can lead to the generation of fewer topics, with stronger word inference that represent public attitudes as expressed by Twitter posts. The results showcase that it is possible to reduce the extracted topics related with COVID-19 tweets from 50 down to just four containing stronger word inferences.

The remainder of the paper is structured as follows. Section 2 reviews examples of topic and opinion extraction using SM data, focusing on Twitter posts related with COVID-19. Section 3 analyzes our methodology, while Section 4 discusses results. Finally, Section 5 summarizes achievements, and presents future research directions.

2. Related work

This section presents research regarding topic and frequent wordset extraction utilizing SM data and recent attempts related to COVID-19.

In [7] authors highlight the primary role of Twitter to facilitate short text messages for a variety of purposes, whilst proposing a novel topic detection technique. This technique allows real-time retrieval of top emergent topics in communities. Text content is extracted from tweets and modeled according to a life circle introducing an aging theory. Identified terms are labeled as emerging, if they frequently occur in a specified time interval and have been rarely used in the past. The authors also used the Page Rank algorithm to rank the importance of the content based on its source. The study's findings are validated by a navigable graph. It connects emerging terms accompanied by keywords, under user specified time slots for various use cases.

Analyzing content from Twitter and attempting to summarize its information might become a quite challenging task. This process can be accomplished by extracting topical key phrases. A context-sensitive Page Rank method considers locality and utilizes a probabilistic scoring function to rank keywords. It also deals with concepts such as relevance and interestingness of key phrases before proceeding to ranking key phrases. The approach was validated by experimenting on a Twitter dataset, showing the efficacy of the topical key phrase extraction process [8].

Twitter offers microblogging services. It generates a great number of instant messages, creating opportunities for opinion summarization. Celebrities and brands are entities that generate a large volume of tweets. A proposal about an entity-centric and topic-based opinion summarization framework is presented in [9]. The methodology for summarizing topics based on extracted opinions comprises of (i) mined topics from hashtags, (ii) grouping hashtags on a weekly basis, (iii) similarity calculation among them and (iv) utilization of the Affinity Propagation algorithm to group hashtags into coherent topics. Then a dependent sentiment classification approach identifies the opinion for a specific target of tweets generating insights. The integration of topic, opinion, insights and other factors (e.g., language styles) forms an optimization framework for extracting opinion summaries.

SM pose a great source of user generated information context. That context exposes user interaction, streams of content, friendships and more. In Twitter, this user content can be extracted by exploiting conversation patterns and lists of user generated Twitter data. In such an approach, mined user context can generate user topics of interest. The validity of this attempt displays an 84% precision regarding the indication that topic information can be extracted from just the user context [10].

Since SM contain rich and abundant information, there is also the need for successful filtering and extraction of trending topics and events. A variety of methods exists for this process, exposing different qualitative results. A comparative research identified six topic detection methods, validated with three Twitter datasets regarding events. Variables, such as the nature of the event, the activity over time, sampling and pre-processing related data, affect each method. Standard Natural Language Processing (NLP) methods do well on very specific topics, while novel methods need to be employed for handling heterogeneous streams of concurrent events. A novel topic detection method based on topic co-occurrence and ranking, seems to perform well, considering the aforementioned conditions [11].

A study about criminal incidents exploits SM data to predict such events that previously relied only on historic crime records, geospatial data and demographics [12]. In SM there might be context that involves incidents of interest, inferring to possible upcoming events. This approach combines NLP of Twitter posts, dimensionality reduction with LDA and a linear prediction model. The evaluation of results is performed attempting to predict hit-and-run crimes, showcasing that it performs better than a baseline model for multiple days.

For effective text mining in real-time data from SM, the process of stemming may be valuable. New text mining challenges arise due to the great amount of text data produced by SM. Most of the text mining techniques apply to pre-defined datasets that are processed without applying limitations to computational complexity and execution times, while not paying attention to event triggers. A work that proposes a lightweight event detection method that uses wavelet signal analysis of hashtags is presented in [13]. LDA along with Gibbs Sampling become a part of a proposed strategy for detecting events while Continuous Wavelet Transformation identifies mention hypes for user specified hashtags. Results show that the proposed approach can summarize Twitter events in streaming environments.

Another novel technique for topic modeling utilizing Twitter data is presented in [14]. It creates groups of tweets occurring in the same conversation between two users. Therefore, a new scheme takes form, allowing tweets and their replies to be aggregated into a document. The users who post them are marked as co-authors. An experimental dataset for topic modeling occurs by utilizing LDA and Author–Topic Model (ATM) to create pools of tweets. A comparative analysis shows that the proposed approach outperforms others in the quality of formed clusters and document retrieval.

SM are being used extensively around the world and are a source of news, opinions and many more on a daily basis. Therefore, opportunities for an automated tool that identifies topics and user sentiment arise. A prototype tool that attempts to identify the pulse of Arabic users is introduced in [15]. Twitter data are used for extracting unigram words appearing more than 20 times in each corpus. Then, they are fed as features for grouping

tweets with bisecting k-means clustering. Results show that the quality of identified topics reaches 72.5%.

Twitter is commonly used for the dissemination of online events that happen in real-time. Frequent Pattern mining was utilized to detect topics in Twitter data. In that case topics comprise groups of words; yet the possibilities for utility pattern generation are omitted. Such a method that attempts to detect emerging topics utilizing Utility Pattern mining along with Frequent Pattern mining is proposed in [16]. Tweets are grouped based on time windowing and a utility of words is defined based on the growth rate and frequency. Then, postprocessing extracts topic patterns to be stored in a Topic-tree data structure. Experimental evaluation is conducted in three datasets, demonstrating better results (5% higher topic recall) and faster execution times compared to other topic detection techniques.

SM platforms such as Twitter produce sparse and noisy text that could be organized into an ontology containing multiple topics. The data can be processed in real-time generating new opportunities for an industrial application that deals with the topic modeling issue in a feasible and effective manner. Such a system providing functionalities like non-topical tweet detection, automatic labeled data acquisition, diagnostic and corrective learning and high precision topic inference is presented in [17]. Topic inference is achieved with a training algorithm for text classification. It uses a mechanism that associates text with external information sources with 93% precision and adequate topic coverage.

The COVID-19 pandemic severely impacts societies and people at the economic, psychological and health level. Governments, organizations, and individuals make use of SM attempting to mitigate this impact. There is a need to extract knowledge regarding content topics that emerge from SM platforms, to inform policy makers and health experts about public opinions and needs. A study utilized 2.8 million tweets identifying 12 topics that were clustered into four themes: (i) the origin of the virus, (ii) the source of the tweet, (iii) its economic and society impact, and (iv) mitigation of the risk infection options [18]. Ten topics had positive sentiment, whilst two had negative, related to deaths and racism. The implications suggest that SM should be utilized to communicate useful health information to the public. Worldwide health systems should focus on disease detection, monitoring and surveillance systems that take advantage of the information generated by the SM.

COVID-19 is a trending topic on SM and more specifically on Twitter. Governments and policy makers take measures attempting to contain the virus spread. Authors in [19] attempt to analyze and report on public reactions according to data extracted from Twitter. They utilized VOSviewer for extracting clusters of COVID-19 related tweet sentiment that form topics, labeled as public sentiments for (i) USA, (ii) Italy, Iran and a vaccine, (iii) doomsday and science credibility, (iv) India, (v) COVID-19's emergence, (vi) Philippines, and (vii) US Intelligence Report. In addition, the most frequent itemsets were synonyms of COVID-19 while the most frequent and confident association rules involved words related with testing, lockdown and China.

Lockdowns around the world were imposed against the pandemic. SM have been paramount for sharing information about this crisis. A mixed-methods analysis of tweets for the period of May 10 to May 24, 2020, was conducted utilizing the MAXQDA software along with the Twitter API for COVID-19 related data for New York. Content analysis unveiled primary topics from unstructured textual data, exposing six themes. These are surveillance, prevention, treatments, testing and cure, symptoms and transmission, fear and financial loss. Accessing public concerns in real-time during the pandemic exposes new fears regarding public health [20].

According to the literature, there is a variety of attempts for topic and opinion extraction from Twitter with various problem applications, themes and use cases (healthcare, celebrities, crime, event detection etc.). In addition, there is a wide mix of methods used. For example, Page Rank [7,8], Affinity Propagation [9], LDA and linear prediction model [12], LDA with Gibbs Sampling and Continuous Wavelet Transformation [13], LDA with Author–Topic Model [14], k-means clustering [15], Utility Pattern mining along with Frequent Pattern mining [16], VOSviewer [19] and MAXQDA software [20]. Moreover, the evaluation is performed in different datasets under different circumstances, rendering a comparative validation/analysis impossible. In general, topic extraction techniques generate wordsets that do not infer to topics in a clear manner. At the same time, words can often appear in multiple extracted topics. Topic extraction involving LDA frequently requires to predefine the number of topics, whilst enforcing a non-hierarchical analysis which does not allows data sharing among internal algorithmic iterations. Finally, most LDA related attempts address static topic extraction, i.e. topics do not evolve over time.

Our approach envisions to enhance the output of any topic extraction technique by performing ARM on its output. It tackles the issues of: (i) appointing the same words within tweets to multiple topics by replacing a probabilistic identification of strong words with a rule-based identification. (ii) requiring a predefined number of topics by performing topic coherence and topic stability analysis. (iii) non-hierarchical analysis by considering unique wordsets in a uniform manner (throughout the whole dataset) by creating a single pool of words resulting from multiple LDA executions, retrieving the strongest of them for generating topics (using ARM on the output of LDA). (iv) issues related with static topic extraction with no evolution of topics over time by predefining the period for retrieving topics related with COVID-19, but also not considering time as an investigated feature in line with our research design.

To the best of our knowledge, this approach is unique. For validating its results, we utilize LDA, a common topic extraction method, combined with ARM for identifying the strongest wordsets that form topics.

3. Research methodology

3.1. Summary

This work utilizes Twitter data retrieved over a period of 153 days, from 27th of February 2020 up to 28th of August 2020, in a worldwide scale. It aims to provide insights about the pandemic, exposing capabilities related with topic extraction from text data.

Our research methodology consists of four steps: (i) Preprocessing, (ii) Topic Extraction, (iii) Association Rule Mining (ARM) and (iv) Knowledge Extraction. Fig. 1 outlines the key components of this methodology, as well as the process flow for the methodology. Each of these steps are detailed in the following subsections. Knowledge extraction is elaborated separately as it culminates the results of this study.

3.2. Dataset

The dataset for this research was gathered using a crawler to retrieve tweets from Twitter's search functionality.[1] The search keywords included COVID-19 common synonyms such as "coronavirus", "covid", "covid-19" and "corona". The tweets under scrutiny date from 27/2/2020 until 28/8/2020, summing up to 2,146,243 unique tweets. The tweets have been filtered so that

[1] https://twitter.com/explore.

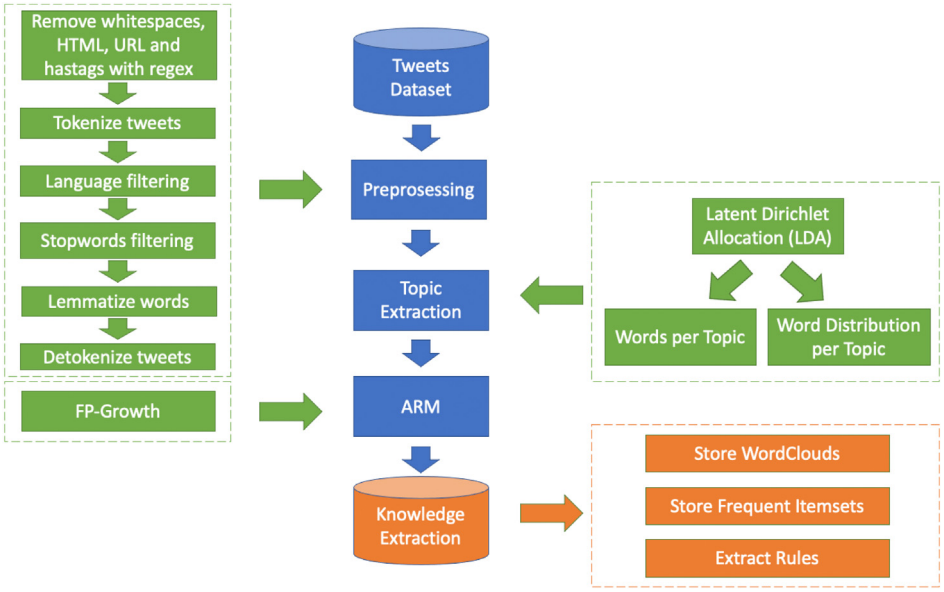

Fig. 1. Flowchart for our methodology.

they only contain English text. Although crawling bypasses some of Twitter API's[2] drawbacks, such as the max number of retrieved tweets, it generates other issues like the need for better text preprocessing.

The implemented method only crawled tweet text, since we did not intend to associate tweets with users, datetimes or tweet post counts. We focused on extracting themes or topics in a uniform manner, for the investigated period of 153 days.

3.3. Preprocessing

Data preprocessing is a step that typically involves data transformations prior to analysis. This work deals with text preprocessing in English, in order to prepare the data for topic extraction and ARM (Fig. 1). The employed sub-processes of preprocessing include incorporation of regex for removing whitespaces, HTML, URL and hashtag elements, Language text filtering (English), stopword removal, tokenization, and detokenization.

More specifically, we employ regular expressions to remove whitespaces, HTML, URL and hashtags with the re library (Regular expression operations,[3]). The tokenization of text is employed by utilizing the nltk.tokenize package.[4] Next, stopword removal is implemented utilizing in a sequential manner the build-in English lexicon versions from three libraries/packages/modules spacy.[5] genism,[6] and nltk.corpus.stopwords[7] The reason for doing so, is to increase the overall stopword lexicon without having to manually append words. To perform a morphological analysis of the words, we use lemmatization from nltk.stem.wordnet.[8] That way we extract the lemma of the words in most of the cases. Detokenization takes place by incorporating the Penn Treebank detokenization implementation from nltk.tokenize.treebank.[9] After preprocessing, the tweets take the form shown in Table 1 while the overall usable tweets were reduced from 2.146.243 to 2.062.864 (96,12% of the initial dataset).

3.4. Topic extraction

To perform topic extraction, we need to employ document (tweet) clustering. This process allows for the abstraction and analysis of lots of data. Once the tweets are clustered, we can allocate new tweets to existent clusters (topics) based on a text similarity metric.

LDA [5] is a topic modeling algorithm that stores and identifies topics and text distribution from a pool of existent text documents, such as tweets. Therefore, once tweet preprocessing concludes, LDA can extract topics. The LDA steps are:

1. Set a number k of topics to be identified.
2. Randomly place each tweet word in one temporal topic.
3. Iterate, processing all tweets and words computing (i) the probability the currently indexed tweet to be appointed to a specific topic, according to how many of the words of this tweet are already appointed to the same topic as the currently indexed word and (ii) the percentage of the tweets appointed to the same topic as the currently indexed word.

Steps 1–3 are executed n times, with n being a predefined value before the algorithm commences. Then, each tweet is appointed to a unique topic according to its most dominant words. For LDA parameter tuning we used the default parameters of scikit-learn.[10] For n, we used 10 iterations, the default parameter value.

The value for k is usually set by calculating topic coherence for a variety of topic numbers. HDP-LDA can be used instead of LDA, despite both being subject to misinterpretation [21]. LDA requires user specified k, while HDP-LDA operates with an unbounded number k, defined by data. For the purposes of this study, we created multiple LDA models with different k values, utilizing the preprocessed tweets dataset (Section 3.3). We employed Jaccard similarity [22] to perform a topic stability analysis [23], but we also performed topic coherence analysis [24]. We exploit this

2 https://developer.twitter.com.
3 https://docs.python.org/3/library/re.html.
4 https://www.nltk.org/_modules/nltk/tokenize.html.
5 https://spacy.io/.
6 https://pypi.org/project/gensim/.
7 https://www.nltk.org/_modules/nltk/corpus.html.
8 http://www.nltk.org/_modules/nltk/stem/wordnet.html#WordNetLemmatizer.
9 https://www.nltk.org/_modules/nltk/tokenize/treebank.html.
10 https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html.

Table 1
Examples of original and preprocessed tweets.

Original tweet	Preprocessed tweet
Thank you for taking the time to consume, digest, and distill this information for all of us. Just want to give you some positive reinforcement from a fan of Boomers and ultimately history and proven experience. #knowledge #coronavirus	Thank take time consume digest distill information want positive reinforcement fan boomers ultimately history prove experience
No I am more worried about the 1% fatality rate of the #coronavirus	Worry fatality rate
Which zombie apocalypse film does our government's response to #Coronavirus most closely resemble?	Zombie apocalypse film governments response closely resemble
# Coronavirus Vaccine 'at Least a Year' Away, Health Official Says #USA #Republicans President Donald #Trump told reporters we were "very close" to a coronavirus vaccine, causing confusion as to the state of a vaccine. https://www.newsweek.com/anthony-fauci-coronavirus-vaccine-year-away-public-availability-1489214	Vaccine year away health official say president donald tell reporters close coronavirus vaccine cause confusion state vaccine
So let me get this straight, world governments can't stop #coronavirus from spreading but if we pay more tax we can change the planets temperature #ClimateChangeHoax	Let straight world governments stop spread pay tax change planets temperature
The #coronavirus is not a time for politics. #COVID19	Time politics
"Radiologists understanding of clinical and chest CT imaging features of coronavirus disease 2019 (COVID-19) will help to detect the infection early and assess the disease course. https://pubs.rsna.org/doi/10.1148/radiol.2020200490 #Corona #coronavirus #CoronaOutbreak #COVID19 #COVID-19	Radiologists understand clinical chest ct image feature coronavirus disease covid help detect infection early assess disease course
Coronovirus could be answer we've been looking for to tame medicare costs and prevent the collapse of #SocialSecurity #coronavirususa #coronavirus #CoronavirusOutbreak #CoronaVirusUpdates #boomers pic.twitter.com/EX5sQ6XXX7	Coronovirus answer weve look tame medicare cost prevent collapse
'It's post-apocalyptic': how #coronavirus has altered day-to-day life https://www.theguardian.com/world/2020/feb/21/post-apocalyptic-how-coronavirus-has-altered-day-to-day-life-wuhan-north-england?CMP=share_btn_tw	Postapocalyptic alter daytoday life

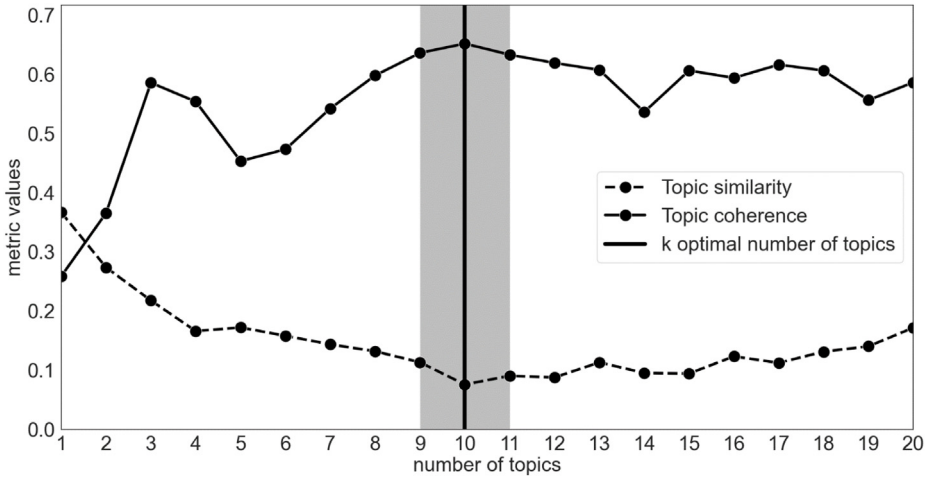

Fig. 2. Similarity, coherence and optimal number of topics.

combinatorial approach and present results in the form of two metrics. This process acts as an indicator for identifying semantic similarity across high scoring words, within all possible k number of topics. According to Fig. 2, the optimal number of topics k = 10 is derived from the maximum difference between coherence and similarity.

To sum up, first LDA is incorporated to our methodology to extract topics. This allows a high-level analysis that narrows down public attitudes during the COVID-19 epidemic period under scrutiny (27/2/2020 up to 28/8/2020). That way, we extract strong keywords with a commonly used method (LDA) representing public attitudes/opinion. Next, we perform ARM on these strong keywords attempting to enhance the topic extraction process while mitigate issues of the LDA approach (as reported in Section 1).

3.5. Association rule mining

ARM is a DM technique that allows the discovery of relationships between variables in databases. In general, ARM refers to items and itemsets (sets of items) and associations among them. In this paper, we customize and employ ARM notation and terminology to fit the SM domain. We substitute transactions with words per topic, items with words and itemsets with wordsets (i.e. sets of words). Thus, we attempt to find the strongest word rules within tweets by using metrics such as support, confidence, lift and leverage [25]. There is a variety of ARM algorithms to choose from, such as Apriori [26] or FP-Growth [27] that require to manually set the minimum support levels for extracting frequent wordsets. Other implementations, such as ARMICA [28], utilize techniques like the heuristic Imperialism Competitive Algorithm (ICA) to extract frequent wordsets without the need to specify support levels.

For the purposes of this paper, we utilize four measures for extracting interesting word rules from tweets, minimum support and confidence, as well as lift and leverage.

Let $I = \{i_1, i_2, i_3 \ldots, i_n\}$ be a set of n binary attributes called words. Let $P = \{t_1, t_2, t_3 \ldots, t_n\}$ be a set of transactions called the pool of tweets. Each transaction in P has a unique id and within it resides a subset of the words in I. A generated rule is

a suggestion of the form X⇒Y where X, Y⊆I and X ∩ Y = ∅. In that case the wordset X is the antecedent or left-hand-side (LHS) and Y the consequent, or right-hand-side (RHS) of the rule [29].

For selecting the most interesting rules from a set of rules for a pool of tweets, we must use thresholds regarding significance and interestingness. The metrics to set these thresholds are minimum support and minimum confidence, respectively.

Support $Supp(X)$ of a wordset X is the proportion of the tweets within a dataset that contains that set of words [30].

Confidence is an estimate of the probability of identifying words in the RHS of a rule within the tweet given that these words also satisfy the LHS. Confidence is essentially a metric that defines whether a rule is true and its frequency. It is given by the following formula [30]:

$$Conf\ (\mathrm{X} \Rightarrow \mathrm{Y}) = \frac{Supp(\mathrm{X} \cup \mathrm{Y})}{Supp(X)} \tag{1}$$

Lift calculates the degree of communality of antecedent and consequent wordsets that appear concurrently in case they are statistically independent. Lift values range between $[0, \infty]$. When lift is equal to 1, the wordsets (antecedent and consequent) are not related. A value between (0, 1) suggests a negative association, while values ranging between $[1, \infty]$ indicate even greater association as the value increases [31].

$$Lift(\mathrm{X} \Rightarrow \mathrm{Y}) = \frac{Conf\ (\mathrm{X} \Rightarrow \mathrm{Y})}{Supp(Y)} \tag{2}$$

Leverage calculates how different is the probability of a rule compared to the probability in case the antecedent and consequent wordsets are statistically independent. Its value ranges between $[-1, 1]$. When the value is equal to 0 this indicates that the wordsets are independent. The greater positive value of leverage suggests an even stronger positive relation, while more negative values indicate stronger negative relations of the wordsets (antecedent and consequent) [25].

$$Leverage\ (\mathrm{X} \Rightarrow \mathrm{Y}) = Supp\ (\mathrm{X} \Rightarrow \mathrm{Y}) - Supp\ (\mathrm{X}) * Supp(Y) \tag{3}$$

For experimenting with ARM on COVID-19 related tweets we used the FP-Growth algorithm. It is an exhaustive ARM algorithm able to produce the same results as Apriori, yet it is much faster [32]. That is a positive characteristic, since our experimentation involves thousands of unique words inside millions of tweets.

FP-Growth finds how many times words appear in the dataset of tweets and places them to a header table. A FP-tree structure is constructed through the insertion of these instances. The words in each of the instances get sorted in a descending order, based on the frequency of appearance in the dataset, enabling a faster tree traversal. At this point a threshold for traversal is set and all words that do not match the required conditions are discarded. That way, large wordsets can be constructed faster; by processing the compressed dataset version in a recursive manner, without creating candidate words and validating them throughout the full dataset. The recursive process concludes finding the longest sets of words pertaining the minimum coverage and the generation of association rules starts [27].

4. Results

This section presents results related to retrieved WordClouds, Frequent Wordsets and Association Rules. As discussed in Sections 3.3 and 3.4, we used LDA to extract topics from the database of preprocessed 2,062,864 tweets. The goal was to form clusters of topics that represent public feelings, opinions, attitudes or discover inferences regarding them. For that reason, LDA was performed several times. Table 2 shows results from these simulations.

Table 2
Simulations summary, 10 topics and extracted words per execution.

Simulation	min_threshold	number_of_extracted_words
1	1000	2713
2	2000	1597
3	3000	1153
4	4000	886
5	5000	713

We set a minimum threshold ensuring that if words appear in less than 1000, 2000, 3000, 4000 and 5000 tweets respectively, they are ignored. We empirically set these thresholds to range around 0.05% and 0.25% of the overall sum of tweets (2,062,864). As a result, words that seem to be too common to have a strong meaning for the topics are discarded. We report on the number of extracted words per execution. That way, we were able to observe how the number of extracted words decreases as the minimum threshold increases. In addition, we performed an analysis to set the number of extracted topics to be 10 as explained in Section 3.4. Next, we retrieved the 20 most common words for each topic, along with their weights (number of appearances) to form WordClouds.

We do that in order to rank and visualize the most frequently appearing words per generated topic.

Table 3 indicatively shows the top 10 words from LDA simulation_no1. The complete ranking list per LDA simulation along with the weights per word (number of appearances) can be found in Appendix A.

4.1. WordClouds

It is quite common to use WordClouds to visualize the strongest words per topic. For example, Fig. 3 depicts Topic#3 results from simulation#1.

Similar figures are employed to visualize the words for each topic, such as Topic#4 (Fig. 4).

For enhancing result visualization, we used a graph for each LDA simulation instead of just using WordClouds for depicting the topics retrieved, (Fig. 5). Each topic is depicted as a hub linked with nodes, corresponding to words belonging to that topic, generating a network of words and topics. The size of each node depends on the word's weight, i.e. its number of appearances in the dataset. The size of each hub results from the average number of appearances of its words. A graph visualization for each simulation is presented in Appendix B.

We performed five simulations attempting to generate topics or clusters that contain the stronger words retrieved from the tweets dataset. Yet, it is unclear what the theme of each topics is or whether some topics infer to the same theme.

One way to retrieve the theme for each topic is to set a threshold for appearances beyond which words can be considered influential for setting the theme of the topic. For example, according to Table 4 Topic#3 has a sum of 431,579 word appearances, while Topic#4 has a sum of 547,267 word appearances. If we arbitrary set a min threshold of 10% for considering a word important, this would mean that any words that have less appearances than 43,158 should be omitted. Respectively, for Topic#4 this number would be 54,727 appearances.

In that case, the extracted themes for each topic should be, Topic#3 (home, stay, work) and Topic#4 (null). If we lower the threshold to 5%, we may include many words that might introduce noise to the theme inference. At 5% we have the following outcome, Topic#3 (home, stay, work, safe) and Topic#4 (like, time, know, go, social, look, think, watch, good). Therefore, we need a mechanism to search and decide which words are considered more important amongst all the retrieved topics, and clarify

Table 3
Ten most common words per topic for simulation#1.

Rank	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
1	covid	coronavirus	pandemic	home	like	case	need	people	virus	close
2	hand	outbreak	impact	stay	time	new	help	die	mask	school
3	corona	read	covid	work	know	test	fight	days	trump	service
4	wash	pandemic	business	safe	go	deaths	people	infect	spread	measure
5	vaccine	say	market	time	social	report	health	patients	china	order
6	ms	covid	new	test	look	total	world	symptoms	stop	public
7	dr	question	company	get	think	number	crisis	kill	people	lockdown
8	clean	ill	crisis	family	watch	positive	pandemic	covid	dont	spread
9	virus	amid	help	quarantine	good	confirm	time	virus	news	open
10	drug	uk	team	individuals	day	coronavirus	support	disease	corona	pay

Fig. 3. LDA simulation_no1 Topic#3 WordCloud.

Fig. 4. LDA simulation_no1 Topic#4 WordCloud.

this selection based on the occurrences within the dataset. Also, we observe that words such as "people" or "get" may appear in multiple topics. To that end, we perform ARM on the set of strong words per topic. We attempt to narrow these down possibly to single words, and identify, in a more precise manner, what is the public attitude regarding the pandemic during the investigated period.

4.2. Frequent wordsets

As a first step, we append to a list all extracted words appointed to topics resulting from all simulations. That way we merge all simulation results into a dataset that contains the strongest words that can be exploited from the tweets dataset. The top-30 most frequent words in this dataset are shown in Fig. 6. The top five words in a descending order are "people", "coronavirus", "covid", "need" and "pandemic".

For identifying frequent wordsets we utilized FP-Growth (Section 3.5). We experimented with a variety of values for the minimum support level in order to extract frequent wordsets.

According to Table 5 there is only one frequent wordset when setting minimum support to 50%. We dropped it down to 3% and stopped, since the identified wordsets with such low support levels become too many. Our aim is to identify the most frequent wordsets with the stronger possible associations.

4.3. Association rules

We utilized lift and leverage, as described in Section 3.5, to extract the strongest rules. These measures show how the probability of a rule to hold can relate to the expected probability, when the antecedent and consequent wordsets are independent from each other. The difference is that lift can compute the strongest wordset associations for wordsets with lower support (less frequent wordsets), whilst leverage prioritizes wordsets with higher

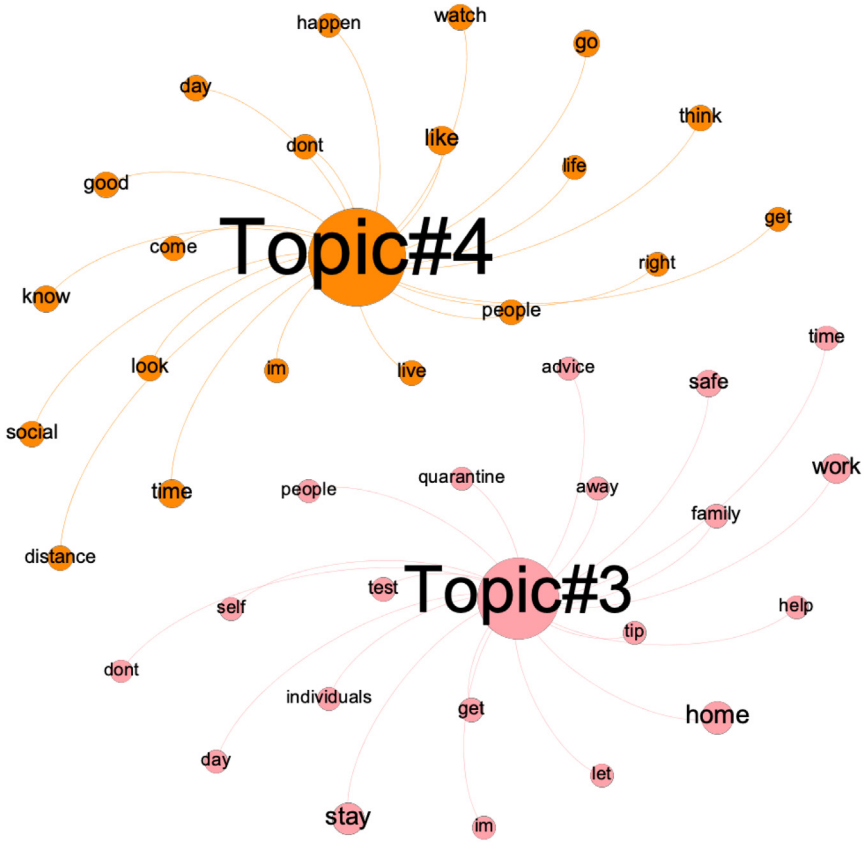

Fig. 5. Graph visualization of simulation_no1, Topic#3 and Topic#4.

Table 4
Simulation#1 Topic#3 & Topic#4 word ranking.

Rank	Topic 3 words	Topic 3 word appearances	Topic 4 words	Topic 4 word appearances
1	home	78 711	like	53 379
2	stay	70 744	time	44 102
3	work	56 798	know	34 043
4	safe	36 521	go	33 540
5	time	20 496	social	31 223
6	test	16 538	look	31 207
7	get	14 618	think	30 573
8	family	13 733	watch	30 250
9	quarantine	11 941	good	29 796
10	individuals	11 262	day	26 768
11	people	11 107	people	26 198
12	let	11 001	live	24 323
13	help	10 832	dont	22 888
14	im	10 715	distance	22 381
15	day	10 506	come	20 529
16	advice	10 039	life	19 270
17	self	9910	im	17 741
18	dont	8797	happen	17 246
19	away	8742	right	16 013
20	tip	8567	get	15 797

support levels. As we want to extract the strongest rules within the dataset, we focused on ranking the strongest rules, based on high leverage values, shown in Table 6.

We grouped words into topics by combining results from the top-50 most common words and association rules with the strongest support, confidence, lift and leverage. We identified strong rules by filtering out rules with leverage less than 9%, as they tend to have the same support, confidence, and lift, rendering the rules identical. The strongest 138 association rules can be found in Appendix C.

We grouped the strongest wordsets observing these rules and used graphs to visualize the final topics extracted from the dataset. This process forms four topics, depicted in Fig. 7. Node size for topics signifies its weight or word count of appearances, while node size for words signifies its frequency.

Topic#0 infers that people (“people”) discuss personal opinions (“im”, “think”, “like”, “know”), negative attitude (“dont”) and inferenced motivations such as “go”, or “get”. Topic#1 infers that users post tweets using COVID-19 synonyms, such as “corona”, “coronavirus”, “covid”, “virus” to comment about updates (“update”) on new (“new”) cases (“case”) or death(s). It should be noted that COVID-19 hashtags were filtered during data pre-processing. In addition, according to leverage values of rules associated with “death” and “deaths”, it is more likely

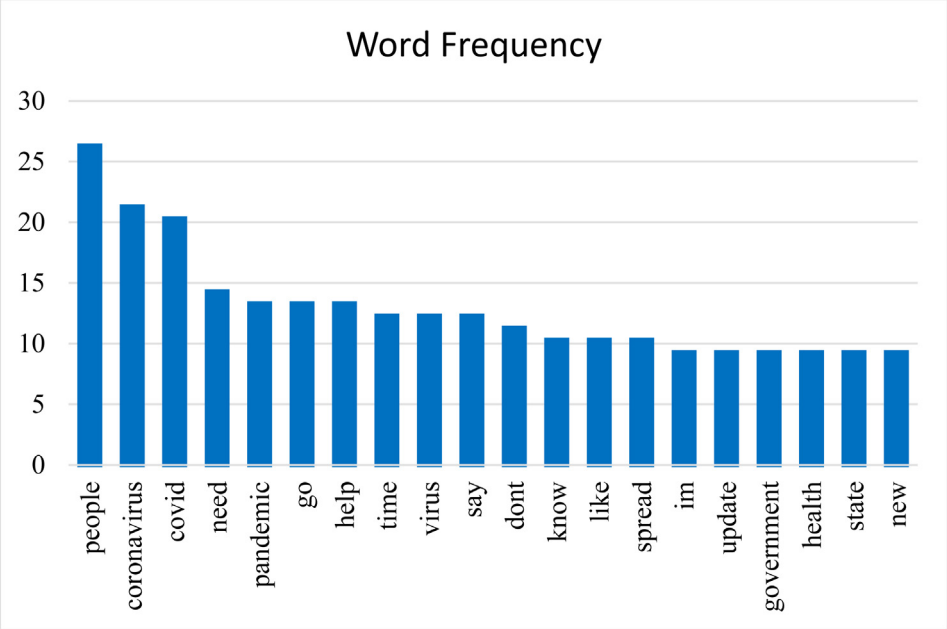

Fig. 6. Top 20 frequent words.

Table 5
Number of frequent wordsets with different support levels.

Min_support level	Frequent wordsets
50%	1
40%	3
30%	4
20%	19
9%–10%	16 622
7%–8%	33 480
5%–6%	135 696
3%–4%	570 924

Table 6
Number of association rules according to leverage.

Min_leverage level	Rules
10%	58
9.4–9.5%	102
9.3%	114
9.2%	130
9.1%	138
9%	4,598,926

that "deaths" appear instead of "death". If stemming was applied during preprocessing "death" and "deaths" would stand as the same lemma: "death". We chose not to apply stemming, as we are interested in this level of detailed information, for example, if a tweet refers to a single death or its plural form; yet we applied lemmatization. Topic#2 has a more arbitrary inference, since it contains the words "need", "look" and "time". It might be the case that users discuss the time needed for developing a cure. Yet, this cannot be proven by this analysis since these associations were not identified within the results.

Topic#3 has a rather clear inference containing the words "pandemic" and "crisis". Therefore, people label or mention in posts COVID-19 as a pandemic and probably comment on the numerous drawbacks of a crisis.

It is also evident that tweets that belong to Topic#0 and Topic#1 appear much more often than Topic#2 and Topic#3, as indicated by the size of their nodes.

Our analysis focuses on Twitter data related with COVID-19 retrieved between February and August of 2020. We first performed data preprocessing and then applied a common topic extraction technique (LDA) to retrieve the themes of discussions regarding the pandemic. Then, we attempted to improve topic extraction by narrowing down the number of top retrieved topics utilizing ARM. We aimed at discovering the greatest possible values, while being driven by data observation. These values are related with the interpretation of ARM measures, each used for a different purpose to identify the strongest rules between wordsets inside topics. The experimentation was conducted with various values for these measures as presented in this section. The results showcase that the task of topic extraction can be enhanced and further generalized by combining LDA with ARM leading to less yet more representative topics.

5. Conclusions

This paper attempts to introduce improvements in topic extraction from SM data. It showcases a methodology that narrows down the selection of wordsets to be included in extracted topics. This is achieved by utilizing a topic extraction method (LDA) and then ARM to identify word frequency appearances in these topics.

We analyzed 2,146,243 unique tweets for a period of 153 days, from 27/2/2020 up to 28/8/2020, focusing on discussions during the pandemic. Thus, a public dataset for a similar task is not available. We performed data preprocessing and topic extraction with various simulation parameters. For topic extraction, we used the LDA method. The output of LDA simulations displayed the existence of discrepancies regarding topic extraction. For example, the same strong words can appear in multiple topics at the same time or extracted topics may contain many trivia words that make topic theme inference more difficult (see Fig. 8).

To address these issues, we performed ARM to identify new topics that suggest SM user attitudes in a more precise and distinct manner. For this purpose, we generated wordset rules utilizing common ARM measures, such as support, confidence, lift and leverage. Out of the 50 topics retrieved by the LDA topic extraction method, we narrowed these down to four topics by removing trivia wordsets, while showcasing few and more representative strong wordsets that infer each topic theme. We

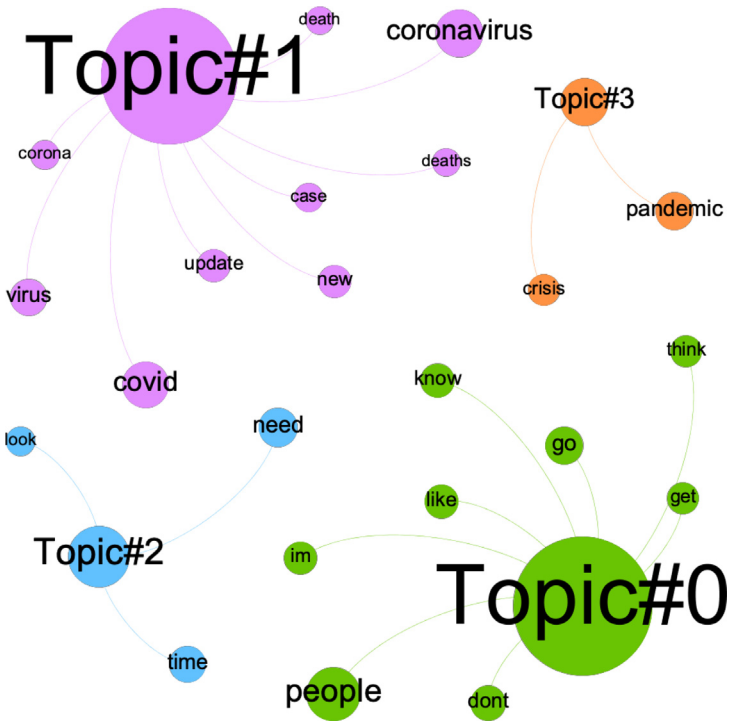

Fig. 7. Final Topic Extraction resulting from ARM rules.

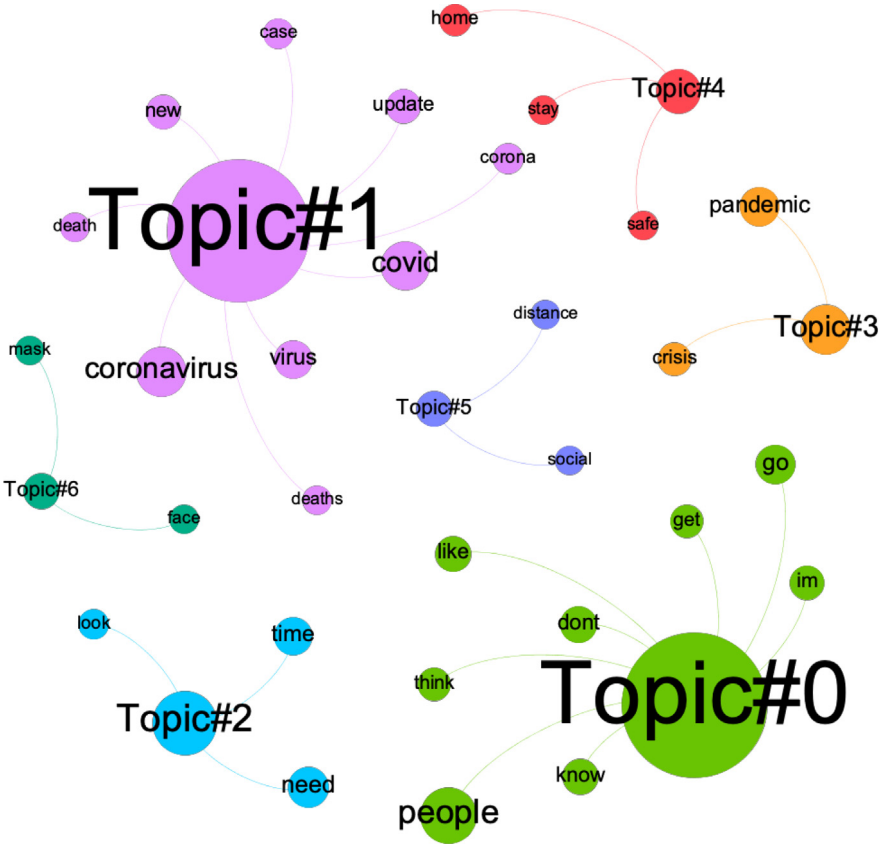

Fig. 8. Final Topic Extraction resulting from ARM rules with relaxed metrics' filtering.

conclude that the utilization of ARM as an extra filter for the results of common topic extraction techniques, such as LDA, can aid producing more representative and accurate results regarding SM user attitudes.

5.1. Implications

The proposed topic extraction methodology enhances the extraction of insights regarding SM user opinions, attitudes, and

Table 7
20 most common words per topic for LDA simulation_no = 1.

Rank	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights	Topic 5 words	Topic 5 weights	Topic 6 words	Topic 6 weights	Topic 7 words	Topic 7 weights	Topic 8 words	Topic 8 weights	Topic 9 words	Topic 9 weights
1	covid	24 884	coronavirus	97 834	pandemic	29 520	home	78 711	like	53 379	case	175 921	need	82 681	people	57 381	virus	56 481	close	35 222
2	hand	24 862	outbreak	23 294	impact	20 776	stay	70 744	time	44 102	new	97 337	help	48 615	die	42 803	mask	37 971	school	25 535
3	corona	15 035	read	17 661	covid	20 384	work	56 798	know	34 043	test	74 308	fight	42 479	days	30 492	trump	32 767	service	19 689
4	wash	14 589	pandemic	17 564	business	19 805	safe	36 521	go	33 540	deaths	66 692	people	37 625	infect	21 544	spread	28 767	measure	18 411
5	vaccine	14 339	say	15 758	market	17 994	time	20 496	social	31 223	report	57 204	health	28 402	patients	16 445	china	26 661	order	18 382
6	ms	13 098	covid	15 148	new	16 528	test	16 538	look	31 207	total	48 146	world	25 334	symptoms	15 221	stop	24 581	public	17 740
7	dr	12 896	question	14 807	company	14 765	get	14 618	think	30 573	number	43 930	crisis	22 265	kill	15 074	people	22 499	lockdown	17 456
8	clean	9194	ill	14 584	crisis	13 455	family	13 733	watch	30 250	positive	41 477	pandemic	19 457	covid	13 565	dont	22 174	spread	16 486
9	virus	8490	amid	14 224	help	13 174	quarantine	11 941	good	29 796	confirm	41 127	time	18 425	virus	12 572	news	21 813	open	16 004
10	drug	8472	uk	13 445	team	13 073	individuals	11 262	day	26 768	coronavirus	38 979	support	18 286	disease	12 537	corona	18 651	pay	15 023
11	est	6423	minister	12 343	support	12 964	people	11 107	people	26 198	update	35 087	medical	18 008	say	11 888	check	16 649	people	14 286
12	pandemia	6368	news	12 199	learn	12 264	let	11 001	live	24 323	death	32 060	thank	17 460	hand	11 739	say	15 581	government	11 818
13	go	5521	latest	12 185	global	12 144	help	10 832	dont	22 888	covid	27 159	care	16 722	person	11 197	wear	15 523	place	10 877
14	unavailable	5362	travel	11 788	economic	11 671	im	10 715	distance	22 381	state	24 057	workers	15 606	risk	11 140	chinese	15 189	food	10 729
15	youth	5134	countries	10 952	die	11 466	day	10 506	come	20 529	recover	20 658	money	15 234	care	11 055	think	14 114	state	10 234
16	earn	5118	sign	10 768	response	11 267	advice	10 039	life	19 270	rise	19 168	government	14 289	spread	11 045	f*ck	13 611	shop	9481
17	cure	5053	update	10 526	pm	11 238	self	9910	im	17 741	italy	17 319	country	12 959	doctor	10 064	go	13 426	store	8959
18	ecoins	4985	flight	10 424	businesses	10 757	dont	8797	happen	17 246	toll	16 794	doctor	12 902	infection	9783	face	12 967	shut	8889
19	gone	4890	answer	9525	work	10 700	away	8742	right	16 013	india	16 669	paper	11 132	cough	9066	know	12 641	pour	8886
20	hate	4812	check	7983	plan	10 549	tip	8567	get	15 797	health	15 437	protect	10 485	second	8711	buy	12 373	students	8880

Table 8
20 most common words per topic for LDA simulation_no = 2.

Rank	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights	Topic 5 words	Topic 5 weights	Topic 6 words	Topic 6 weights	Topic 7 words	Topic 7 weights	Topic 8 words	Topic 8 weights	Topic 9 words	Topic 9 weights
1	die	57 325	good	38 502	people	52 825	coronavirus	79 194	say	34 340	time	35 665	help	39 187	home	78 470	covid	21 780	case	175 922
2	people	38 207	lockdown	35 033	trump	40 886	china	42 523	people	32 111	come	28 982	crisis	31 208	stay	70 738	ill	16 821	new	105 304
3	im	17 694	close	34 359	dont	40 688	news	39 032	health	30 884	need	28 423	support	30 851	work	43 022	ms	13 098	test	92 928
4	man	17 168	social	33 502	think	36 477	world	37 090	live	25 328	thank	28 117	free	24 199	mask	38 539	cure	12 093	deaths	66 655
5	infect	14 478	time	30 177	know	34 130	pandemic	35 550	medical	23 043	let	27 850	need	24 033	safe	38 221	coronavirus	9839	report	52 008
6	year	14 433	school	25 376	like	31 250	outbreak	33 467	doctor	20 193	like	27 339	pandemic	22 122	face	24 634	est	9836	total	47 841
7	house	11 887	distance	23 947	want	22 655	spread	26 460	patients	18 678	help	21 934	business	20 117	corona	21 836	past	9642	number	43 842
8	paper	11 132	go	23 560	go	22 451	read	25 658	pm	17 420	right	21 159	impact	19 398	spread	21 191	non	9015	positive	42 096
9	toilet	9984	day	22 481	question	19 312	virus	22 514	minister	17 261	fight	20 831	covid	17 363	hand	20 726	pour	8886	confirm	41 127
10	old	9595	take	18 858	say	19 222	covid	19 824	save	17 249	world	20 768	service	17 259	protect	20 374	check	7983	coronavirus	39 431
11	wait	9462	weeks	17 345	hand	19 094	global	19 527	care	17 009	love	20 743	new	14 590	virus	17 436	day	7266	update	37 001
12	vote	8552	people	16 470	watch	18 367	market	19 439	spread	16 101	look	19 429	businesses	14 176	wear	15 727	plus	6781	death	33 540
13	self	8388	days	16 114	kill	17 747	countries	15 223	fight	16 030	know	19 284	fund	13 612	people	15 155	pandemia	6369	covid	27 211
14	men	8312	open	15 536	get	17 629	chinese	14 796	virus	15 391	im	18 944	emergency	13 590	dont	14 167	go	5959	state	20 900
15	white	7378	years	13 369	real	16 846	amid	14 200	covid	15 322	share	18 498	company	13 002	follow	13 650	place	5815	recover	20 705
16	additions	7024	week	11 459	tell	15 314	trump	13 476	risk	15 152	get	18 237	offer	12 959	wash	12 921	im	5815	rise	17 366
17	right	7017	today	11 287	make	15 092	warn	13 259	government	14 775	hope	18 015	economy	12 853	need	12 370	sure	5375	march	16 473
18	flu	6683	morning	11 058	try	14 817	vaccine	12 924	india	14 164	feel	16 502	coronavirus	12 663	buy	11 217	others	5342	toll	15 924
19	f*ck	6534	stop	10 684	need	14 639	say	12 586	disease	12 853	great	14 693	people	12 584	help	10 342	inside	5323	rate	15 317
20	virus	6526	start	10 543	ask	14 515	fear	11 875	treat	10 645	video	13 068	response	12 541	avoid	9758	youth	5134	record	13 576

Table 9
20 most common words per topic for LDA simulation_no = 3.

Rank	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights	Topic 5 words	Topic 5 weights	Topic 6 words	Topic 6 weights	Topic 7 words	Topic 7 weights	Topic 8 words	Topic 8 weights	Topic 9 words	Topic 9 weights
1	people	78 951	stay	70 688	spread	56 075	test	92 899	close	28 030	world	43 763	virus	56 842	time	78 946	work	78 334	case	175 922
2	get	48 693	home	55 192	health	44 307	say	32 755	support	27 973	china	32 739	news	42 270	need	57 245	coronavirus	40 953	new	100 676
3	like	36 208	hand	39 831	social	33 625	positive	28 060	school	25 478	new	19 543	trump	42 201	help	49 637	die	23 139	deaths	66 693
4	go	34 594	mask	38 540	public	28 807	pandemic	23 738	open	19 067	lockdown	16 195	corona	30 625	people	35 385	covid	20 147	report	57 549
5	dont	31 795	safe	37 751	distance	25 407	covid	22 081	covid	18 390	uk	14 286	watch	27 106	look	27 713	home	15 448	total	50 897
6	im	27 834	face	25 968	measure	21 049	medical	19 930	business	17 571	people	14 117	live	23 990	like	24 544	ms	13 098	number	44 399
7	die	27 074	wash	19 902	people	19 665	coronavirus	19 383	fight	17 198	hit	13 550	know	21 016	thank	24 008	company	11 605	coronavirus	41 893
8	know	25 957	buy	16 169	travel	18 803	patients	18 972	ill	16 821	government	13 542	read	20 736	share	23 966	force	9976	confirm	41 127
9	think	22 968	wear	15 727	minister	17 723	workers	17 620	pm	15 975	time	13 493	market	19 619	love	19 597	employees	9750	death	36 727
10	kill	21 765	dont	12 816	outbreak	17 704	need	17 520	service	14 953	come	13 407	media	16 693	think	18 292	self	9382	update	35 474
11	quarantine	20 173	protect	10 874	lockdown	17 117	dr	16 423	online	14 600	country	13 320	president	14 301	hope	17 513	person	8568	covid	33 269
12	let	19 823	use	10 215	risk	16 147	question	16 051	latest	14 512	end	12 898	good	12 699	know	17 403	men	8312	recover	20 719
13	right	17 413	touch	10 171	say	15 324	care	15 586	businesses	13 719	pandemic	12 712	video	12 449	great	16 774	cure	7450	rise	18 755
14	way	16 673	people	9540	coronavirus	15 117	response	15 147	update	13 095	coronavirus	11 510	lie	10 707	life	16 676	isolate	7161	toll	16 795
15	live	16 337	clean	9210	government	13 943	crisis	14 558	join	12 910	paper	11 125	pandemic	10 642	feel	16 027	additions	7024	italy	15 762
16	save	15 997	panic	9162	amid	13 929	trump	12 636	coronavirus	12 760	war	10 927	like	10 524	good	15 546	pandemia	6369	data	15 053
17	happen	15 974	avoid	8638	cancel	13 643	health	12 372	free	12 675	state	10 852	coronavirus	10 502	day	15 418	im	6118	rate	14 997
18	thing	15 904	follow	8492	take	13 299	state	12 064	help	12 391	city	10 589	impact	10 101	learn	14 158	meet	5447	state	14 058
19	want	14 963	water	8392	prevent	11 906	result	11 691	check	12 289	toilet	9957	chinese	9911	things	13 780	unavailable	5356	positive	14 040
20	flu	13 972	shop	8311	state	11 698	ask	10 844	march	11 269	crisis	9187	want	9785	important	13 599	others	5342	india	13 839

Table 10
20 most common words per topic for LDA simulation_no = 4.

Rank	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights	Topic 5 words	Topic 5 weights	Topic 6 words	Topic 6 weights	Topic 7 words	Topic 7 weights	Topic 8 words	Topic 8 weights	Topic 9 words	Topic 9 weights
1	time	77 313	help	50 796	world	56 633	home	78 063	question	19 783	test	100 147	coronavirus	81 292	dont	67 133	covid	38 634	case	175 905
2	like	41 220	work	49 432	trump	54 683	people	74 285	cancel	18 208	positive	37 004	spread	55 705	people	40 222	health	32 913	new	103 876
3	look	30 423	fight	45 783	know	39 104	stay	70 741	get	17 833	state	34 796	lockdown	34 045	say	34 849	read	26 924	deaths	66 693
4	day	21 163	hand	39 343	right	32 310	die	55 660	week	16 150	say	28 178	social	33 630	im	32 842	risk	24 755	report	55 675
5	come	21 049	need	33 602	china	31 424	safe	34 103	days	14 224	pm	25 214	distance	25 897	think	25 155	check	22 438	total	50 897
6	good	20 992	support	30 249	live	29 162	virus	22 782	ask	13 715	india	19 797	outbreak	20 404	know	25 116	doctor	21 419	number	44 066
7	best	20 268	thank	29 924	watch	25 672	work	19 088	ms	13 098	march	19 642	measure	19 722	want	19 669	latest	20 417	coronavirus	41 656
8	market	19 676	mask	27 261	go	25 616	infect	18 986	answer	11 005	crisis	18 108	amid	18 920	like	18 818	information	19 951	confirm	41 127
9	today	18 472	face	26 288	people	22 209	quarantine	18 961	months	9594	minister	17 683	public	17 364	love	18 577	update	18 680	death	37 177
10	people	18 353	pandemic	25 565	president	19 348	corona	17 665	air	9170	close	16 707	follow	15 771	year	18 382	patients	18 304	covid	35 783
11	need	15 603	crisis	20 675	virus	19 322	ill	16 821	light	8711	health	15 150	prevent	15 608	get	17 884	advice	17 026	update	31 880
12	buy	15 533	wash	19 889	way	17 297	live	11 737	flight	8151	april	14 490	order	12 346	virus	15 161	impact	15 837	recover	20 718
13	great	15 487	workers	19 854	need	16 838	save	10 946	news	8032	government	14 298	covid	12 205	flu	14 471	coronavirus	15 071	italy	19 487
14	stock	14 953	people	19 015	pandemic	16 666	man	10 617	tweet	7759	try	12 617	avoid	11 616	go	14 316	care	14 386	rise	19 197
15	life	14 556	pay	18 857	happen	15 558	stop	10 426	postpone	7747	announce	12 419	government	11 203	tell	13 915	medical	13 519	toll	16 799
16	hope	14 530	join	16 288	like	15 314	f*ck	10 094	wait	7435	plan	11 427	pandemic	10 805	worry	13 791	pandemic	13 091	news	15 416
17	go	13 123	protect	14 866	let	15 162	self	9964	track	7403	covid	11 392	travel	10 374	hear	12 706	share	12 249	rate	15 004
18	feel	12 517	food	14 261	end	14 944	non	9015	link	7074	government	10 418	close	9815	kid	12 685	article	12 037	china	14 124
19	online	11 881	company	13 018	lie	14 843	healthy	8724	events	6639	result	10 409	media	9742	individuals	12 577	uk	11 935	record	13 704
20	help	11 679	donate	12 317	time	14 799	person	8568	coronavirus	6528	school	9821	take	9416	house	11 841	important	11 898	south	12 878

Table 11
20 most common words per topic for LDA simulation_no = 5.

Rank	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights	Topic 5 words	Topic 5 weights	Topic 6 words	Topic 6 weights	Topic 7 words	Topic 7 weights	Topic 8 words	Topic 8 weights	Topic 9 words	Topic 9 weights
1	coronavirus	97 408	new	70 554	work	52 803	pandemic	74 917	case	158 748	people	79 625	time	67 393	close	36 453	test	93 419	stay	70 601
2	covid	88 860	day	32 103	come	33 210	crisis	37 761	deaths	60 990	virus	76 865	like	49 124	die	32 494	positive	28 183	home	53 343
3	live	39 322	state	27 134	think	25 471	health	27 696	report	53 223	trump	50 604	im	42 410	watch	30 149	social	23 842	hand	39 831
4	update	31 125	coronavirus	25 766	people	25 203	response	25 725	total	50 897	know	40 046	get	42 382	school	25 570	support	22 106	safe	38 183
5	read	29 171	china	24 171	go	25 086	world	23 876	new	49 314	corona	33 760	people	29 892	pm	23 842	people	22 016	spread	35 473
6	latest	23 191	italy	23 218	need	24 780	good	22 929	confirm	41 002	say	27 434	pay	24 708	lockdown	23 502	service	19 032	mask	35 104
7	impact	20 320	news	18 954	thank	17 485	change	20 910	number	40 367	dont	25 802	feel	22 075	video	22 895	medical	18 417	protect	27 861
8	spread	17 853	march	17 243	great	17 372	look	19 450	death	30 334	kill	22 677	care	20 906	government	21 783	travel	18 099	help	21 617
9	economy	17 819	case	17 174	buy	17 295	global	17 987	recover	20 720	world	22 370	leave	20 432	take	16 815	distance	17 377	face	21 379
10	news	16 364	try	16 355	house	17 050	listen	14 562	health	20 685	call	21 310	go	17 581	coronavirus	13 919	provide	17 053	wash	19 902
11	help	15 772	hit	15 433	food	17 047	need	13 174	update	18 478	right	20 072	work	16 628	vaccine	13 491	government	16 960	dont	17 873
12	outbreak	15 018	april	14 376	time	16 296	outbreak	13 062	coronavirus	17 540	go	20 013	workers	16 421	plan	13 350	staff	16 036	use	15 023
13	save	14 394	south	14 038	job	15 914	lead	12 395	rise	16 645	question	19 612	dont	15 932	measure	13 001	fight	15 923	prevent	14 950
14	world	12 417	record	13 598	run	15 189	link	12 195	covid	16 271	die	19 163	sick	15 487	man	11 520	help	15 467	stop	13 858
15	daily	11 463	days	12 987	stock	14 915	time	11 531	positive	13 916	want	17 702	need	14 210	year	10 666	sign	14 650	wear	13 831
16	prepare	10 849	countries	11 422	company	14 615	best	10 804	toll	13 236	think	17 115	f*ck	13 604	say	10 321	government	14 620	family	13 386
17	urge	10 088	lockdown	9992	open	14 161	fight	10 531	hours	12 546	china	17 060	look	12 666	no	10 063	patients	14 389	ms	13 098
18	information	10 005	york	9813	weeks	14 046	read	10 298	india	12 496	ill	16 821	home	12 498	est	9829	india	13 747	avoid	12 115
19	need	9959	uk	9573	years	13 829	risk	10 099	reach	12 309	stop	15 674	right	12 471	play	9698	doctor	13 593	love	11 797
20	cause	9647	city	9366	happen	13 647	current	9735	bring	12 163	lie	15 528	worry	12 426	quarantine	8954	uk	12 933	share	11 539

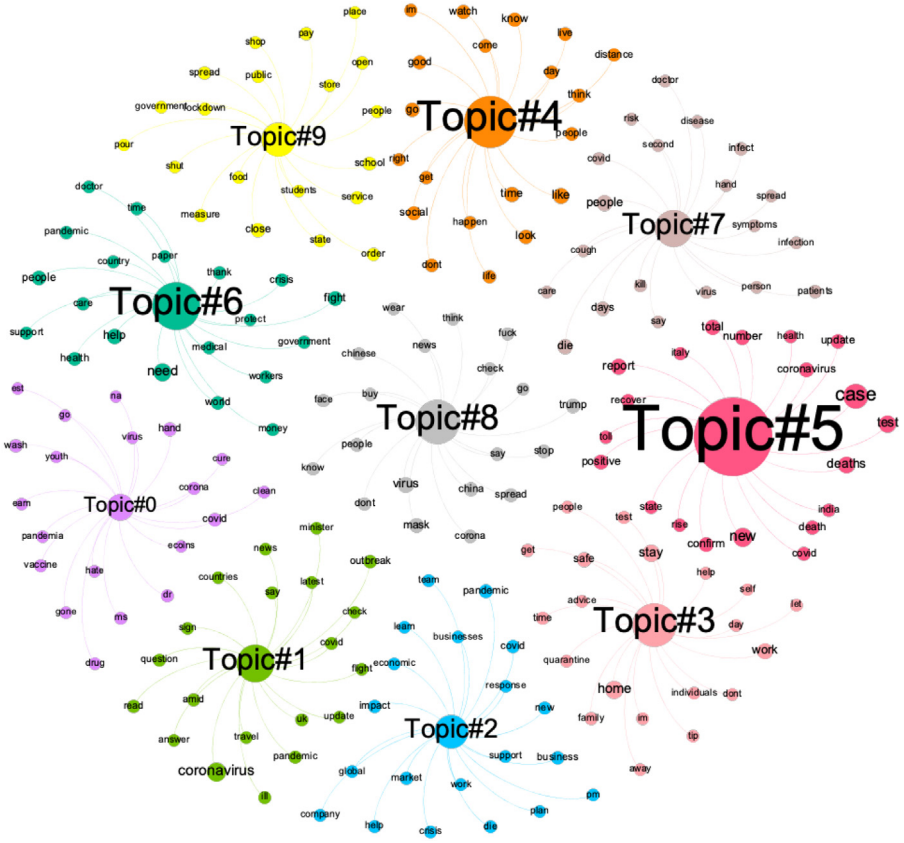

Fig. 9. LDA filtering with 2713 features.

discussions. These may refer to worldwide events such as a pandemic, and it could be implemented in a plethora of topics since there is not a predefined ontology or vocabulary. Our methodology utilizes as a use case the Twitter platform although it could be expanded to additional SM platforms. This would generate better prospects for a holistic compare and contrast analysis on topic extraction for multi-SM applications. For example, we could implement the proposed topic extraction methodology for Facebook, Instagram, Pinterest and more. During the COVID-19 pandemic SM were flashed with unstructured and unfiltered messages presenting opinions and ideas often resulting in negative outcomes and actions. It is usually due to the lack of a mechanism that allows the review of all this free-flow information by experts. This highlights the need for a methodology/tool being able to identify and categorize these data flows under generic topics. At the same time, it may be useful to have a more precise overview of the discussions of topics in SM. On the other hand, SM may also provide valuable medical relevant content. Policy makers such as governments or medical parties can take advantage of a more accurate topic extraction method to engage with the SM public opinion. With such a tool they could grasp in a more precise yet generalized manner the themes of SM discussions. Then if deemed necessary they can intervene by communicating with the public with information and guidelines from experts.

5.2. Limitations

We discuss here limitations that may introduce bias to our methodology. For performing the proposed analysis, we retrieved data using a Twitter crawler as described in Section 3.2. This is just one SM, over a specific period resulting in a number of tweets, since the COVID-19 outburst. More specifically, from 27/2/2020 until 28/8/2020 we collected 2.146.243 unique tweets in a worldwide scale. Inevitably, this dataset (at some point) generates restrictions for a more robust analysis. More SM data sources could be utilized, for example retrieving COVID-19 data from other SM platforms and combining them into one dataset.

There are also data biases due to the preprocessing techniques utilized. Although multiple data preprocessing steps were implemented, it is almost impossible to completely clean the text. For example, we cannot ascertain synonyms of strong words that are used in slang or jargon. The most representative such word in this analysis was COVID-19 and its synonyms. Although we filtered the hashtags within the tweets, people use words such as "corona", "covid", "coronavirus", or simply "virus" and numerous other miss-spellings. Also, there are always missing data that may cause faulty representation of results. This issue could be mitigated by further improving the implemented preprocessing process.

The decisions made regarding the minimum thresholds of support, confidence, lift and leverage, could also introduce bias. During our proposed analysis, we extracted the most frequent wordsets attempting to form new topics, yet with the higher possible leverage and support. Preferably, we aimed for strong rules (high confidence and support values) while taking into consideration leverage and lift. For that reason, the resulting grouping of wordsets to topics may be ambiguous, due to the general subjectivity that this methodology introduces. This threat to validity generates concrete reasons for pursuing extended experimental validation of our proposed methodology regarding ARM.

Yet, the abovementioned limitations and assumptions motivate points for future work as presented in Section 5.3.

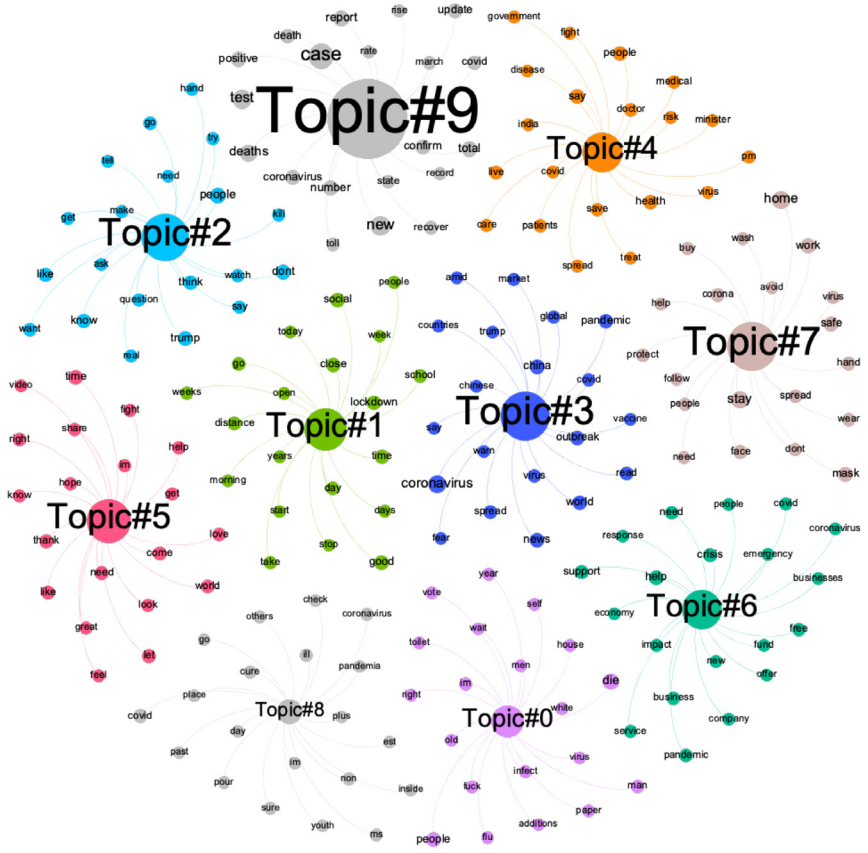

Fig. 10. LDA filtering with 1597 features.

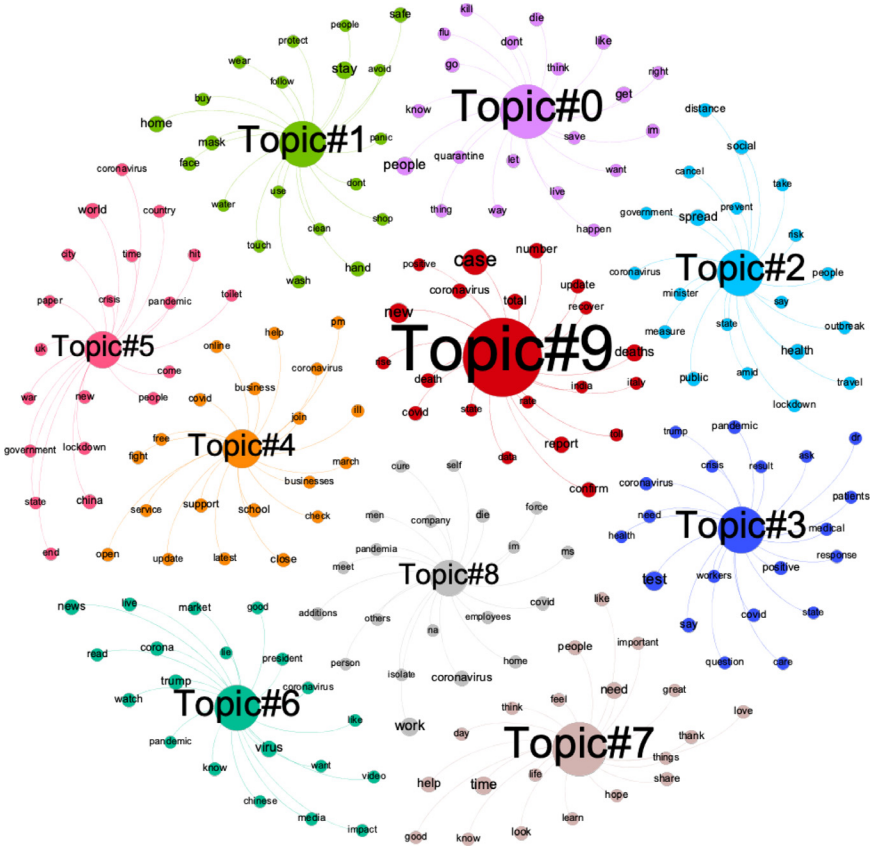

Fig. 11. LDA filtering with 1153 features.

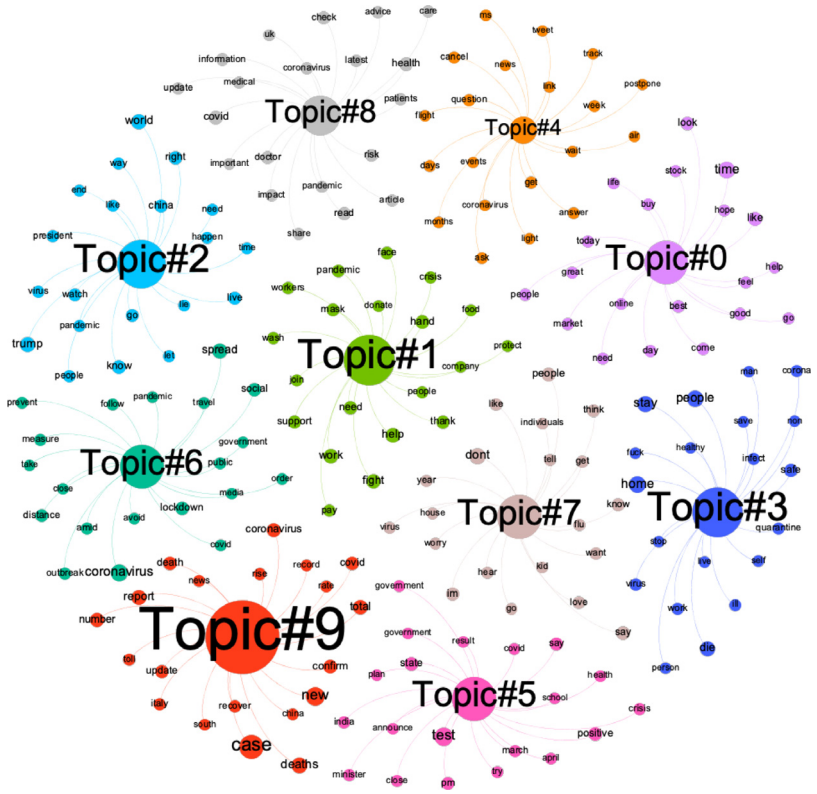

Fig. 12. LDA filtering with 886 features.

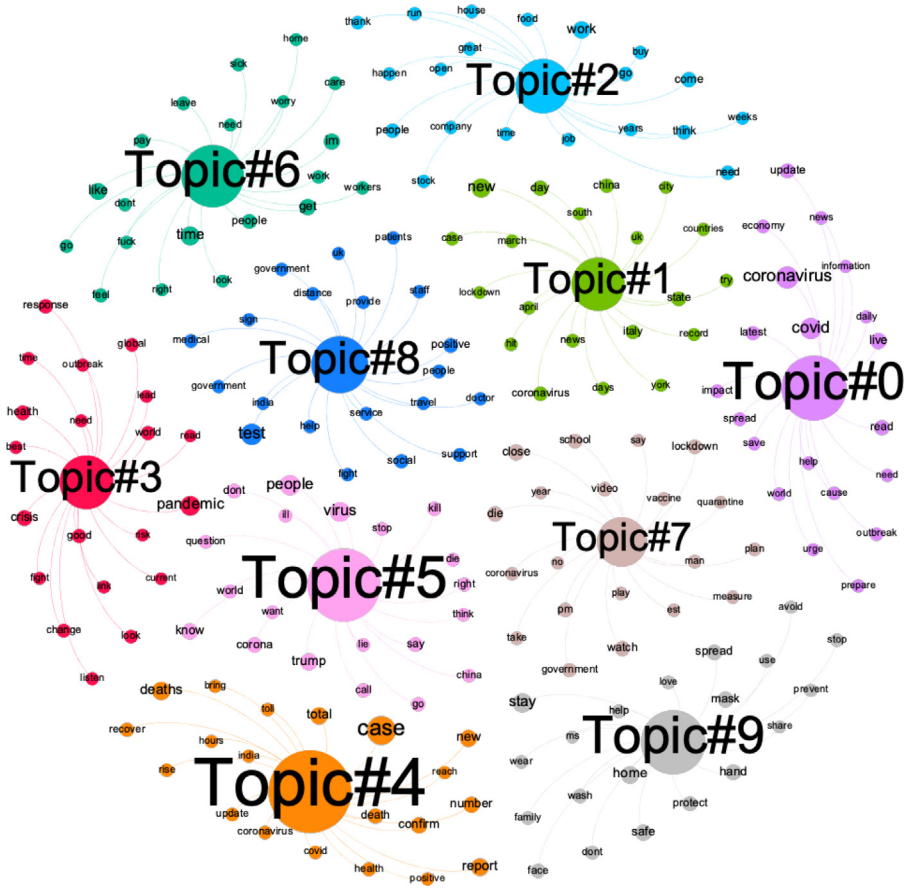

Fig. 13. LDA filtering with 713 features.

Table 12
Top 138 rules showing the strongest leverage, support, confidence, and lift.

Antecedents	Consequents	Antecedent supp	Consequent supp	supp	conf	Lift	Leverage
coronavirus	covid	0,42	0,40	0,30	0,71	1,79	0,132
covid	coronavirus	0,40	0,42	0,30	0,75	1,79	0,132
covid, coronavirus	update	0,30	0,18	0,18	0,60	3,33	0,126
update	covid, coronavirus	0,18	0,30	0,18	1,00	3,33	0,126
know	like	0,20	0,20	0,16	0,80	4,00	0,120
like	know	0,20	0,20	0,16	0,80	4,00	0,120
know, people	think	0,16	0,16	0,14	0,88	5,47	0,114
think	know, people	0,16	0,16	0,14	0,88	5,47	0,114
covid	update	0,40	0,18	0,18	0,45	2,50	0,108
covid	update, coronavirus	0,40	0,18	0,18	0,45	2,50	0,108
know	think	0,20	0,16	0,14	0,70	4,38	0,108
know	think, people	0,20	0,16	0,14	0,70	4,38	0,108
think	know	0,16	0,20	0,14	0,88	4,38	0,108
think, people	know	0,16	0,20	0,14	0,88	4,38	0,108
update	covid	0,18	0,40	0,18	1,00	2,50	0,108
update, coronavirus	covid	0,18	0,40	0,18	1,00	2,50	0,108
go, people	think	0,22	0,16	0,14	0,64	3,98	0,105
think	go, people	0,16	0,22	0,14	0,88	3,98	0,105
coronavirus	update	0,42	0,18	0,18	0,43	2,38	0,104
coronavirus	update, covid	0,42	0,18	0,18	0,43	2,38	0,104
update	coronavirus	0,18	0,42	0,18	1,00	2,38	0,104
update, covid	coronavirus	0,18	0,42	0,18	1,00	2,38	0,104
know, go	think, dont	0,14	0,12	0,12	0,86	7,14	0,103
think, go	know, dont	0,14	0,12	0,12	0,86	7,14	0,103
think, go, people	know, dont	0,14	0,12	0,12	0,86	7,14	0,103
know, go, people	think, dont	0,14	0,12	0,12	0,86	7,14	0,103
think, go	know, dont, people	0,14	0,12	0,12	0,86	7,14	0,103
know, go	think, dont, people	0,14	0,12	0,12	0,86	7,14	0,103
know, dont	think, go	0,12	0,14	0,12	1,00	7,14	0,103
think, dont	know, go	0,12	0,14	0,12	1,00	7,14	0,103
think, dont, people	know, go	0,12	0,14	0,12	1,00	7,14	0,103
know, dont, people	think, go	0,12	0,14	0,12	1,00	7,14	0,103
think, dont	know, go, people	0,12	0,14	0,12	1,00	7,14	0,103
know, dont	think, go, people	0,12	0,14	0,12	1,00	7,14	0,103
think	know, dont	0,16	0,12	0,12	0,75	6,25	0,101
think	know, go, dont	0,16	0,12	0,12	0,75	6,25	0,101
know, people	think, dont	0,16	0,12	0,12	0,75	6,25	0,101
think, people	know, dont	0,16	0,12	0,12	0,75	6,25	0,101
think	know, dont, people	0,16	0,12	0,12	0,75	6,25	0,101
think, people	know, go, dont	0,16	0,12	0,12	0,75	6,25	0,101
know, people	think, go, dont	0,16	0,12	0,12	0,75	6,25	0,101
think	know, go, dont, people	0,16	0,12	0,12	0,75	6,25	0,101
new, coronavirus	case	0,16	0,12	0,12	0,75	6,25	0,101
know, dont	think	0,12	0,16	0,12	1,00	6,25	0,101
know, go, dont	think	0,12	0,16	0,12	1,00	6,25	0,101
know, dont, people	think	0,12	0,16	0,12	1,00	6,25	0,101
know, dont	think, people	0,12	0,16	0,12	1,00	6,25	0,101
think, dont	know, people	0,12	0,16	0,12	1,00	6,25	0,101
know, go, dont, people	think	0,12	0,16	0,12	1,00	6,25	0,101
think, go, dont	know, people	0,12	0,16	0,12	1,00	6,25	0,101
know, go, dont	think, people	0,12	0,16	0,12	1,00	6,25	0,101
case	new, coronavirus	0,12	0,16	0,12	1,00	6,25	0,101
know, think	go, dont	0,14	0,14	0,12	0,86	6,12	0,100
go, dont	know, think	0,14	0,14	0,12	0,86	6,12	0,100
go, dont, people	know, think	0,14	0,14	0,12	0,86	6,12	0,100
know, think, people	go, dont	0,14	0,14	0,12	0,86	6,12	0,100
go, dont	know, think, people	0,14	0,14	0,12	0,86	6,12	0,100
know, think	go, dont, people	0,14	0,14	0,12	0,86	6,12	0,100
go	like, people	0,26	0,16	0,14	0,54	3,37	0,098
go	know, people	0,26	0,16	0,14	0,54	3,37	0,098
go	think	0,26	0,16	0,14	0,54	3,37	0,098
go	think, people	0,26	0,16	0,14	0,54	3,37	0,098
pandemic	crisis	0,26	0,16	0,14	0,54	3,37	0,098
like, people	go	0,16	0,26	0,14	0,88	3,37	0,098
know, people	go	0,16	0,26	0,14	0,88	3,37	0,098
think	go	0,16	0,26	0,14	0,88	3,37	0,098
think, people	go	0,16	0,26	0,14	0,88	3,37	0,098
crisis	pandemic	0,16	0,26	0,14	0,88	3,37	0,098
new	case	0,18	0,12	0,12	0,67	5,56	0,098
new	case, coronavirus	0,18	0,12	0,12	0,67	5,56	0,098
case	new	0,12	0,18	0,12	1,00	5,56	0,098
case, coronavirus	new	0,12	0,18	0,12	1,00	5,56	0,098
know, people	go, dont	0,16	0,14	0,12	0,75	5,36	0,098
think	know, go	0,16	0,14	0,12	0,75	5,36	0,098

(*continued on next page*)

Table 12 (*continued*).

Antecedents	Consequents	Antecedent supp	Consequent supp	supp	conf	Lift	Leverage
know, people	think, go	0,16	0,14	0,12	0,75	5,36	0,098
think, people	know, go	0,16	0,14	0,12	0,75	5,36	0,098
think	know, go, people	0,16	0,14	0,12	0,75	5,36	0,098
think	go, dont	0,16	0,14	0,12	0,75	5,36	0,098
think, people	go, dont	0,16	0,14	0,12	0,75	5,36	0,098
think	go, dont, people	0,16	0,14	0,12	0,75	5,36	0,098
go, dont	know, people	0,14	0,16	0,12	0,86	5,36	0,098
know, go	think	0,14	0,16	0,12	0,86	5,36	0,098
know, go, people	think	0,14	0,16	0,12	0,86	5,36	0,098
know, go	think, people	0,14	0,16	0,12	0,86	5,36	0,098
think, go	know, people	0,14	0,16	0,12	0,86	5,36	0,098
go, dont	think	0,14	0,16	0,12	0,86	5,36	0,098
go, dont, people	think	0,14	0,16	0,12	0,86	5,36	0,098
go, dont	think, people	0,14	0,16	0,12	0,86	5,36	0,098
know	think, dont	0,20	0,12	0,12	0,60	5,00	0,096
know	think, go, dont	0,20	0,12	0,12	0,60	5,00	0,096
know	think, dont, people	0,20	0,12	0,12	0,60	5,00	0,096
dont, people	know, think, go	0,20	0,12	0,12	0,60	5,00	0,096
know	think, go, dont, people	0,20	0,12	0,12	0,60	5,00	0,096
go, people	like	0,22	0,20	0,14	0,64	3,18	0,096
go, people	know	0,22	0,20	0,14	0,64	3,18	0,096
like	go, people	0,20	0,22	0,14	0,70	3,18	0,096
know	go, people	0,20	0,22	0,14	0,70	3,18	0,096
think, dont	know	0,12	0,20	0,12	1,00	5,00	0,096
think, go, dont	know	0,12	0,20	0,12	1,00	5,00	0,096
think, dont, people	know	0,12	0,20	0,12	1,00	5,00	0,096
think, go, dont, people	know	0,12	0,20	0,12	1,00	5,00	0,096
know, think, go	dont, people	0,12	0,20	0,12	1,00	5,00	0,096
go, people	know, dont	0,22	0,12	0,12	0,55	4,55	0,094
dont	get, people	0,22	0,12	0,12	0,55	4,55	0,094
dont	know, think, go	0,22	0,12	0,12	0,55	4,55	0,094
go, people	think, dont	0,22	0,12	0,12	0,55	4,55	0,094
go, people	know, think, dont	0,22	0,12	0,12	0,55	4,55	0,094
dont	know, think, go, people	0,22	0,12	0,12	0,55	4,55	0,094
know, dont	go, people	0,12	0,22	0,12	1,00	4,55	0,094
get, people	dont	0,12	0,22	0,12	1,00	4,55	0,094
know, think, go	dont	0,12	0,22	0,12	1,00	4,55	0,094
think, dont	go, people	0,12	0,22	0,12	1,00	4,55	0,094
know, think, go, people	dont	0,12	0,22	0,12	1,00	4,55	0,094
know, think, dont	go, people	0,12	0,22	0,12	1,00	4,55	0,094
need	time	0,28	0,24	0,16	0,57	2,38	0,093
time	need	0,24	0,28	0,16	0,67	2,38	0,093
know	go, dont	0,20	0,14	0,12	0,60	4,29	0,092
dont, people	know, go	0,20	0,14	0,12	0,60	4,29	0,092
know	go, dont, people	0,20	0,14	0,12	0,60	4,29	0,092
know	think, go	0,20	0,14	0,12	0,60	4,29	0,092
know	think, go, people	0,20	0,14	0,12	0,60	4,29	0,092
dont, people	think, go	0,20	0,14	0,12	0,60	4,29	0,092
dont, people	know, think	0,20	0,14	0,12	0,60	4,29	0,092
go, dont	know	0,14	0,20	0,12	0,86	4,29	0,092
go, dont, people	know	0,14	0,20	0,12	0,86	4,29	0,092
know, go	dont, people	0,14	0,20	0,12	0,86	4,29	0,092
think, go	know	0,14	0,20	0,12	0,86	4,29	0,092
think, go, people	know	0,14	0,20	0,12	0,86	4,29	0,092
think, go	dont, people	0,14	0,20	0,12	0,86	4,29	0,092
know, think	dont, people	0,14	0,20	0,12	0,86	4,29	0,092
go, people	dont	0,22	0,22	0,14	0,64	2,89	0,092
dont	go, people	0,22	0,22	0,14	0,64	2,89	0,092
time	look	0,24	0,12	0,12	0,50	4,17	0,091
virus	corona	0,24	0,12	0,12	0,50	4,17	0,091
im	get	0,18	0,16	0,12	0,67	4,17	0,091
get	im	0,16	0,18	0,12	0,75	4,17	0,091
look	time	0,12	0,24	0,12	1,00	4,17	0,091
corona	virus	0,12	0,24	0,12	1,00	4,17	0,091

5.3. Future work

In the future, we aim at further mitigating the biases identified in the limitations section. In addition, we plan to:

I. This paper also envisions the developed methodology as a part of a complete Decision Support System (DSS). This will engage in predictive and prescriptive analytics [33] utilizing SM historical data to forecast public attitudes/ sentiment regarding healthcare issues. Implications of such a DSS involve policy makers when taking actions for mitigating issues arising during a worldwide crisis.
II. Expand on algorithmic improvements regarding ARM options [34]. One of the basic concerns of the methodology refers to decisions made regarding the appropriate values of minimum of support and minimum confidence for finding the most frequent item sets and extracting the rules, respectively. We would like to investigate possible options that effectively perform this process in an automated way,

without manually setting the support and confidence levels [35].

III. Create an algorithm that automatically classifies input tweets to the generated topics resulting from this research. That way topic extraction will have a dynamic extraction feature, enabling real-time monitoring and classification of SM data input streams.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Topic extraction results

See Tables 7–11.

Appendix B. Graph visualization topic extraction

See Figs. 9–13.

Appendix C. Strongest association rules

See Table 12.

References

[1] P. Koukaras, C. Tjortjis, D. Rousidis, Social media types: introducing a data driven taxonomy, Computing 102 (1) (2020) 295–340, http://dx.doi.org/10.1007/s00607-019-00739-y.

[2] H. Tankovska, • number of social media users 2025 | statista, jan 28, 2021, 2021, https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (accessed May 31, 2021).

[3] D. Rousidis, P. Koukaras, C. Tjortjis, Social media prediction: a literature review, Multimedia Tools Appl. 79 (9–10) (2020) 6279–6311, http://dx.doi.org/10.1007/s11042-019-08291-9.

[4] I. Vayansky, S.A.P. Kumar, A review of topic modeling methods, Inf. Syst. 94 (2020) 101582, http://dx.doi.org/10.1016/j.is.2020.101582.

[5] D.M. Blei, A.Y. Ng, M.T. Jordan, Latent dirichlet allocation, Adv. Neural Inf. Process. Syst. 3 (2002) 993–1022.

[6] T. Haerder, A. Reuter, Principles of transaction-oriented database recovery, ACM Comput. Surv. 15 (4) (1983) 287–317, http://dx.doi.org/10.1145/289.291.

[7] M. Cataldi, L. Di Caro, C. Schifanella, Emerging topic detection on Twitter based on temporal and social terms evaluation, in: Proceedings of the 10th International Workshop on Multimedia Data Mining, MDMKDD '10, 2010, pp. 1–10, http://dx.doi.org/10.1145/1814245.1814249.

[8] W.X. Zhao, et al., Topical keyphrase extraction from Twitter, in: ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, 2011, pp. 379–388.

[9] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, H. Wang, Entity-centric topic-oriented opinion summarization in twitter, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 379–387, http://dx.doi.org/10.1145/2339530.2339592.

[10] R. Pochampally, V. Varma, User context as a source of topic retrieval in Twitter, 2011, pp. 1–3, in … on Enriching Information Retrieval …, no. Enir, [Online]. Available: http://select.cs.cmu.edu/meetings/enir2011/papers/pochampally-varma.pdf.

[11] L.M. Aiello, et al., Sensing trending topics in twitter, IEEE Trans. Multimedia 15 (6) (2013) 1268–1282, http://dx.doi.org/10.1109/TMM.2013.2265080.

[12] X. Wang, M.S. Gerber, D.E. Brown, Automatic crime prediction using events extracted from twitter posts, in: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, in: LNCS, vol. 7227, 2012, pp. 231–238, http://dx.doi.org/10.1007/978-3-642-29047-3_28.

[13] M. Cordeiro, Twitter event detection: combining wavelet analysis and topic inference summarization, in: The Doctoral Symposium on Informatics Engineering - DSIE'12, Vol. 1, 2012, pp. 11–16, [Online]. Available: http://paginas.fe.up.pt/~prodei/dsie12/papers/paper_14.pdf.

[14] D. Alvarez-Melis, M. Saveski, Topic modeling in Twitter: Aggregating tweets by conversations, in: Proceedings of the 10th International Conference on Web and Social Media, in: ICWSM 2016, vol. 10, 2016, pp. 519–522, (1).

[15] A. Rafea, N.A. Mostafa, Topic extraction in social media, in: Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013, 2013, pp. 94–98, http://dx.doi.org/10.1109/CTS.2013.6567212.

[16] H.J. Choi, C.H. Park, Emerging topic detection in twitter stream based on high utility pattern mining, Expert Syst. Appl. 115 (2019) 27–36, http://dx.doi.org/10.1016/j.eswa.2018.07.051.

[17] S.H. Yang, A. Kolcz, A. Schlaikjer, P. Gupta, Large-scale high-precision topic modeling on twitter, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 1907–1916, http://dx.doi.org/10.1145/2623330.2623336.

[18] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hai, Z. Shah, Top concerns of tweeters during the COVID-19 pandemic: A surveillance study, J. Med. Internet Res. 22 (4) (2020) e19016, http://dx.doi.org/10.2196/19016.

[19] S. Noor, Y. Guo, S.H.H. Shah, P. Fournier-Viger, M.S. Nawaz, Analysis of public reactions to the novel coronavirus (COVID-19) outbreak on Twitter, Kybernetes (2020) http://dx.doi.org/10.1108/K-05-2020-0258.

[20] Z.T. Osakwe, I. Ikhapoh, B.K. Arora, O.M. Bubu, Identifying public concerns and reactions during the COVID-19 pandemic on Twitter: A text-mining analysis, Public Health Nurs. 38 (2) (2021) 145–151, http://dx.doi.org/10.1111/phn.12843.

[21] C. Wang, J. Paisley, D.M. Blei, Online variational inference for the hierarchical Dirichlet process, J. Mach. Learn. Res. 15 (2011) 752–760.

[22] P. Jaccard, The distribution of the flora in the alpine zone, New Phytol. 11 (2) (1912) 37–50, http://dx.doi.org/10.1111/j.1469-8137.1912.tb05611.x.

[23] D. Greene, D. O'Callaghan, P. Cunningham, How many topics? Stability analysis for topic models, in: Joint European Conf. on Machine Learning and Knowledge Discovery in Databases, in: LNAI, vol. 8724, 2014, pp. 498–513, http://dx.doi.org/10.1007/978-3-662-44848-9_32, no. PART 1.

[24] M. Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, in: WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining, 2015, pp. 399–408, http://dx.doi.org/10.1145/2684822.2685324.

[25] G. Piatetsky-Shapiro, Discovery, analysis, and presentation of strong rules, Knowl. Discov. Databases (1991) 229–248, [Online]. Available: http://ci.nii.ac.jp/naid/10000000985/.

[26] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in datamining, Int. J. Sci. Technol. Res. 2 (12) (2013) 13–24.

[27] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, ACM Sigmod Rec. 29 (2) (2000) 1–12.

[28] S. Yakhchi, S.M. Ghafari, C. Tjortjis, M. Fazeli, ARMICA-improved: A new approach for association rule mining, in: Int'l Conf. on Knowledge Science, Engineering and Management, in: LNAI, vol. 10412, 2017, pp. 296–306, http://dx.doi.org/10.1007/978-3-319-63558-3_25.

[29] J. Hipp, U. Güntzer, G. Nakhaeizadeh, Algorithms for association rule mining — a general survey and comparison, ACM SIGKDD Explor. Newsl. 2 (1) (2000) 58–64, http://dx.doi.org/10.1145/360402.360421.

[30] M. Hahsler, B. Grün, K. Hornik, Arules - a computational environment for mining association rules and frequent item sets, J. Stat. Softw. 14 (15) (2005) 1–25, http://dx.doi.org/10.18637/jss.v014.i15.

[31] S. Brin, R. Motwani, J.D. Ullman, S. Tsur, Dynamic itemset counting and implication rules for market basket data, SIGMOD Rec. (ACM Spec. Interest Group Manage. Data) 26 (2) (1997) 255–264, http://dx.doi.org/10.1145/253262.253325.

[32] J. Han, M. Kamber, J. Pei, Data mining: Concepts and techniques, Data Min. Concepts Tech. 5 (4) (2012) 83–124, http://dx.doi.org/10.1016/C2009-0-61819-5.

[33] P. Koukaras, C. Tjortjis, Social media analytics, types and methodology, in: Machine Learning Paradigms, Springer, 2019, pp. 401–427.

[34] S.M. Ghafari, C. Tjortjis, A survey on association rules mining using heuristics, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9 (4) (2019) e1307, http://dx.doi.org/10.1002/widm.1307.

[35] S.M. Ghafari, C. Tjortjis, Association rules mining by improving the imperialism competitive algorithm (ARMICA), IFIP Adv. Inf. Commun. Technol. 475 (2016) 242–254, http://dx.doi.org/10.1007/978-3-319-44944-9_21.