New Approaches to Extract Information from Posts on COVID-19 Published on Reddit

DOI: 10.1142/S0219622022500213

New approaches to extract information from posts on COVID-19 published on Reddit

Abstract

In the last two years, we have seen a huge number of debates and discussions on COVID-19 in social media. Many authors have analyzed these debates on Facebook and Twitter, while very few ones have considered Reddit. In this paper, we focus on this social network and propose three approaches to extract information from posts on COVID-19 published in it. The first performs a semi-automatic and dynamic classification of Reddit posts. The second automatically constructs virtual subreddits, each characterized by homogeneous themes. The third automatically identifies virtual communities of users with homogeneous themes. The three approaches represent an advance over the past literature. In fact, the latter lacks studies regarding classification algorithms capable of outlining the differences among the thousands of posts on COVID-19 in Reddit. Analogously, it lacks approaches able to build virtual subreddits with homogeneous topics or virtual communities of users with common interests.

Keywords: COVID-19; Reddit; Information Extraction; Hierarchical Classification; Backtracking; Social Network Analysis; Community Detection

1 Introduction

COVID-19 is a severe disease that is upsetting the world. It is affecting nearly every aspect of human life, from healthcare to economy, from education to tourism, and so on. That is why it has provoked, and continues to provoke, an enormous debate among experts and ordinary people alike. In this context, it is inevitable that COVID-19 is also one of the most user-focused topics in social networks. This fact has aroused the interest of Social Network Analysts, who have already proposed several studies on how COVID-19 has been treated in the main social networks (see, for example, the studies reported in [17, 80, 24, 20, 9, 57], just to mention a few).

The variety of issues related to COVID-19, along with the variety of social networks and, more generally, social media and journals discussing them, opens up interesting challenges. In fact, it is worth observing how the same issue arouses debates very heterogeneous in content and modalities, depending on the medium in which they take place and the people participating in them. On one hand, we have the major generalist networks, such as Facebook[1] and Twitter[2], which are very widespread. Because of the intrinsic characteristics of these networks, users are led to write their posts very

[1] `https://www.facebook.com`
[2] `https://www.twitter.com`

frequently and "on the fly". Therefore, these networks have the merit of immediately revealing the feelings of their users about the issue they are discussing. However, such feelings could be very fickle, as a user often writes on these networks without careful meditation [35]. As a consequence, it may happen that she/he takes completely different positions on the same issue during the same day as she/he reflects better on the subject she/he is debating. Since these networks are generalist, both common users and specialists in various fields (e.g., virologists, epidemiologists, economists, politicians, etc.) write on them. On the other hand, we have scientific networks and social media. In this case, users are specialists in their fields. Therefore, they are physicians in medical social media and journals, economists in business social media and journals, and so on. Content written in this context is very thoughtful and, in the case of research journals, is also peer-reviewed. Between these two extremes there are several intermediate cases. A very interesting one is the case of generalist social networks that are very popular but not as widespread as Facebook and Twitter. In them, writers do not usually publish their content "out of the blue", as people do on Facebook or Twitter, but periodically, for example at the end of a day [56]. As a result, what is written in these social networks is more meditated than the content published in Facebook or Twitter. However, differently from specialized media, anyone (and not only specialists) can publish on them.

We believe that this last category of networks deserves great attention because of the intermediate nature between the two extremes highlighted above and because of their considerable diffusion. In fact, we could extract from it information different from both the one retrievable from Facebook and Twitter and the one retrievable from specialized social media. One of these networks is Reddit[3]. It is currently one of the most active social media. As a matter of fact, Alexa's top 500 global sites[4] currently ranks it on 21^{st} place of the ranking of the most accessed social media, with more than 430 million active users every month[5]. Reddit can be considered as a heterogeneous collection of forums. Its members can share news and content; furthermore, they can comment and vote on news and content posted by the other members.

In this paper, we aim to extract information from posts on COVID-19 published on Reddit. In particular, we propose three approaches. The first is a hierarchical classification algorithm for the posts on COVID-19 published in Reddit. The second is an algorithm capable of identifying a set of homogeneous themes regarding the COVID-19 disease discussed by users. The third is an algorithm capable of identifying a number of user communities showing homogeneous interests. We applied these three approaches to all the posts related to COVID-19 published in Reddit from January 9^{th}, 2020 to April 30^{th}, 2020. The number of posts considered is almost two and half million. Here we illustrate in detail this activity and the corresponding results obtained.

The three approaches proposed have been conceived with reference to COVID-19. However, we point out that they are general and can be used to extract information about any other issue that may cause an intense posting activity on Reddit.

The outline of this paper is as follows: In Section 2, we examine the related literature. In Section 3, we illustrate the proposed approaches. In Section 4, we describe the experiments we carried out to evaluate them. Finally, in Section 5, we draw our conclusions and look at some possible future

[3] `https://www.reddit.com`

[4] `https://www.alexa.com/topsites`

[5] `https://www.redditinc.com`

developments.

2 Related literature

This section is organized into five subsections. In Subsection 2.1, we present related literature on Reddit. In Subsection 2.2, we examine several approaches for studying the usage of social networks during past and current pandemics and disasters. We dedicate Subsection 2.3 to present approaches for studying the diffusion of information on COVID-19 in Reddit. In Subsection 2.4, we present several approaches for investigating how dramatic events can affect emotionality in social posting w.r.t. COVID-19. Finally, in Subsection 2.5, we highlight some limitations of existing approaches and illustrate what problems we are targeting in this paper.

2.1 Related literature on Reddit

Reddit has been largely investigated in the past literature. An overview of the academic research on this social network can be found in [52]. Here, the authors illustrate the main research directions primarily focused on it. In [16], the authors aim at knowing if the moderation and banning rules of Reddit are effective in terms of decrease of hateful behavior. They find that this reduction is confirmed but many subreddits maintain a high level of hateful speech even after those targeted bans. In [26], the authors aim at addressing the problem of predicting community endorsement in online discussions. To reach this goal, they leverage both the participant response structure and the comment text. In [21], the authors study the characteristics of NSFW (Not Safe For Work) posts in Reddit. They extract three knowledge patterns on the main differences between NSFW and SFW posts. Thanks to these patterns, they are able to better understand the dynamics behind NSFW posts. In [31], the authors present an analysis on crowd and platform manipulations of news. In particular, they investigate known features for predicting news popularity and how those features may change on Reddit.

In [32], the authors study a large dataset of Reddit comments from 11 subreddits with different properties. They introduce several features, like sentiment, relevance and content analysis. The authors of [61] examine Reddit during a period of community unrest affecting millions of users in the summer of 2015. They realize large-scale changes in user behavior and migration patterns to Reddit-like alternative platforms. They find an important pull factor that enabled Reddit to retain users, i.e., the niche content. The author of [81] aims at understanding the relationship between intrinsic article quality and popularity in Reddit and HackerNews. In [14], the authors first propose a definition and an analysis of several stereotypes of both subreddits and authors. Then, they investigate the possible existence of author assortativity in Reddit focusing on co-posters.

In [1], the authors study the differences about lexical, topical and emotional expressions between females and males. To do this, they use a dataset collected from a set of subreddits, where authors commonly self-report their gender (like `r/askmen` or `r/askwomen`). In [45], the authors examine the content of Reddit by applying trace ethnography methods to understand how the architecture and operation mode of this social network impact information visibility. In [83], the authors analyze different characteristics of the content posted on Reddit, in order to determine whether this platform benefits from a freedom from the press or not. To achieve this objective, they build a dataset of posts

published from April 2013 to April 2014 regarding the Boston Marathon bombing.

2.2 Related literature on the use of social networks in pandemics and disasters

Social media have already played a key role during emergencies [37]. Indeed, a social medium can easily spread a message to a huge number of users; therefore, it could be useful for emergency response managers to know what is happening in real time. This has led researchers to analyze the content shared by users during past pandemic outbreaks [18, 75, 79, 55, 82]. Since December 2019, when the first cases of COVID-19 were reported in Wuhan (China), the conversations about it have increased in Twitter and Facebook, as well as in other social network platforms. In [17], the authors started collecting tweets, from January 28^{th}, 2020, continuously monitoring Twitter's trending topics, keywords, and sources associated with COVID-19, to capture conversations related to the outbreak. Social networks are also leveraged to convey misinformation, myths and other low quality news. In [80], the authors analyze 5 types of myths emerged during the crisis: flu comparison, heat kill disease, home remedies, theories about the origin of COVID-19 and vaccine development.

Another interesting analysis concerns the sentiment of people communicating through social networks [50, 89]. In [24], the authors leverage the NRC Word-Emotion Association Lexicon[6], which contains 10,170 lexical items, in order to study the positive and negative polarity of tweets. Furthermore, they classify the eight emotions defined by the psychologist Robert Plutchik [68].

Along with Twitter and Facebook, Reddit is a social platform where users discuss historical phenomena too. Indeed, the past literature provides us with several studies in which researchers analyze the behavior of users through their posts and comments during different events (such as the Boston Marathon bombing [83], the Hurricane Sandy [44], and many others). The ultimate goal is the derivation of interesting patterns that could deeply describe the whole scenario, or that could be leveraged to perform a prediction activity [77, 4, 43, 22, 65, 71, 49, 51]. For instance, in [43], the authors investigate how the Human Papillomavirus (hereafter, HPV) vaccine is mentioned on Reddit over time. The results obtained show that the discussions involving HPV vaccine encompass mostly political debates and discussion on cancer risk for both men and women. In [4], the authors study how a social news platform shapes which event news are relevant for a user.

2.3 Related literature on the use of Reddit during the COVID-19 pandemic

In this section, we illustrate the efforts that researchers made to analyze Reddit during these months of COVID-19 pandemic. To the best of our knowledge, at the time of writing this paper, few studies consider Reddit along with other social networks [20, 9], and even fewer ones focus only on the Reddit perspective.

In [90], the authors analyze discussions on COVID-19 in the `r/China_flu` and `r/Coronavirus` subreddits. In particular, they analyze from which subreddits the founders and early users of these subreddits come. Moreover, they study how subreddits regarding COVID-19 arose from these two and derive a genealogy of subreddits on this topic. They also study the differences between these subreddits, especially in the moderation rules, which led to two different evaluations of them. They

[6] `http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`

also identify a list of keywords highly used by the members of the two subreddits. Finally, they show how the pandemic has changed users' habits on Reddit. In [27], the authors analyze online media coverage regarding COVID-19 in four countries, i.e. Italy, the United Kingdom, the United States and Canada. For this purpose, they use several data sources, including Wikipedia and Reddit. As for the latter, they consider posts and comments published in the `r/Coronavirus` subreddit. Their ultimate goal is to analyze the relationship between media coverage, epidemic progression and user responses on the Web. In [57], the authors investigate posts made by coronavirus-positive patients published in the subreddit `r/COVID19Positive`. Their ultimate goal is getting insights into personal struggles with the virus. Through topic modelling and sentiment analysis approaches, they identify two clusters of positive and negative emotions associated with the evolution of patient symptoms. These clusters could be leveraged to reveal potential mental issues.

2.4 Related literature on the sentiment analysis during the COVID-19 pandemic

In the past literature, there exist several studies aiming to investigate the sentiment of posts published during the COVID-19 pandemic [58, 15, 91]. In [8], the authors propose a new model that uses five classifiers and combines them to improve the overall classification output. They also observe that the appearance of information on SARS-COV-2 is related to the first reported infected cases and deaths. Finally, they show that each country has specific sentiment patterns, and that the maximum negative sentiment values occur when new cases of infections and COVID-19 related deaths appear. In [91], the authors analyze public opinions on COVID-19 after the Wuhan closure and during the spring festival in China. Specifically, they extract topics from social media posts by means of the Latent Dirichlet Allocation (LDA) model and analyze them from temporal and spatial perspectives. In [87], the authors investigate the impact of COVID-19 by extracting and analyzing tweet sentiments. Their analysis leads them to conclude that people post their views or opinions on different topics related to this issue, such as frontline workers, rise in new infected cases, travel restrictions and effects of COVID-19. They also conclude that people appreciate all the frontline efforts and motivate each other to follow all the precautions. In [15], the authors show how COVID-19 related tweets and the World Health Organization were unsuccessful in guiding people during the pandemic. They show that people tweeted mostly in a positive way regarding COVID-19. However, a portion of them only propagated negative tweets, with no useful suggestions. They support their claim with a deep learning based classification model that reaches an accuracy of up to 81%.

Other papers went one step further in this research topic and analyzed how dramatic events can affect emotionality in social posting during the COVID-19 pandemic [7, 38, 48, 60]. In [7], the authors focus on the effects that the pandemic had on ordinary social posting life on Reddit. They figure out that: *(i)* the posting behavior, in terms of submissions, comments and scores, was strongly affected by extreme events; *(ii)* emotion expression was amplified during extreme events; *(iii)* non-standard emotions (such as skepticism to an increase of new cases) can be interpreted as a forerunner for new events. In [88], the authors analyze comments on daily posts containing updates on COVID-19 statistics from a location-specific subreddit. In this way, they are able to study changes in web-based engagement, discussion and emotional expression to new infected cases and vaccination rates. Their study shows that data from social media can be used to better understand concerns and sentiments

surrounding the pandemic at the local level, which enables more targeted and publicly acceptable policies. In [3], the authors monitor the evolution of people's thoughts about vaccines over a one-year period. In their analysis, they figure out that as vaccine diffusion increases, positive changes can be observed in people's thought. In [60], the authors develop a model to analyze the emotional nature of various tweets posted during the COVID-19 pandemic. They use a recurrent neural network for emotional prediction, search for connection between words and mark them with positive and negative emotions. They also show that positivity has strengthened over time, albeit there is also a stronger negative reaction, which could be natural.

2.5 Contribution of this paper to the related literature

The previous sections highlight how the past literature on Reddit is extremely rich and varied. A similar claim applies to the past literature regarding the use of social networks in pandemics and disasters. Going more specifically into the past literature on the use of Reddit during the COVID-19 pandemic, we could see that there are already several papers that have dealt with these topics. In particular, we have mentioned [90, 27, 57]. These three papers present some similarities but also many differences with our approaches.

For instance, [90] focuses on two subreddits, while our approach analyzes posts on COVID-19 present in several subreddits. In addition, [90] proposes an in-depth study of the difference in language between two communities. This issue is not considered in this paper, which, however, proposes an approach for virtual subreddit construction and another for virtual community definition. Furthermore, [27] only considers the `r/Coronavirus` subreddit, along with other data sources different from Reddit. By contrast, our approach considers all posts with the terms "covid" and/or "coronavirus" published on any subreddit. The main goal of [27] is the study of user behavior starting from media coverage, which is different from the goals of our approach. Finally, in [57], the authors mainly extract keywords related to symptoms derived from COVID-19 to reconstruct the disease progression based on them. Instead, the topics considered in our approach are more general. Moreover, the authors of [57] analyze the evolution of the keywords over time because they are interested in the disease progression. We do not make this kind of analysis, because it is out of the scope of our approaches.

Despite the enormous amount of related literature that already exists, it is possible (as with any research field) to identify some limitations and, therefore, some challenging issues to be addressed. In particular, we observe a lack of studies regarding classification algorithms that could outline the differences among the thousands of posts on COVID-19 in Reddit. Analogously, there is a lack of approaches able to build virtual subreddits with homogenous themes or virtual communities of users with common interests. This paper wants to provide a contribution in this setting. In particular, we propose three approaches.

The first one performs a hierarchical classification for the posts on COVID-19 published in Reddit. Each class is characterized by a set of keywords. A post can belong to several classes. The algorithm is semi-automatic, because the definition of the initial class hierarchy is done with the support of a human expert. This works with the most frequent and characterizing keywords found in Reddit posts on COVID-19. The fact that the algorithm is semi-automatic implies a considerable expense of initial resources. This suggests the idea that, once the initial class hierarchy is built, the algorithm should

become automatic and incremental. As can be seen by examining the approaches described in [40], the main problem with most of the hierarchical classification and clustering algorithms is the lack of backtracking. This implies that if, at a certain level of the hierarchy, a post is assigned to the wrong class, this error propagates to all the next levels. In other words, there is no mechanism such that if a classification error is found at a certain level of the hierarchy, one can go back to the previous levels and make the suitable corrections [30]. The algorithm we propose, in addition to being incremental and automatic (once the initial hierarchy is built), provides some backtracking mechanisms that allow the correction of classification errors made previously, the correction of errors in the class hierarchy itself, as well as its evolution over time.

The second approach aims to identify a set of homogeneous themes regarding the COVID-19 disease discussed by users. Starting from a set of real subreddits, it defines homogeneous virtual ones. To this end, it first identifies the most frequent keywords in Reddit in the time interval of interest. Next, it creates a virtual subreddit for each of these keywords, which, therefore, acts as a seed. In fact, the subreddit has the keyword as its initial topic. Then, starting from it, our approach identifies and assigns other topics to the subreddit. To this end, it selects the other keywords that co-occur most frequently with the seed keyword in the period of interest. The virtual subreddits thus obtained can attract users interested in finding all the posts related to a certain theme in one place. Such homogeneous virtual subreddits can be seen as the leaf nodes of a class hierarchy, similar to the one considered by the first approach. Therefore, it is possible to apply the incremental algorithm of the first approach to them. This aims at obtaining a hierarchy of virtual subreddits and allowing them to evolve automatically over time, based on the evolution of the post content.

The third approach aims to identify a number of user communities showing homogeneous interests. It builds virtual communities of users interested in the same topics and can be employed, for example, to recommend other users with similar interests. It could also be adopted by Reddit itself to proactively promote communities of users with similar interests. Again, these communities can be seen as the leaves of a class hierarchy. Therefore, once they have been identified, it is possible to apply the incremental algorithm described in the first approach to them. This aims at obtaining a hierarchy of user communities with homogeneous interests and allowing it to evolve automatically over time, based on the evolution of the post content published by users.

3 Description of the proposed approaches

3.1 Approach to classify posts based on topics

3.1.1 Approach description

In Reddit, the COVID-19 disease is dealt from many points of view. Therefore, it seems useful to think about defining a classification of COVID-19 posts in Reddit based on their content. This classification cannot be exclusive because a post can belong to more than one class. Furthermore, it can be hierarchical [47] because, by adopting different abstraction levels, two or more classes of a lower level can be grouped into a class of a higher level.

Given the novelty of the COVID-19 disease and the various terms used to describe it, the definition of the initial class hierarchy can be only semi-automatic. In other words, the support of the human

expert is needed to identify at least the leaf classes of the hierarchy. The human expert examines the main keywords associated with the posts as they are derived from any text mining approach (such as the ones described in [67, 53, 74, 41, 39]). Starting from this examination, she/he identifies the leaf classes and, then, associates a set of representing keywords with each of them. Two or more classes sharing a minimum number of keywords are considered siblings and can be "merged" into a single class at the higher abstraction level. The set of keywords of the new class will be equal to the union of the sets of keywords of the starting classes. Proceeding this way, after several abstraction levels, the model will result in a single tree, if there is at least one keyword common to all classes, or a forest of trees, if not.

Once the initial hierarchy is built, the assignment of posts to the corresponding classes can be done automatically. For this purpose, it is necessary to identify a measure of similarity between the keywords of a post and those of a class, and a mechanism that, based on this measure, decides whether or not a post belongs to a certain class. As far as the measure of similarity is concerned, we thought to adopt the Jaccard coefficient taking the semantic relationships (e.g., synonymies, homonymies) between keywords into account.

In particular, if CS_i indicates the set of keywords of the class C_i, and PS_k denotes the set of keywords of the post P_k, the enhanced Jaccard coefficient J_{ik}^+ between C_i and P_k is defined as:

$$J_{ik}^+ = \frac{|CS_i \sqcap PS_k|}{|CS_i \sqcup PS_k|}$$

where $\sqcap$ (resp., $\sqcup$) denotes the enhanced intersection (resp., union) between the keywords in such a way as to take into account the synonymies and homonymies as stored in a suitable thesaurus, like Babelnet [59]. J_{ik}^+ belongs to the real interval $[0, 1]$.

We are now able to define an automatic approach for determining whether a post belongs to a class. Since multiple class memberships are allowed, i.e., a post can belong to more than one class, for leaf classes it is sufficient to define a threshold th_J and to establish that P_k belongs to C_i if $J_{ik}^+ \geq th_J$. The higher th_J, the fewer the classes which P_k will belong to. From a theoretical point of view, it is appropriate for the value of th_J to be low in order to encourage a post to belong to multiple classes. Based on this idea, we performed experiments to find the optimal value of this threshold. Due to space constraints, we do not report such experiments in detail. We only say that at the end of them we found that the optimal value of th_J is 0.25. If C_i is a non-leaf class, P_k belongs to C_i if it belongs to at least one child of C_i.

The content of a social network is very dynamic, so a classification cannot remain unchanged over time. As new posts arrive, new keywords emerge, which can stimulate the appearance of new classes. At the same time, other keywords become obsolete, which can lead to the disappearance of some classes or their inclusion into others. Finally, two or more classes may have to be merged into one class because they have become very similar. All this led us to the definition of an incremental and automatic algorithm for updating the original classification. This algorithm is important because it is well known that one of the weak points of most hierarchical clustering or hierarchical classification algorithms is the lack of backtracking [30]. Instead, our approach is provided with some backtracking mechanisms and, therefore, is able to fix any possible classification error performed in the past and to support the evolution of the hierarchy over time.

In order to operate, our algorithm needs a parameter capable of measuring the cohesion degree of a class. Since a class is determined by its keywords, it is necessary to identify a measure of cohesion among keywords. This problem has been highly investigated in the past literature on information systems [70]. A possible solution is to associate a similarity coefficient σ_{st} with each pair of keywords (kw_s, kw_t), derived through an appropriate thesaurus such as WordNet [54] and, then, to solve a maximum weight matching problem. This maximizes the average α of the similarity coefficients of the pairs of the class keywords, with the constraint that each keyword can belong to at most one pair. We will not dwell on the formalization and technical details of this solution; the interested reader can find it in [66, 23]. Here, it is sufficient to say that, given a class C_i characterized by a set CS_i of keywords, the average α_i described above is an indicator of the cohesion degree of C_i. α_i belongs to the real interval $[0, 1]$; the higher α_i, the higher the cohesion.

We are now able to describe our (automatic) algorithm for incremental update. It receives a current classification (which consists of a hierarchy of classes and the assignments of the past posts to them) and a new post P_q to be classified and returns the updated classification. First, for each leaf class C_i of the hierarchy, it computes the enhanced Jaccard coefficient J_{iq}^{+} between the sets of keywords of C_i and P_q. After the computation of all the enhanced Jaccard coefficients between P_q and any leaf class of the hierarchy, three cases might happen, namely:

- $J_{iq}^{+} < th_J$ for each leaf class C_i. This means that P_q cannot be assigned to any class. This can happen under two very different circumstances, namely: *(i)* P_q is the first post on a new topic, in which case it is likely that, in the near future, several other posts will contain the same keywords as P_q; *(ii)* P_q is an outlier, i.e., a post totally detached from the others. To deal with both cases, our algorithm adds a new leaf class C_q to the hierarchy. The keywords of C_q will be the ones of P_q. Clearly, P_q is assigned to C_q. At this point, our algorithm activates a counter that increases each time a new post is examined. Before this counter reaches a maximum value c_{max}, if at least another post is assigned to C_q, then the latter is kept in the hierarchy and will gradually grow, giving rise to its ancestors in the hierarchy. On the contrary, if none of the c_{max} posts following P_q is assigned to C_q, then P_q was an outlier, so C_q is removed and P_q remains unclassified.

- $J_{iq}^{+} \geq th_J$ for exactly one leaf class C_i. In this case, P_q is assigned to C_i and all the keywords of P_q not present in C_i are associated with that class. At this point, the cohesion coefficient α_i of C_i is re-computed. If this is less than a certain threshold $th_{\alpha_{min}}$, then we proceed to split C_i into two classes by solving an optimization problem that aims at maximizing the cohesion coefficient of the two classes thus obtained. The two classes have the same parent class, and this class will be the original parent class of C_i. This will result in the potential assignment of new keywords to it, which could lead to a decrease of its cohesion degree. If this were to happen, it would be necessary to split the parent class too. In the worst case, this process may continue until the root of the hierarchy has to be split. Note that this is a first backtracking mechanism present in our algorithm. It solves the problem regarding the existence of an excessively heterogeneous class. This could happen because of an error in the construction of the initial hierarchy or because the objects incrementally assigned to it have made its heterogeneity level greater than the maximum acceptable value.

- $J_{iq}^{+} \geq th_J$ for two or more classes of the hierarchy. In this case, P_q is assigned to all classes for which the above condition is true. Let C_i and $\overline{C_i}$ be the classes having the maximum and submaximum values of the enhanced Jaccard coefficient with P_q, respectively. Our algorithm verifies if C_i and $\overline{C_i}$ continue to be sufficiently distinct or must be merged into a single class. For this purpose, it computes the cohesion coefficients α_i of C_i, $\overline{\alpha_i}$ of $\overline{C_i}$ and α^* of the class C^* that would be obtained by merging C_i and $\overline{C_i}$. If $\alpha^* > \alpha_i$ and $\alpha^* > \overline{\alpha_i}$ then C_i and $\overline{C_i}$ are merged into C^*. This merge process could propagate to the parents of C_i and $\overline{C_i}$ and, gradually, to the ancestors, possibly reaching the root of the hierarchy. For each class which P_q is assigned to, it is necessary to make the check seen in the previous case to verify if that class, after the assignment of P_q to it, is sufficiently cohesive or must be split into two classes. In this last case, the same tasks described for the previous case must be performed. This is a second backtracking mechanism present in our algorithm. It is activated when there are two classes similar to each other that should be merged into a single class. This could happen because of an error in the construction of the hierarchy or because the objects incrementally assigned to the two classes have made them more and more similar to each other.

Once verified in which scenario it falls, our approach proceeds accordingly and obtains a new version of the hierarchy. In Figure 1, we report a flowchart that schematizes the behavior of our approach.

3.1.2 Approach discussion

An important element of the previous algorithm is represented by the two backtracking mechanisms, which allow the correction of possible problems in the hierarchy. In principle, these problems may exist due to construction errors or, more likely, because the incremental assignment of new posts to classes has led to the need of suitably restructuring the initial hierarchy. We observe that, in the literature, there are a few rare cases of a hierarchical classification algorithm provided with backtracking mechanisms. Our approach belongs to this strand. For example, in [92], an approach for hierarchical classification of a set of documents with backtracking is proposed. It assigns a document to one or more categories of a predefined hierarchy. This approach could be applied to Reddit posts as an alternative to ours. However, in our approach, backtracking mechanisms not only allow us to repair a misclassification, as done in [92], but also to modify the hierarchy, if necessary. In our opinion, this last property is important as it allows us to correct not only errors in class assignment but also errors in the hierarchy structure. Moreover, it lets the hierarchy evolve incrementally with the evolution of the posts classified in it.

Our algorithm is very rigorous, as it provides a version of the class hierarchy and post assignments to classes in real time. However, it could be computationally expensive. To reduce its computational costs, we might consider processing a set $PSet$ of posts, instead of just one post, before making any changes to the classes. Clearly the bigger $PSet$, the greater the gain in computational resources, and the greater the information loss caused by not updating classes in real time. A good trade-off we found was setting $PSet$ to the posts on COVID-19 published each day on Reddit.

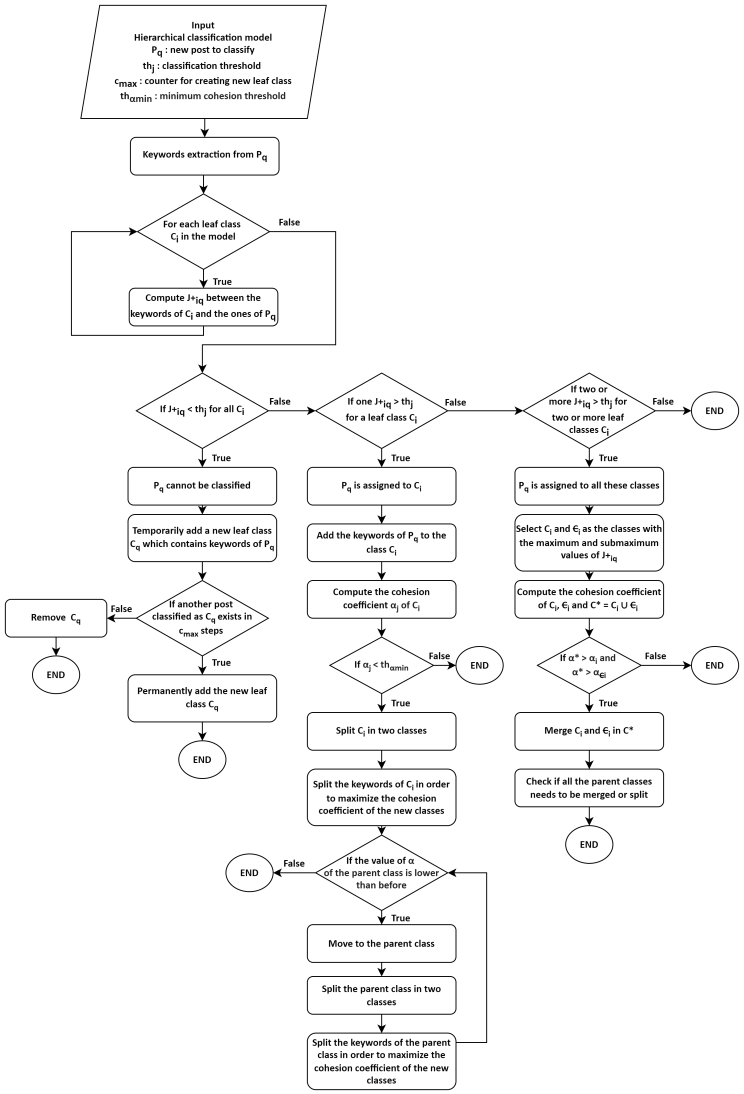

Figure 1: A flowchart representing our approach to classify posts based on topics

3.2 Approach to build virtual subreddits with homogenous topics

3.2.1 Approach description

Network Analysis techniques play a fundamental role in this approach. However, most of the algorithms based on Network Analysis are notoriously expensive, so we had to operate on a sample of available posts, rather than on all of them. Therefore, given a sample S_i, we constructed a suitable network $\mathcal{S}_i$ supporting our approach. In particular:

$$\mathcal{S}_i = \langle N_i, E_i \rangle$$

N_i is the set of nodes of $\mathcal{S}_i$. There is a node n_{i_j} for each post P_{i_j} of S_i. Since there is a biunivocal correspondence between the nodes of N_i and the posts of S_i, in the following of this section, we will use

these two terms interchangeably. E_i represents the set of arcs of $\mathcal{S}_i$. There is one arc $(n_{i_j}, n_{i_k}, w_{jk})$ if there is at least one keyword in common between the posts P_{i_j} and P_{i_k}[7]; w_{jk} denotes the corresponding number of common keywords.

Our approach is parametric with respect to an integer number X. Given the sample S_i, it considers the set KS_i of the X keywords most present in the posts of S_i and builds (at most) X virtual subreddits, $R_1, \cdots, R_X$, one for each keyword. Given the j^{th} keyword $kw_j \in KS_i$, the corresponding virtual subreddit R_j will have associated a set RS_j of keywords (obviously including kw_j) and a set $PostS_j$ of posts. Our approach proceeds as follows:

- For each keyword $kw_j \in KS_i$:
  - It builds the subreddit R_j by initially setting $RS_j = \{kw_j\}$ and $PostS_j = \emptyset$.
  - It builds the set $\overline{KS_j}$ of the X keywords that co-occur most frequently with kw_j in the posts of S_i.
  - It sets $RS_j = RS_j \cup \overline{KS_j}$.
  - For each keyword $kw_{j_h} \in \overline{KS_j}$: *(i)* it builds the set $\overline{KS_{j_h}}$ of the X keywords that co-occur most frequently with kw_{j_h} in the posts of S_i; *(ii)* it sets $RS_j = RS_j \cup \overline{KS_{j_h}}$.

  Note that, once we arrive at the keywords of the set $\overline{KS_{j_h}}$, we do not proceed with finding other keywords that co-occur with them. From the Network Analysis point of view, this means that we stop at the neighbors of the neighbors of kw_j. This practice of stopping at the second separation degree is very common in Network Analysis [85], as well as in the context of the derivation of semantic similarities [66, 23]. It represents an effective answer to the need of having virtual subreddits with homogeneous themes but, at the same time, wide enough to attract many users.

- Now, our approach has identified X homogeneous virtual subreddits $R_1, \cdots, R_X$, one for each keyword of KS_i. However, it could happen that two of these subreddits, say R_k and R_h, are very similar to each other, in the sense that they share most of the associated keywords (and, consequently, of the assigned posts). In this case, it would be better to merge R_k and R_h into a single subreddit R_{kh}. To make this verification and, if necessary, to merge R_k and R_h, our approach proceeds as follows:
  - Let RS_k and RS_h be the set of keywords of R_k and R_h, respectively. It computes the enhanced Jaccard Coefficient J^+_{kh} between RS_k and RS_h.
    - If $J^+_{kh} < th'_J$ then R_k and R_h are not homogeneous enough to be merged[8].
    - If $J^+_{kh} \geq th'_J$ then R_k and R_h must be merged into a single subreddit R_{kh} whose set RS_{kh} of keywords is obtained as $RS_{kh} = RS_k \cup RS_h$.

[7] The identification of common keywords takes synonymies and homonymies into account by following the thesaurus-based approach mentioned in Section 3.1.1.

[8] th'_J is a high threshold in such a way that if $J^+_{kh} \geq th'_J$ then RS_k and RS_h are very similar. For instance, th'_J could be set to $1 - th_J$, where th_J is the same threshold seen in Section 3.1.1.

- At this point, there are at most X virtual subreddits, each with homogeneous topics sufficiently distinct from the ones of the other subreddits. The last step of our approach consists in assigning the corresponding posts to each subreddit. In this regard, we recall that a post can be assigned to more subreddits if its content is compatible with the corresponding keywords. In order to assign posts to subreddits, our approach proceeds as follows:

  - For each virtual subreddit R_k previously built:
    - For each available post P_q:
      - It computes the enhanced Jaccard Coefficient J^+_{kq} between the set RS_k of keywords associated with R_k and the set PS_q of keywords associated with P_q. If $J^+_{kq} > th_J$ then P_q is assigned to R_k.

In Figure 2, we report a flowchart that schematizes the behavior of our approach.

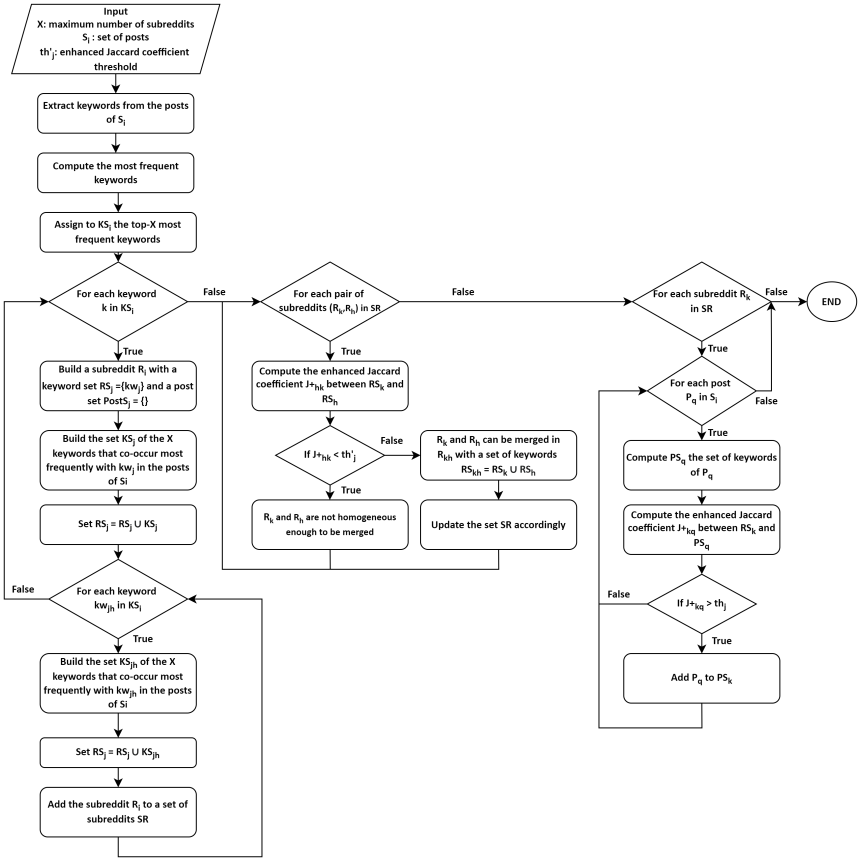

Figure 2: A flowchart representing our approach to build virtual subreddits with homogeneous topics

3.2.2 Approach discussion

The virtual subreddits thus obtained can obviously attract users interested in finding all the posts related to a given topic in one place. Therefore, they can become very attractive not only for current Reddit users but also for new users interested in deepening a certain topic. Indeed, the former would find a new service available, the latter would find the topics of interest in Reddit in a comprehensive way and in a single place, thanks to the presence of the corresponding virtual subreddit.

It is worth pointing out that applying the approach described in Section 3.1.1 to the virtual subreddits returned by the approach described in this section could make them capable of evolving over time. Furthermore, it would be possible to build a classification hierarchy from virtual subreddits, in which these last would represent the corresponding leaf nodes.

We observe that our approach shares several similarities with document/semantic clustering methods [46]. A discussion on these methods can be found in [76]. In this paper, the authors group them into four categories, based on Latent Semantic Analysis, lexical chains, graphs and ontologies, respectively. Our approach shares the most similarities with graph based ones. In [76], six approaches of this family are mentioned. In the following, we give a brief description of each of them highlighting the similarities and differences with our own.

[28] describes a semi-supervised approach for clustering biomedical documents. It uses local information, derived from suitable documents, global information, derived from the MEDLINE collection, and other semantically specific information. Both the approach of [28] and ours operate by making use of keywords in similarity evaluation. However, they have some differences in that: *(i)* the approach of [28] is particularly focused on the biomedical context; *(ii)* it is semi-supervised, while ours is unsupervised; *(iii)* it imposes some constraints on the observations to be clustered. [84] defines an approach to evaluate similarities between documents in different languages. To this end, it represents multilingual documents through the concepts most commonly found in them. The clustering of documents based on concepts proposed in [84] shares some similarities with the clustering of posts based on the keywords of our approach. However, the approach in [84] was designed to analyze complex multilingual documents and to resolve translation ambiguities. Instead, our approach targets generally short texts (i.e., posts) with the goal of clustering them. Therefore, it is less general than the approach of [84] but, being more specific to a given context (i.e., Reddit posts), it can better exploit its features. [73] proposes a new approach for Multilingual Document Clustering using a tensor-based model that can handle the high dimensionality of these documents. Compared to the approach of [73], our own is more tailored to a single goal and, thus, more able to take full advantage of the characteristics of the target context. As an additional difference, the approach of [73] computes document similarities based on phrases, while our approach computes post similarities based on keywords. [72] proposes an approach that classifies a text based on the relationships present in it. To this end, it uses a graphical representation that makes the clusters easier to interpret by contextualizing their terms. In fact, the main goal of this approach is assigning a semantics to clusters. This objective is achieved by associating each cluster with its dominant topic. Both the approach of [72] and ours use keywords of the texts involved as a basis for measuring their similarity. However, they have some differences. In fact, the approach of [72] is complex, having as objective the analysis of the relationships between terms represented through graphs, which are, then, exploited to perform clustering. By contrast,

our approach is tailored to posts, which can be considered very simple documents, but is capable of processing tens of thousands of them. [42] proposes an approach for extracting keywords from a text represented through a graph modeling its terms and their relationships. This approach uses a measure of centrality (e.g., PageRank) to carry out its tasks. Both the approach of [42] and ours are designed to operate in online contexts, characterized by a large number of documents or texts to be analyzed. There are also some differences between them. Indeed, the approach of [42] uses centrality measures, which are complex to compute. Moreover, its main focus is the extraction of keywords from texts rather than the next clustering activity. [33] proposes a document clustering approach based on frequent senses. It searches for frequent subgraphs that reflect the frequent senses of a sentence. The subgraphs thus discovered are used to generate document clusters. The main difference between the approach of [33] and ours is that the former represents a sense by means of a subgraph, while the latter represents a post by means of keywords. Operating on graphs instead of on keywords takes much more time and is well suited to classify a limited number of complex documents. Instead, it is hardly applicable to our context, where there are simple, but very numerous, posts to be clustered.

In [29], the authors propose another survey for clustering semantic documents. In the following, we present the approaches described therein that shares the most similarities with our approach and, for each of them, we highlight the similarities and differences with ours. [5] proposes a clustering approach to distinguish relevant information from irrelevant one in a document. Both this approach and ours are designed to operate with many data. The main differences between them are that the approach of [5] uses ontologies and was primarily conceived for the medical field, where well defined ontologies already exist. Instead, our approach can be applied on posts about any topic, even those for which well-defined ontologies do not exist. [6] proposes a clustering approach based on frequent concepts, rather than frequent keywords. These concepts are derived from the documents through a pre-processing activity. The approach of [6] is very accurate but is suitable for a context where the number of documents to be classified is limited, which is very different from our reference context. [78] proposes an approach to classify documents based on the terms present in them and the corresponding lexical relationships. To this end, it associates a tag with each document and enriches its representation through a bag of words. Both this approach and ours are based on keywords and consider the lexical relationships involving them (in our approach this is done by using the operator J^+ instead of the operator J). The main difference between them is that the approach of [78] is designed for clustering a limited number of complex documents.

3.3 Approach to build virtual communities of users with homogeneous interests

3.3.1 Approach description

Our approach to build virtual communities of users having homogeneous interests is based on Network Analysis too. Therefore, also in this case, we use a support social network. Specifically, given a sample S_i, we construct a social network $\mathcal{S}'_i$:

$$\mathcal{S}'_i = \langle N'_i, E'_i \rangle$$

N'_i is the set of nodes of $\mathcal{S}'_i$. There is a node n_{i_j} for each author A_{i_j} who submitted at least one post of S_i. Since there is a biunivocal correspondence between the nodes of N'_i and the authors of the

posts of S_i, we will use these two terms interchangeably in this section. E'_i represents the set of arcs of $\mathcal{S}'_i$. There is an arc $(n_{i_j}, n_{i_k}, w_{jk})$ if the authors A_{i_j} and A_{i_k} used the same keyword in at least one post of S_i published by them. The weight w_{jk} of the arc indicates the number of keywords used by both A_{i_j} and A_{i_k} in some of their posts of S_i. Again, we took synonymies and homonymies between keywords into account using the same guidelines seen in Sections 3.1 and 3.2.

A first issue to address in the definition of our approach is to find a rule allowing us to identify bots (i.e., automatic Reddit users that posted news crawled from different sources). For this purpose, we analyzed the behavior of bots in Reddit and observed that they generally had a high number of keywords associated with them. Therefore, we decided to consider as bots all those authors who had more that B keywords associated with them. We carried out some tests to identify the optimal value of B and found that it is equal to 8.

Knowing the number of keywords in each arc is an important starting point to reach our goal. However, it is not sufficient. Actually, it is necessary to go in more detail considering the specific sets of keywords associated with network arcs. As a matter of fact, going to this level of detail, we observed that some sets of keywords were repeated in many arcs. This fact is important because it represents the key to construct our virtual communities of users with homogeneous interests [69]. In fact, in principle, all the nodes connected by arcs having the same set of keywords could be regarded as a community of users sharing the same set of interests.

Starting from this reasoning, our approach operates as follows:

- It identifies all the sets of keywords associated with the network arcs.
- It removes the sets of keywords consisting of less than three elements, because we considered them insignificant as indicators of common interests for a community of users.
- It removes the sets of keywords occurring less than three times because we believe that, with such a low number of occurrences, the coincidence of interests between authors expressed by them could be incidental.
- It computes the distribution of the remaining sets of keywords against the number of occurrences.
- It selects all the sets of keywords belonging to the first quartile of the distribution determined in the previous step. For each of these sets, it constructs the subnetwork consisting of only the arcs belonging to it. The nodes of this subnetwork represent a community of users with homogeneous interests defined by the keywords of the set.

In Figure 3, we report a flowchart that schematizes the behavior of our approach.

3.3.2 Approach discussion

Each subnetwork represents an output of our approach and, therefore, a virtual community of users with homogeneous interests. The virtual communities thus obtained can be useful to create a collaborative filtering recommender system aiming at suggesting to a user other ones with similar interests. Moreover, our approach could be adopted by Reddit itself to propose a new functionality aiming at

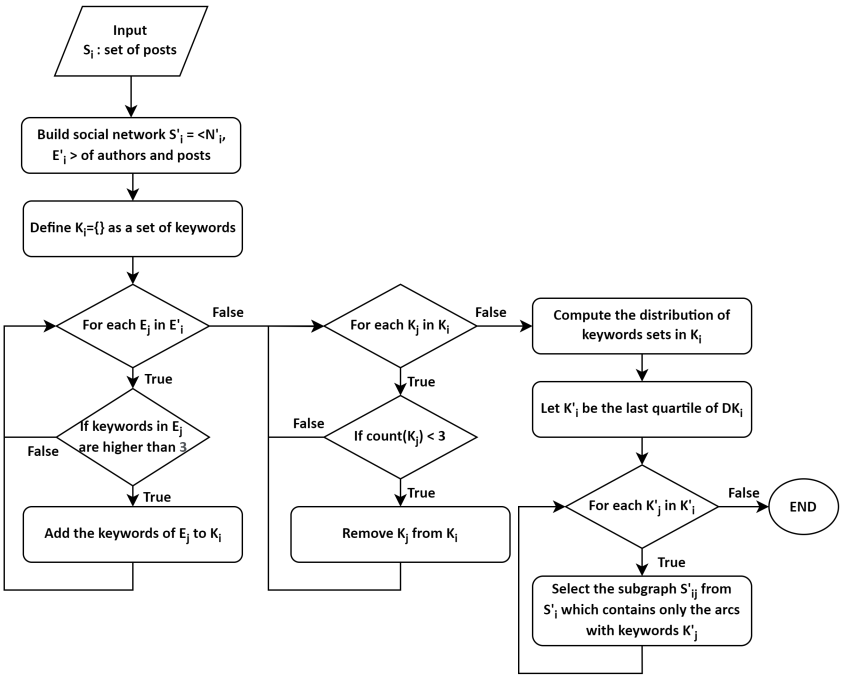

Figure 3: A flowchart representing our approach to build virtual communities of users with homogeneous interests

creating communities of users with common interests [64]. Again, we note that applying the approach described in Section 3.1.1 to the virtual communities of users returned by this approach could make returned communities able to evolve over time. Moreover, also in this case, it would be possible to build a classification hierarchy from the virtual communities. These last would represent the leaf nodes of the hierarchy.

The approach described here shares several similarities with the approaches to cluster a node-attributed network or a semantic document network.

A survey on community detection methods in node-attributed social networks can be found in [19]. Among the approaches described in this survey, the ones closest to ours are those presented in [2] and [11]. In [2], the authors propose a graph embedding approach to cluster content-enriched graphs. The idea behind this approach is to embed each node of a graph in a continuous vector space, in which structural and attributive information located at the vertices can be encoded into a unified latent representation. Analogously to our approach, the one of [2] considers the graph structure during clustering activities. In [11], the authors propose a community detection and characterization algorithm that includes the contextual attribute information of graph nodes. Its goal is to compute the context of communities and discover new ones. For this purpose, it uses a coordinate-based algorithm that updates the community label assignment of nodes. Analogously to our approach, it uses the Jaccard coefficient to evaluate the context of a node. The way the approaches of [2] and [11] operate allows them to achieve high accuracies. However, they are heavy for a context like ours characterized

by very simple graphs but with a huge number of nodes and arcs. From this point of view, our approach, which considers only the structure of the graph and very little other information, is lighter and is able to process even graphs with tens of thousands of nodes, which are those of interest for our context.

In [13], the authors propose a survey on approaches to clustering the nodes of a graph with attributes. Both the approaches described in [13] and ours focus on finding homogeneous communities within the network. However, there are important differences between them. Indeed, the approaches described in [13] handle multi-dimensional graphs whose nodes and arcs can have attributes. This makes these approaches particularly suitable in handling very complex contexts where they prove to be very accurate. However, the processing times required by them are high; so, they cannot be applied in presence of large networks, such as those characterizing our scenario.

4 Experiments

In this section, we illustrate the experiments we performed. In particular, in Section 4.1, we describe our dataset and the Exploratory Data Analysis we carried out on it. Then, in Sections 4.2 - 4.4, we present the experiments we conducted to evaluate our three approaches.

4.1 Dataset description

The dataset we used in the activities described in this paper was derived from the `pushshift.io` website, which is one of the main data repositories related to Reddit content. Specifically, `pushshift.io` collects Reddit posts and comments and provides a suitable website and an API for accessing them. It simplifies the query process of historical Reddit data. Furthermore, it provides several features, like a full-text search on comments and submissions. Overall, it stores all the posts and comments published on Reddit from June 2005 to today [10]. Leveraging the API provided by it, we downloaded all the posts published in Reddit from January 9^{th}, 2020 to April 30^{th}, 2020. Then, we stored them in a `.csv` file. Afterwards, we performed a set of cleaning operations, aimed to obtain a dataset ready for our analyses. Specifically, we maintained all the posts whose title contained the words "covid" and/or "coronavirus". Then, we deleted the posts consisting only of images and videos. Finally, among the remaining posts, we selected only the ones having a title written in English. To identify them, we leveraged the English corpus available in the `nltk` (i.e., Natural Language Toolkit[9]) library of Python. Specifically, we iterated over each lemma of the post title and verified if it was present in the corpus. If all the lemmas of the post title satisfied this condition, we considered the corresponding title as written in English and added it to our dataset. This last task aimed to avoid working with a multi-language dataset, which was out of our scope. At the end of these cleaning operations, our dataset consisted of 2,498,768 posts. For each post we considered the following features:

- `id`: the post's identifier;
- `author`: the post's author;

[9] `https://www.nltk.org/`

- `title`: the post's title;
- `created`: the date the post was created;
- `subreddit`: the subreddit where the post was published;
- `num_comments`: the number of comments received by the post;
- `num_crossposts`: the number of times the post was crossposted;
- `score`: the score of the post (equal to the number of upvotes minus the number of downvotes);
- `upvote_ratio`: the ratio of upvotes to the total number of votes.

The number of subreddits involved is 70,280 while the number of authors is 567,914. We note that the average number of authors per subreddit and the average number of posts per author are low, in that they are equal to 8.08 and 4.40, respectively. The average number of posts per subreddit is 35.55.

We performed our analyses on a server equipped with 16 Intel Xeon E5520 CPUs and 96 GB RAM. We used Ubuntu 18.04.3 as operating system. Moreover, we chose Python 3.6 as programming language, its Pandas Library to carry out ETL (i.e., Extraction, Transformation and Loading) tasks and its NetworkX library to perform network-based operations.

4.1.1 Exploratory Data Analysis

Before carrying out our tests on the three approaches proposed in this paper, we performed an Exploratory Data Analysis (EDA, for short) on our dataset. To this end, we carried out the following tasks:

- Analysis of the distributions of `created`, `subreddit`, `num_comments`, `num_crossposts`, `score` and `upvote_ratio`.
- Analysis of the possible outliers and management of the possible missing values on all the features.
- Analysis of the possible correlations between the features.
- Detection of interesting patterns and models.

In the following of this section, we describe each of these tasks.

Analysis of feature distributions

In Figures 4 - 6, we report the distributions of `created`, `subreddit`, `num_comments`, `num_crossposts`, `score` and `upvote_ratio`. A first analysis of them highlights that the distribution of posts over time is irregular with the presence of two peaks. The first of them is at the end of January, the period when the COVID-19 epidemic reached its peak in China, South Korea, and other Asian countries. The second peak, much higher than the first, is around Mid-March, when the virus began to spread enormously in Europe. The distributions of posts against subreddits, comments, crossposts, score and upvote ratio follow power laws.

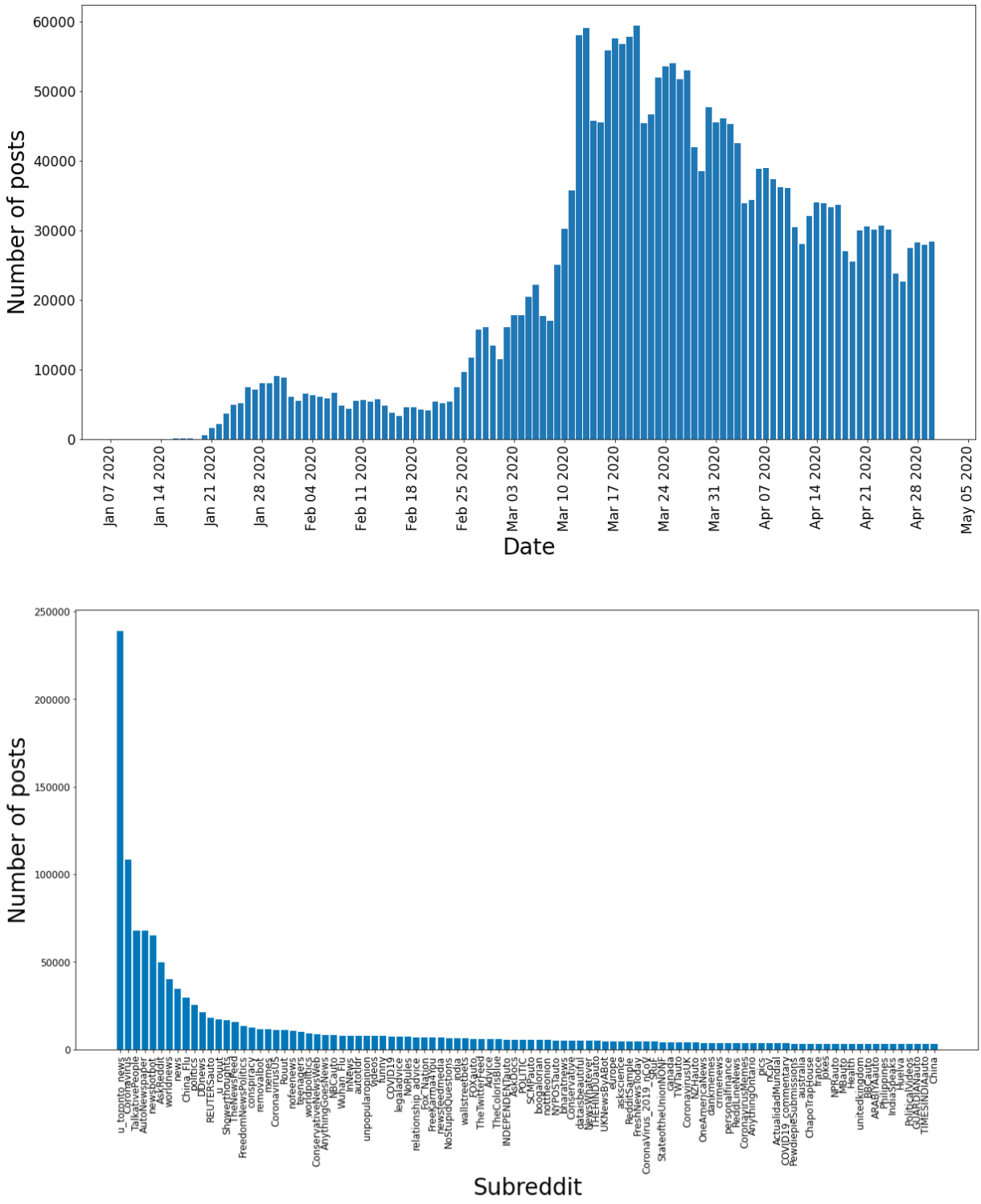

Figure 4: Distribution of the features `created` (normal scale) and `subreddit` (normal scale)

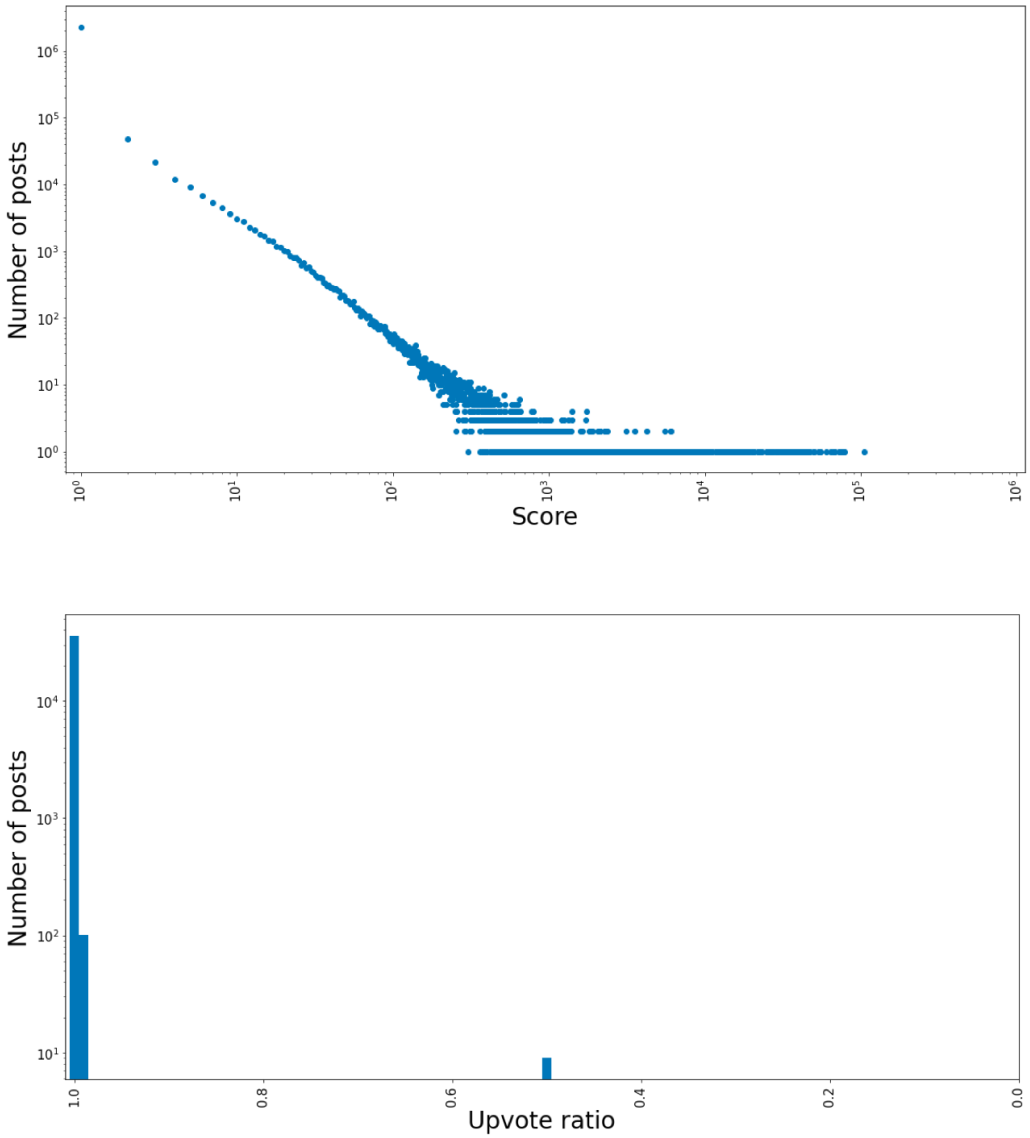

Figure 5: Distribution of the features `num_comments` (log-log scale) and `num_crossposts` (log-log scale)

Analysis of possible outliers and management of missing values

As we mentioned previously, our dataset was downloaded from `pushshift.io`. Reddit data undergoes ETL activities before being stored in that repository. As a result, there are no missing or incorrect values (e.g., a negative number of comments) in `pushshift.io` and, therefore, in our dataset. Any other value assumed by one of the features of our interest (e.g., a very high value of the number of comments) cannot be considered in principle as an outlier, given the power law distribution characterizing them.

Analysis of the possible correlations between the features of the dataset

In Figure 7, we report the correlation matrix of the features of our dataset. This matrix has a row

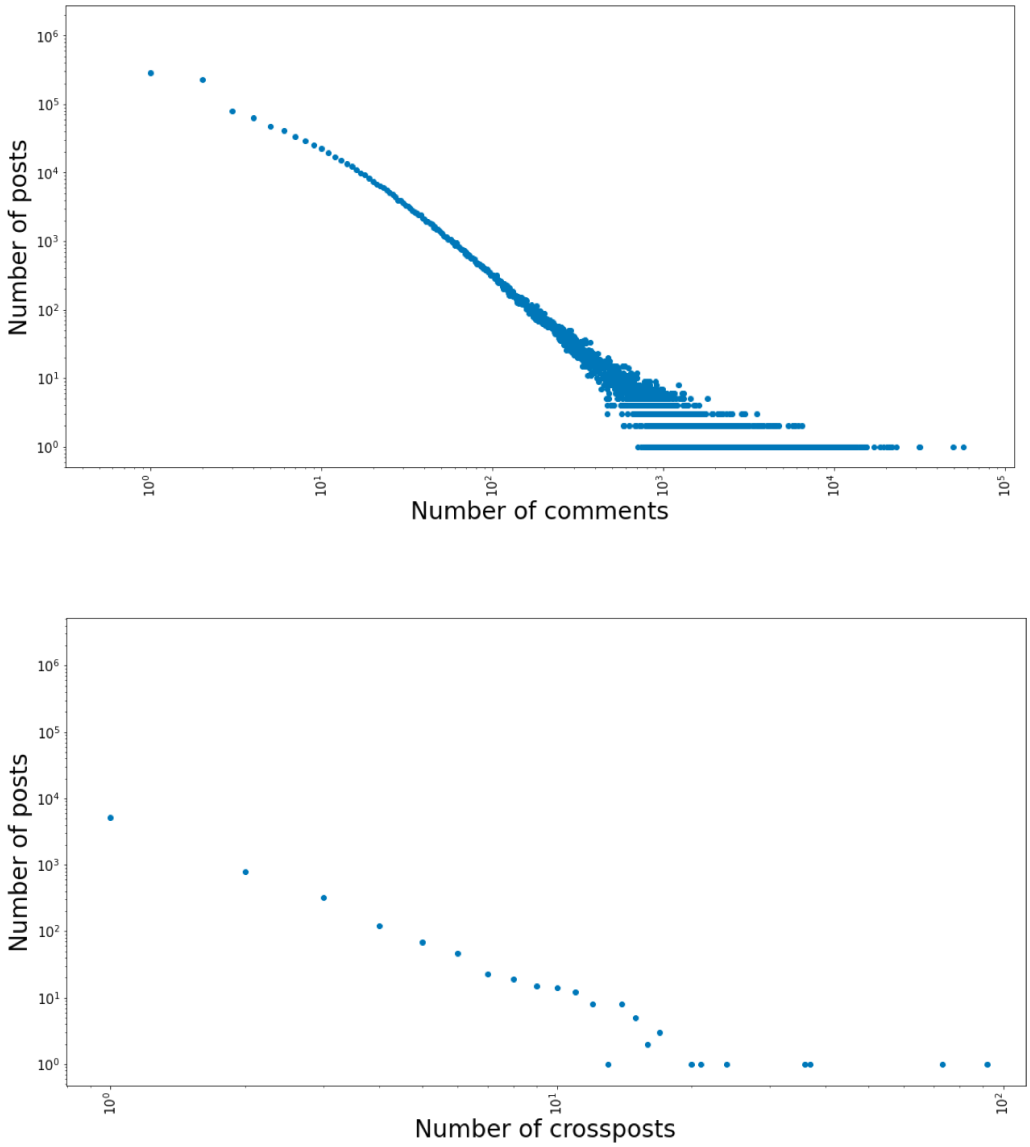

Figure 6: Distribution of the features `score` (log-log scale) and `upvote_ratio` (semi-log scale)

and a column for each feature. Its generic element $[i, j]$ denotes the value of the Pearson correlation between the features associated with the i^{th} row and the j^{th} column. We recall that the Pearson correlation coefficient is a parameter whose values range in the real $[-1, 1]$. When it is 1 there is a strong direct correlation; when it is -1 there is a strong inverse correlation; when it is 0 there is no correlation. From the analysis of this matrix, we can see that there is a certain correlation between `score` and `num_crossposts` and between `score` and `upvote_ratio`. The latter was expected because the percentage of upvotes influence the score of a post. Instead, the former is an unexpected information derived thanks to our analysis.

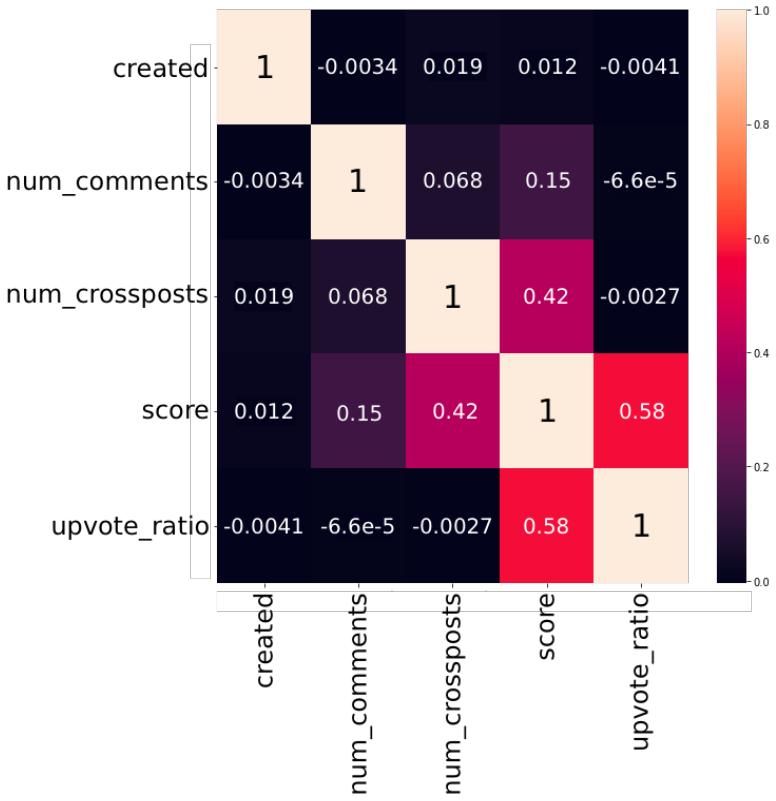

Figure 7: Correlation matrix of the features of our dataset

Detection of interesting patterns and models from the dataset

As a final Exploratory Data Analysis task on our dataset, we performed a search for patterns and models that might be useful both for understanding the data available and for the next experiments.

First of all we computed the distribution of authors against posts. It is shown in Figure 8. This figure suggests us that it follows a power law. Recall that a quantity is said to follow a power law when the probability of measuring a particular value of it varies inversely as a power of that value. This distribution is also known as Zipf's law or Pareto Distribution [63]. It can be characterized through two parameters, namely: α, which represents the steepness of the curve, and δ, which denotes the smoothness of the slope change. In this specific case, the presence of the power law distribution means that very few authors submit a very high number of posts, while most authors submit a very little number of posts. In order to quantitatively confirm that the distribution of Figure 8 follows a power law, we ran two different Kolmogorov-Smirnov tests on it. The first one was based on the null hypothesis H_{01} = "The distribution is log-normal", the second one on the null hypothesis H_{02} = "The distribution is power law". We found that, given 567,914 observations, the critical value $D_{crit} = 0.025$. The first test returned a statistic $D_1 = 0.42$, with a p-value = 0.31, which led us to reject H_{01}. The second test returned a statistic $D_2 = 0.023$, with a p-value = 0.018, which confirmed H_{02}. In conclusion, we could say that our distribution follows a power law, in particular a Type 1 power law. Then, we computed its α and δ parameters and obtained that $\alpha = 2.1157$ and $\delta = 0.0201$.

By operating in the same way, we also computed: *(i)* the distribution of posts against subreddits;

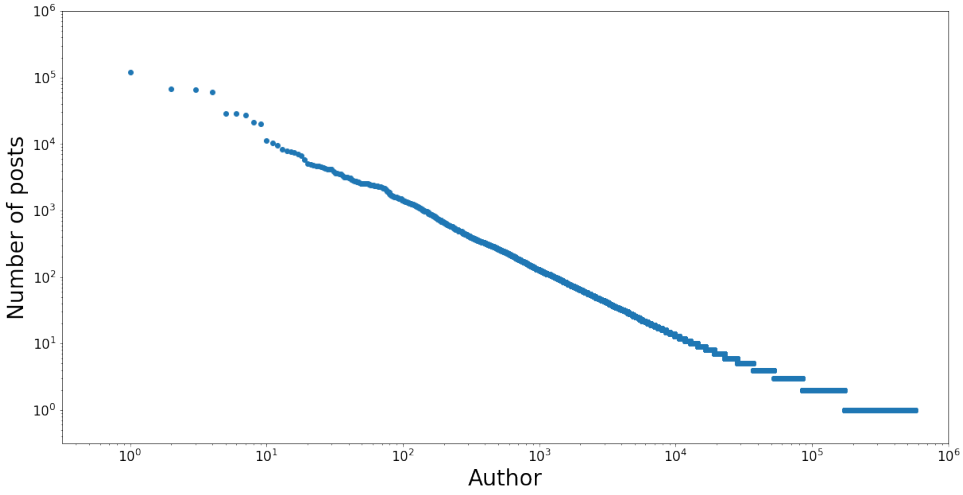

Figure 8: Distribution of authors against posts (log-log scale)

(ii) the distribution of authors against subreddits; *(iii)* the distribution of posts against score; *(iv)* the distribution of comments against posts; *(v)* the distribution of crossposts against posts. We verified that all these distributions follow a power law.

The fact that the distributions of posts against score, number of comments and number of crossposts follow a power law could lead us to be pessimistic about the overall quality of published posts. Actually, this is not necessarily the case, because the power law distribution is the most common one in social networks [85]. Therefore, we decided to perform a further verification by computing the fraction of posts with an `upvote_ratio` less than 1. We saw that only 110 posts of the 2,498,768 examined ones (i.e., the 0.00044% of them) have an `upvote_ratio` less than 1. This confirms the validity of our conjecture that we should not be pessimistic about the results on posts previously obtained. All in all, the vast majority of the posts on COVID-19 were appreciated by the Reddit community.

So far we have considered four indicators of post quality and we have seen that three of them follow a power law, while the fourth one is almost always positive. We found it very interesting to check if the posts with the highest values for each of the four indicators were always the same or not. For this reason, we selected the top 500 posts for each quality parameter and computed their intersection. We could see that it contained only 13 posts. This result is very important because it tells us that there are no absolute best posts; instead, the various quality parameters capture different aspects. The only intersection worthy of attention regarded the top 500 posts with the highest score and the top 500 posts with the highest number of crossposts. In this case, we obtained that the intersection included 158 posts. This is not surprising because Figure 7 shows that there is a fairly high correlation between these two features.

After carrying out structural analysis, our attention focused on content. To this end, we considered the titles of the posts and carried out a lemmatization activity on them, removing stop words and punctuation marks. After these tasks, we computed the number of occurrences for each keyword. In Figure 9, we report the most frequent keywords along with the corresponding number of occurrences.

As we can see from this figure, there is a keyword (i.e., "Coronavirus") that is by far the most

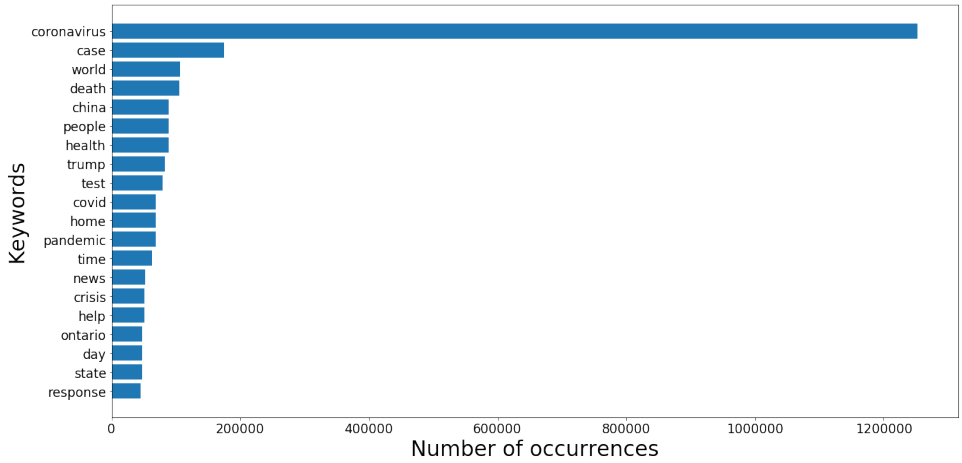

Figure 9: Most frequent keywords in post titles and corresponding number of occurrences

frequent one. Actually, this information is quite obvious, and therefore not very significant. Instead, if we consider all the other keywords in the same figure, we can observe that most of them are characterized by a comparable and high number of occurrences. This reveals that COVID-19 is dealt within Reddit from various points of view, from health to economy, from politics to technology, and so on. This property can represent an empirical justification of the classification approach described in Section 3.1.

Finally, as a last task, we carried out the clustering of the keywords described above. First, we trained a FastText [36, 12] word embedding model in order to have a 100-dimensional vector representation of the keywords. Then, we used the elbow method to identify the recommended number of clusters and found that this number is equal to 5. In order to observe clusters in the bi-dimensional plane, we computed the Principal Component Analysis [86] of the word embedding vectors. We report the resulting scatter plot in Figure 10. This figure is interesting because it reveals how keywords can be grouped in very homogeneous clusters. This property can represent an empirical justification of the approaches illustrated in Sections 3.2 and 3.3.

4.2 Approach to classify posts based on topics

The first step of our experimental campaign for evaluating this approach was the construction of the initial classification. For this purpose, we used all posts in our dataset from January 9^{th}, 2020 to March 31^{st}, 2020. To perform this classification, we required the support of a human expert. She was a sociologist who has been working in the field of Social Network Analysis for more than 10 years. She has been following the dynamics of information diffusion on Reddit for more than 6 years and on other popular online social networks (in particular, Facebook and Twitter) since the beginning of her work. The sociologist was supported by an epidemiologist, in interpreting technical medical terms found in some posts. We were very careful in selecting the human expert and her consultant epidemiologist, because we were aware that their decisions were very important since they would represent the ground truth in the evaluation of our approach. The initial classification is shown in Figure 11.

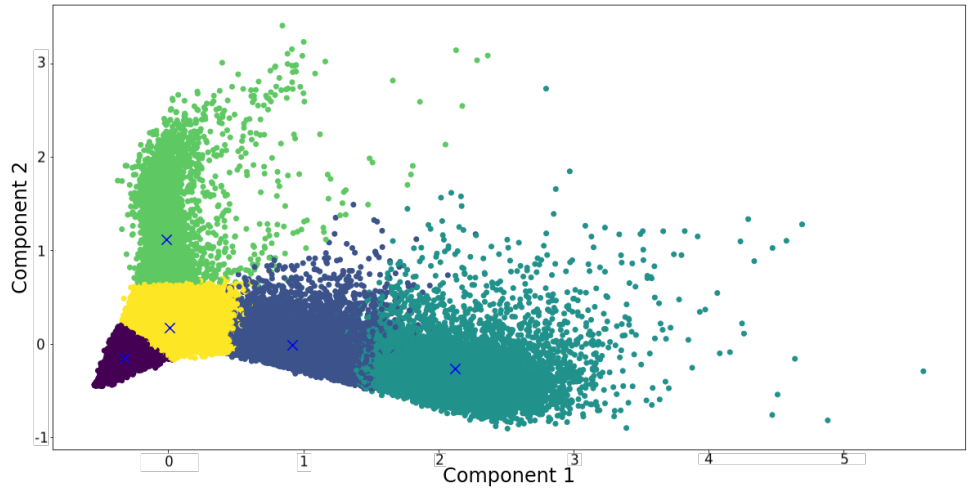

Figure 10: Clustering of the keywords derived from post titles

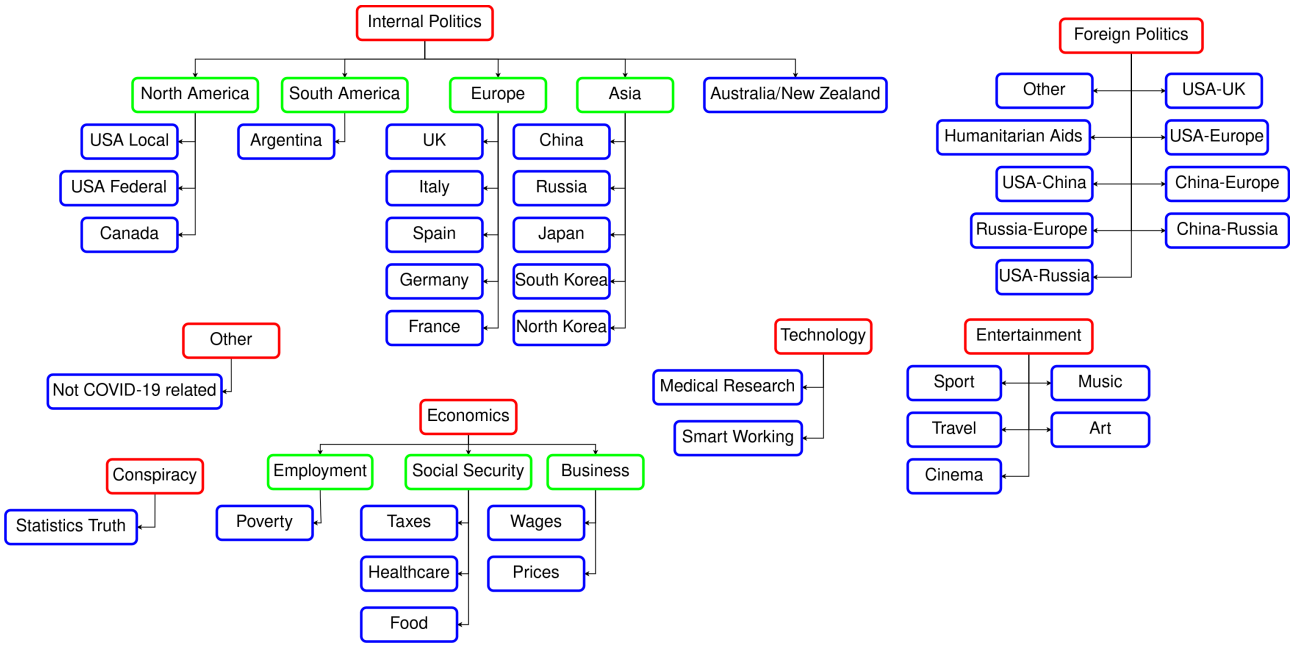

Figure 11: The initial classification for the posts on COVID-19 in Reddit

With regard to it, we have the following parameter values: *(i)* number of posts available: 1,745,073; *(ii)* number of leaf classes: 40; *(iii)* number of posts assigned to at least one class: 1,605,347 (equal to 91.99% of all the posts available); *(iv)* average number of keywords associated with the leaf classes: 11.45; *(v)* average number of classes a post was assigned to: 4.07.

After this initial classification, we provided our algorithm with the posts on COVID-19 published in Reddit in April 2020. We carried out a session of the algorithm for each day of April. During each session, we gave in input the classification of the previous day and the set of posts on COVID-19 published in the current day. The classification obtained at the end of April is shown in Figure 12.

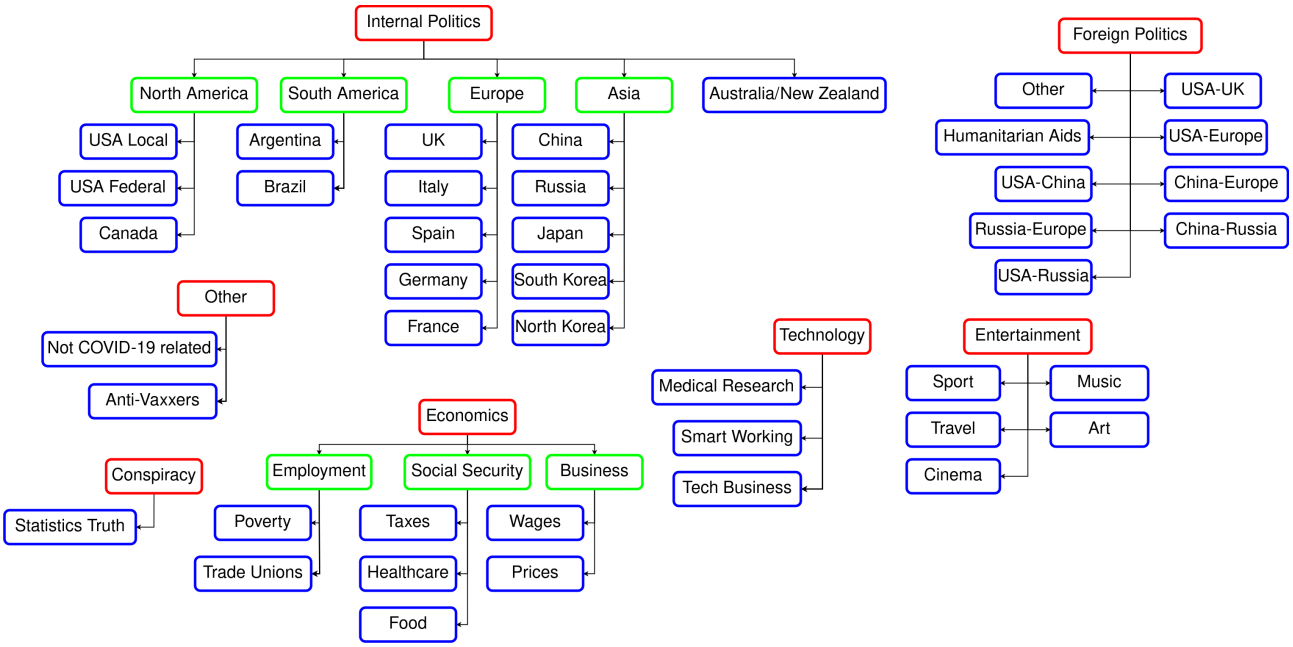

Figure 12: The final classification for the posts on COVID-19 in Reddit

With regard to this final classification, we have the following parameter values: *(i)* number of posts available: 2,498,768; *(ii)* number of leaf classes: 43; *(iii)* number of posts assigned to at least one class: 2,396,744 (equal to 95.92% of all the posts available); *(iv)* average number of keywords associated with the leaf classes: 11.25; *(v)* average number of classes a post was assigned to: 4.26.

To evaluate the quality of the classification returned by the proposed algorithm we adopted the classic parameters employed in these cases, namely Precision, Recall and F-Measure [34]. In order to carry out these measurements, we used the decisions of the human expert as the ground truth. However, the processing capabilities of the human expert are limited, so it was not possible to operate on all the posts available, but only on a subset of them. Therefore, we randomly selected two samples, S_1 and S_2, each containing 500 posts of the initial classification. We also considered two samples, S_3 and S_4, each containing 500 posts of the final classification.

Given the sample S_h, $1 \leq h \leq 4$, the Precision denotes how many of the post assignments to classes made by our approach were also made by the human expert. The Recall indicates how many of the post assignments to classes made by the human expert were also made by our approach. The F-Measure is the harmonic mean of Precision and Recall. The values of Precision, Recall and F-Measure for the four samples under consideration are shown in Table 1.

The analysis of this table reveals that:

- Our approach returns very accurate results, with both the initial and the incrementally updated classifications. Regarding these results, we observe that the values obtained using our approach are very high compared to those generally obtained when content mining techniques are adopted. In our opinion, this is caused by two reasons. The first is that our approach is not completely automatic because the leaf classes of the initial hierarchy are determined with the support of

Sample	*Parameter*	*Value*
S_1	Precision	0.92
	Recall	0.86
	F-Measure	0.89
S_2	Precision	0.94
	Recall	0.85
	F-Measure	0.89
S_3	Precision	0.97
	Recall	0.94
	F-Measure	0.95
S_4	Precision	0.96
	Recall	0.97
	F-Measure	0.96

Table 1: Precision, Recall and F-Measure for the four samples under consideration

the human expert who evaluates, and possibly corrects, the results produced by the text mining algorithm. While the presence of the human expert has a negative impact on timing, there is no doubt that it can have a very positive impact on accuracy. The second reason is that, since the posts used for training were published from January 9^{th}, 2020 to March 31^{st}, 2020, while those used for testing were published in April 2020, it is plausible that there is a strong similarity between the training and testing data.

- The results obtained are stable because, if we take two different samples for each classification, they change very little. More specifically, if we consider the two samples S_1 and S_2, both derived from the initial classification, we have that: *(i)* Precision is always very high, above 0.90; its variation occurring when switching from S_1 to S_2 is 2.18%. *(ii)* Recall is always high, above 0.80; its variation occurring when switching from S_1 to S_2 is 1.17%. *(iii)* F-Measure is always high, equal to 0.89, and does not change when switching from S_1 to S_2.

  We now consider the samples S_3 and S_4 both derived from the final classification. We have that the variation of Precision (resp., Recall, F-Measure) occurring when switching from S_3 to S_4 is 1.03% (resp., 3.09%, 1.04%).

  As we can see, when we switch from S_1 to S_2 or from S_3 to S_4, the variations in the values of all parameters are negligible.

- Incremental updates allow our approach to obtain even more accurate results, especially for Recall. This last fact is not surprising because updates are made on the basis of the posts published. Indeed, if we compare the values of the parameters before and after classification, we can see that they always show an improvement. In particular: *(i)* Precision increases by 3.76%, passing from an average value of 0.930 to an average value of 0.965; *(ii)* Recall increases by 11.70%, passing from an average value of 0.855 to an average value of 0.955; *(iii)* F-Measure increases by 7.30%, passing from an average value of 0.890 to an average value of 0.955.

Everything we have seen in this section allows us to conclude that our approach is really capable of classifying posts related to COVID-19 and of keeping this classification updated.

4.3 Approach to build virtual subreddits with homogeneous topics

Analogously to what we performed for the experiments related to the previous approach, we decided to select two samples randomly, in order to verify whether the results we will obtain are stable. In particular, we considered two samples, S_1 and S_2, each including 52,352 randomly selected posts. Their main characteristics are reported in Table 2. Figures 13 and 14 illustrate the distribution of posts against authors and comments, whereas Figure 15 reports the trend of the number of posts over time. As we can see, despite the total randomness they were built with, the differences between the two samples are very low. Therefore, it was reasonable assuming that the results we obtained from them would have been stable. In any case, we did not trust this hypothesis alone but, for each result obtained, we made the appropriate stability check to see if it was very similar in the two samples.

Parameter	*Value in* S_1	*Value in* S_2
Number of posts	52,352	52,352
Number of authors	23,874	23,807
Number of subreddits	7,820	7,825
Timestamp of the first post	2020-01-09 05:35:31	2020-01-09 04:59:13
Timestamp of the last post	2020-04-30 23:59:55	2020-04-30 23:58:25
Average number of comments per post	9.210	9.702
Average score of posts	5.168	4.401
Average number of keywords per post	3.031	3.053

Table 2: Main characteristics of the two samples S_1 and S_2

After building the two networks, we computed some basic parameters of them. These are shown in Table 3.

Parameter	*Value in* $\mathcal{S}_1$	*Value in* $\mathcal{S}_2$
Number of nodes	52,352	52,352
Number of arcs	29,498,151	29,332,207
Density	0.0215	0.0214
Average clustering coefficient	0.702	0.699
Average weight of arcs	1.035	1.035

Table 3: Some basic parameters of the networks $\mathcal{S}_1$ and $\mathcal{S}_2$

The analysis of the values of these basic parameters provides us with valuable information. In fact, we can see that the density of $\mathcal{S}_1$ and $\mathcal{S}_2$ is low, while the corresponding average clustering coefficient is high. This kind of configuration for these two parameters is not very common in Network Analysis. In fact, usually, both of them are low or both are high. Instead, in this case, the presence of a low density indicates that each post shares keywords with only few other ones. This can be justified considering that the topics covered in the COVID-19 posts are various, as we have seen in Section 4.2. The presence of a high clustering coefficient is an indicator of closed triads [85]. This implies that, if the post P_{i_j} shares keywords with the post P_{i_k}, and P_{i_k} shares keywords with the post P_{i_h}, then P_{i_j} and P_{i_h} will also share keywords [62, 25]. This suggests that, actually, there may be groups of keywords in common among a "cluster" of posts. These keywords are exactly the reference point for the construction of virtual subreddits with homogeneous themes. In fact, the cluster of posts with the keywords in common represents the core of the virtual subreddit.

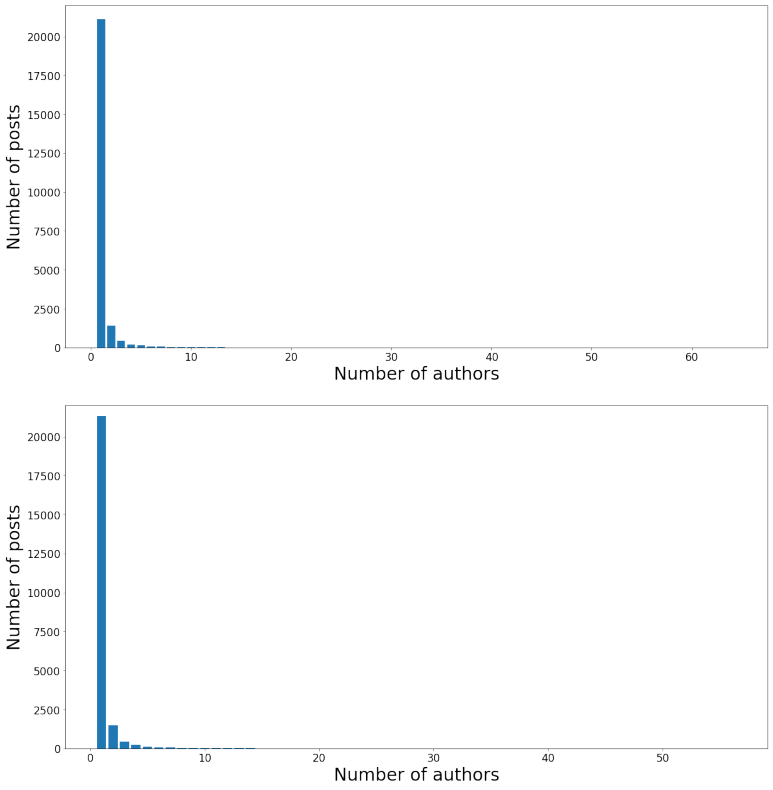

Figure 13: Distribution of posts against authors for S_1 (on top) and S_2 (on bottom)

As a starting point in the definition of our approach, we determined the distribution of the keywords in the posts of the samples S_1 and S_2[10]. Indeed, our general idea is to use each of the most common keywords in the sample as an aggregation point for attracting new homogeneous keywords, together with the corresponding posts where they are present.

We applied our approach to the two samples S_1 and S_2 presented at the beginning of this section. We set $X = 10$ because, due to the steep distribution followed by the keywords characterizing the posts of the samples, the first 10 keywords already "cover" 73.33 % of the posts of S_1 and 73.03 % of the posts of S_2. The 10 keywords identified for S_1 and S_2, sorted by the number of posts in which they occur, are shown in Table 4.

Finally, Tables 5 and 6 report the subreddits derived from these keywords. For each subreddit, they report the set of the corresponding keywords and the number of posts assigned to it. Observe that, in Table 5, the subreddits R_1 and R_6 were found very similar and, according to the rules of our approach, were merged into a unique subreddit $R_{1,6}$.

We observe that many of the keywords present in Tables 5 and 6 belong to two or more virtual subreddits. In other words, each virtual subreddit in one of these tables shares keywords with one or more of the other virtual subreddits. This is due to the high clustering coefficient characterizing

[10] Also in the computation of this distribution we removed the word "Coronavirus" (for the reasons discussed in Section 4.1.1) and took the synonymies and homonymies into account.

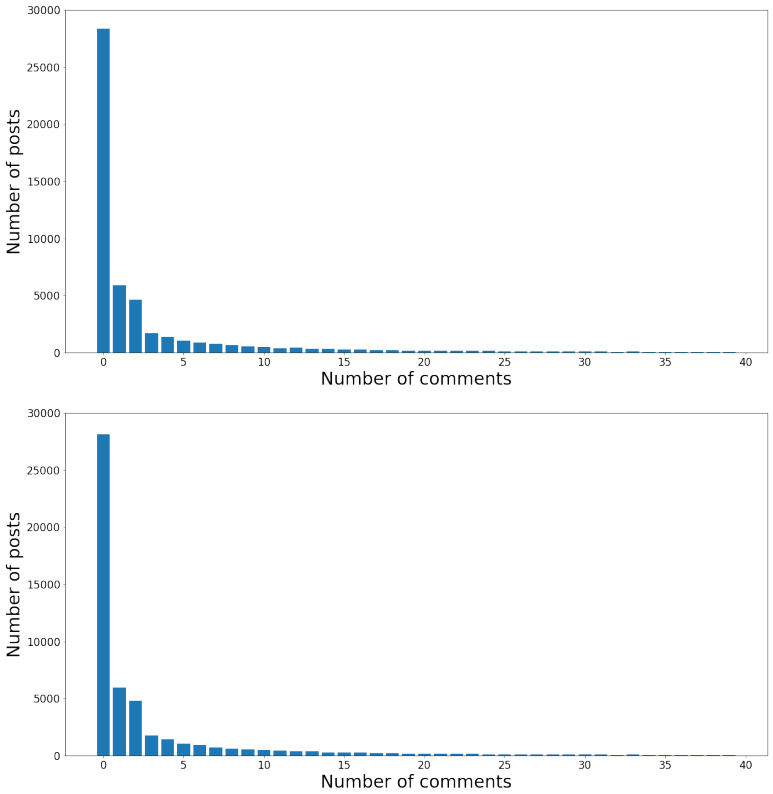

Figure 14: Distribution of posts against comments for S_1 (on top) and S_2 (on bottom)

S_1	S_2
case (6073)	case (5910)
world (5432)	world (5514)
people (4836)	health (4862)
trump (4793)	death (4811)
health (4774)	people (4784)
death (4750)	trump (4752)
china (4525)	china (4488)
ontario (4491)	ontario (4451)
test (4422)	test (4348)
home (4296)	home (4312)

Table 4: The 10 keywords identified for S_1 and S_2

the networks $\mathcal{S}_1$ and $\mathcal{S}_2$ and discussed in our comments to Table 3. In that part, we pointed out that a high clustering coefficient implies that posts tend to be connected forming closed triads and that the posts in a triad share groups of common keywords. All this is reflected by the fact that virtual subreddits, which are ultimately sets of posts, share several keywords with each other. This is further amplified by the fact that our approach allows a post to belong to multiple virtual subreddits. Each virtual subreddit obtained through our approach often differs from the others not so much for the exclusivity of topics or posts but for the greater or smaller emphasis that it assigns to one or more topics with respect to others.

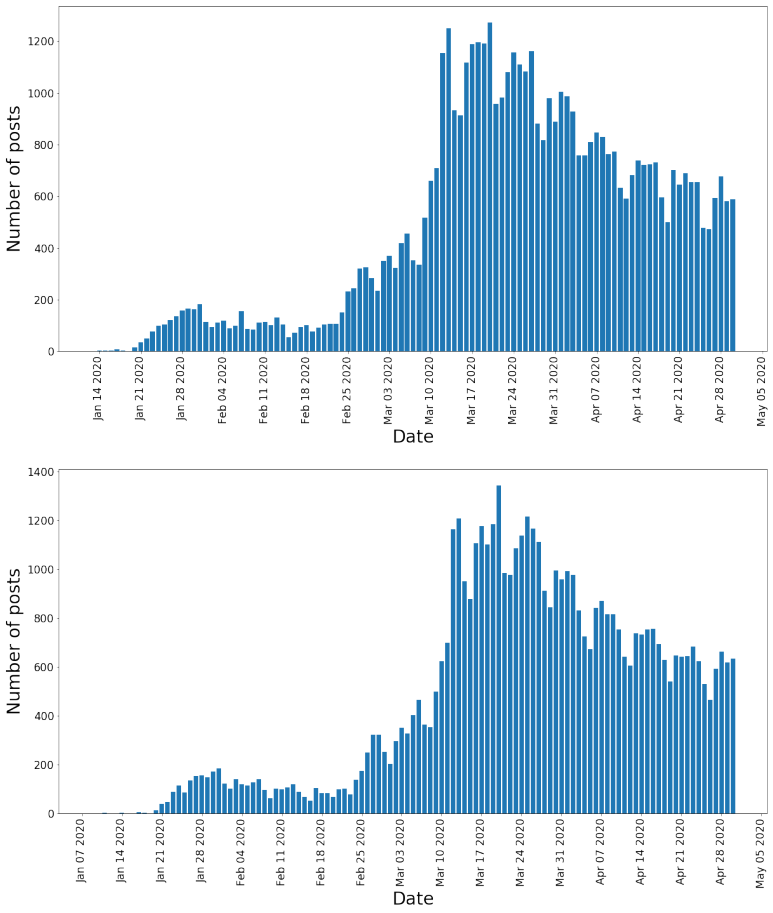

Figure 15: Trend of the number of posts over time for S_1 (on top) and S_2 (on bottom)

4.4 Approach to build virtual communities of users with homogeneous interests

In the experiments to evaluate this approach, we decided to work on the samples S_1 and S_2 described in Section 4.3.

After building the two networks, we computed some basic parameters of them. They are shown in the second and third column of Table 7. Their examination immediately revealed a problem, namely the great variance in the number of keywords per author. Analyzing in more detail the average values and the ones associated with the quartiles, it emerged that this variance was due to the presence of some outlier authors. Examining them carefully, we realized that they were bots (i.e., automatic Reddit users that posted news crawled from different sources), so they were not of interest for the goal we were pursuing. Therefore, we decided to remove them. The basic parameters of the new networks $\mathcal{S}_1''$ and $\mathcal{S}_2''$, obtained after the removal of bots from $\mathcal{S}_1'$ and $\mathcal{S}_2'$, are shown in the fourth and fifth columns of Table 7. The distribution of the arcs of $\mathcal{S}_1''$ and $\mathcal{S}_2''$ against the number of associated keywords is reported in Table 8.

In order to give an idea of how our approach works, we describe its application to the networks $\mathcal{S}_1''$ and $\mathcal{S}_2''$. The sets of at least 3 keywords occurring most frequently in the arcs of $\mathcal{S}_1''$ and $\mathcal{S}_2''$, along

Virtual subreddit	*Keyword(s) from which it originated*	*Set of keywords associated with it*	*Number of assigned posts*
$R_{1,6}$	case, death	world, ontario, city, death, number, health, toronto, week, report, worker, rate, case, total, china, rise, country, home, resident, official, people, patient, test, day, toll, york	34,071
R_2	world	world, china, case, death, people, report, trump, health, virus, ontario, number, day, test, home, toronto, response, organization, time, country	18,205
R_3	people	people, health, case, death, world, worker, test, ontario, report, home, toronto, number, day, china, trump, country, virus, help, spread	18,816
R_4	trump	trump, president, news, state, house, health, world, china, case, death, report, country, people, response, government, ontario, toronto, claim, administration, test, virus	18,280
R_5	health	health, case, ontario, death, report, number, total, world, day, city, official, toronto, trump, home, minister, test, china, people, worker, help	18,278
R_7	china	china, world, case, health, death, trump, report, country, people, ontario, number, day, toll, home, toronto, virus, test, news, flight, wuhan	18,036
R_8	ontario	ontario, case, death, report, number, health, total, world, day, home, toronto, nursing, people, test, worker, hospital, patient, icu, week	15,492
R_9	test	test, ontario, case, death, home, total, health, worker, minister, world, report, people, employee, kit, china, mask, help, result, time, day, hospital, member, staff	17,833
R_{10}	home	home, death, case, report, ontario, health, number, toronto, world, nursing, resident, retirement, test, stay, life, total, worker, help, city, people, day, work	16,272

Table 5: The virtual subreddits constructed for S_1

Virtual subreddit	*Keyword(s) from which it originated*	*Set of keywords associated with it*	*Number of assigned posts*
R_1	case	case, death, report, ontario, home, health, world, rate, number, toronto, total, test, worker, china, state, york, people, country, organization, trump	17,499
R_2	world	world, china, case, death, people, health, report, trump, response, ontario, organization, test, number, country, state, home, rate, toronto, outbreak, time, york, day	18,871
R_3	health	health, official, case, death, trump, state, ontario, toronto, number, report, total, world, country, china, organization, time, people, home, rate, minister, test, patient, worker, outbreak	18.728
R_4	death	death, case, ontario, report, number, health, total, world, country, state, china, toll, rise, home, city, toronto, outbreak, test, worker, resident, people, organization, rate, day	17,028
R_5	people	people, health, case, world, death, ontario, organization, test, report, home, toronto, staff, china, country, time, trump, number, help, day	18,245
R_6	trump	trump, president, news, test, world, house, response, ontario, china, health, case, death, organization, report, people, call, administration, state, claim	16,762
R_7	china	china, world, health, case, death, organization, country, time, report, trump, people, ontario, number, state, home, toronto, test, virus, response, flight	18,244
R_8	ontario	ontario, case, death, report, number, health, total, world, state, home, toronto, work, people, staff, hospital, worker, test, patient, day, week, province	16,414
R_9	test	test, ontario, case, death, home, total, health, worker, world, minister, organization, report, employee, work, hospital, kit, china, toronto, result, day, time, people, staff, member, country	17,632
R_{10}	home	home, death, case, report, ontario, health, world, number, toronto, stay, life, country, response, nursing, resident, staff, total, test, worker, week, outbreak, spread, help, work, people, day, china	18,902

Table 6: The virtual subreddits constructed for S_2

with the corresponding number of occurrences, is shown in Figure 16. Some fundamental parameters about them are reported in Table 9. In Figure 17 (resp., 18), we show four communities derived from $\mathcal{S}_1''$ (resp., $\mathcal{S}_2''$) to give an idea of them. In Table 10, we report the density and clustering coefficient of $\mathcal{S}_1''$ and $\mathcal{S}_2''$, as well as the average values of these parameters for the networks associated with the communities returned by our approach. As we can see, both the average density and the average clustering coefficient of the networks returned by our approach are higher, or much higher, than the

Parameter	*Value in* $\mathcal{S}'_1$	*Value in* $\mathcal{S}'_2$	*Value in* $\mathcal{S}''_1$	*Value in* $\mathcal{S}''_2$
Number of nodes	23,835	24,084	22,204	22,457
Number of arcs	6,956,916	6,972,620	3,326,119	3,392,481
Density	0.0245	0.0240	0.0130	0.0135
Average clustering coefficient	0.7374	0.7332	0.7010	0.6960
Average weight of arcs	1.20	1.19	1.02	1.02
Average number of keywords per author	6.640	6.618	2.763	2.780
Standard deviation of the number of keywords per author	159.5453	159.8500	1.7472	1.7425
Maximum number of keywords per author	21,185	21,293	8	8
Minimum number of keywords belonging to the first quartile	4	4	4	4
Minimum number of keywords belonging to the second quartile	2	3	2	2
Minimum number of keywords belonging to the third quartile	1	1	1	1
Minimum number of keywords per authors	1	1	1	1

Table 7: Some basic parameters of the networks $\mathcal{S}'_1$, $\mathcal{S}'_2$, $\mathcal{S}''_1$ and $\mathcal{S}''_2$

Number of keywords	*Number of arcs in* $\mathcal{S}''_1$	*Number of arcs in* $\mathcal{S}''_2$
1	3,270,276	3,331,012
2	54,292	59,693
3	1,400	1,606
4	107	122
5	35	29
6	6	10
7	2	5
8	1	4

Table 8: Distribution of the arcs of $\mathcal{S}''_1$ and $\mathcal{S}''_2$ against the number of associated keywords

ones of $\mathcal{S}''_1$ and $\mathcal{S}''_2$. This is an indicator that our approach is really capable of finding new user communities with homogeneous interests.

Parameter	*Value in* $\mathcal{S}''_1$	*Value in* $\mathcal{S}''_2$
Average number of occurrences of the sets of keywords	7.02	6.98
Standard deviation of the number of occurrences of the sets of keywords	13.95	17.27
Maximum number of occurrences of the sets of keywords	116	185
Minimum number of occurrences of the sets of keywords belonging to the first quartile	6	6
Minimum number of occurrences of the sets of keywords belonging to the second quartile	3	3
Minimum number of occurrences of the sets of keywords belonging to the third quartile	3	3
Minimum number of occurrences of the sets of keywords	3	3

Table 9: Some fundamental parameters of the sets of at least 3 keywords occurring at least 3 times in the arcs of $\mathcal{S}''_1$ and $\mathcal{S}''_2$

Networks	(Average) Density	(Average) Clustering Coefficient
$\mathcal{S}''_1$	0.0130	0.7010
$\mathcal{S}''_2$	0.0135	0.6960
Networks representing author communities derived from $\mathcal{S}''_1$	0.9498	0.9267
Networks representing author communities derived from $\mathcal{S}''_2$	0.9382	0.9114

Table 10: Values of (average) density and (average) clustering coefficients for $\mathcal{S}''_1$ and $\mathcal{S}''_2$ and the networks associated with the communities obtained by applying our approach

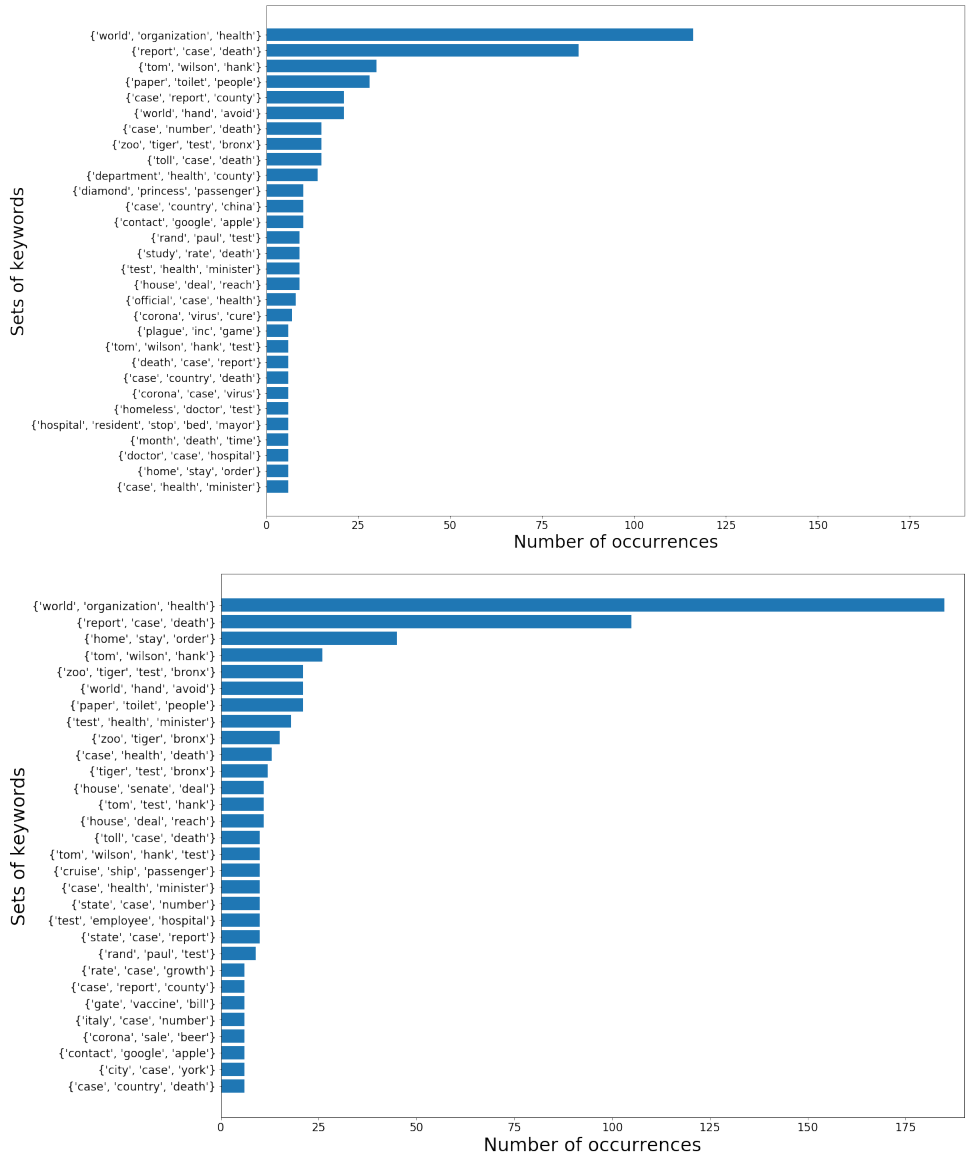

Figure 16: Most frequent sets of at least 3 keywords occurring at least 3 times in the arcs of $\mathcal{S}_1''$ (on top) and $\mathcal{S}_2''$ (on bottom) and corresponding number of occurrences

5 Conclusion

In this paper, we presented three approaches to extract information from posts on COVID-19 published on Reddit. The first approach is semi-automatic and incremental. It aims at building, and then updating, a classification of posts on Reddit. This classification allows us to define a hierarchy of classes each characterized by a set of keywords. The second approach is automatic and allows the identification of a set of themes concerning COVID-19. Each theme deals with homogeneous topics and has homogeneous posts associated with it. It can also be seen as the core for the realization of a

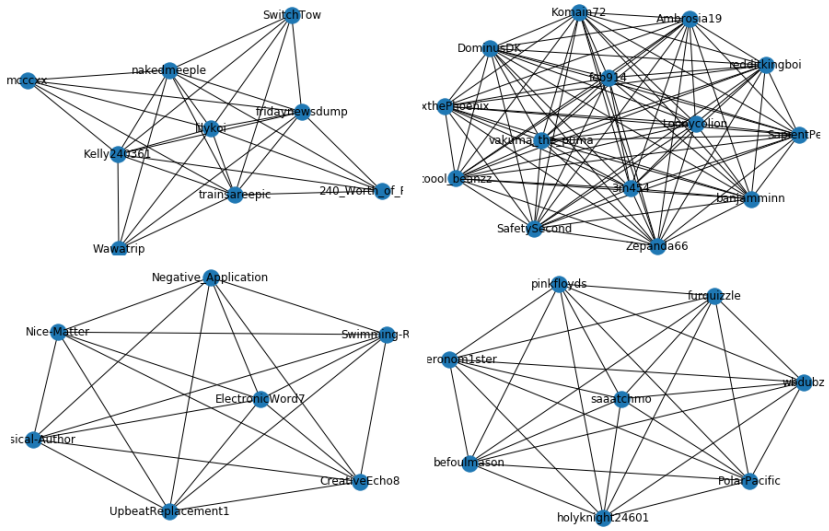

Figure 17: Four communities of authors with homogeneous interests derived from $\mathcal{S}_1''$

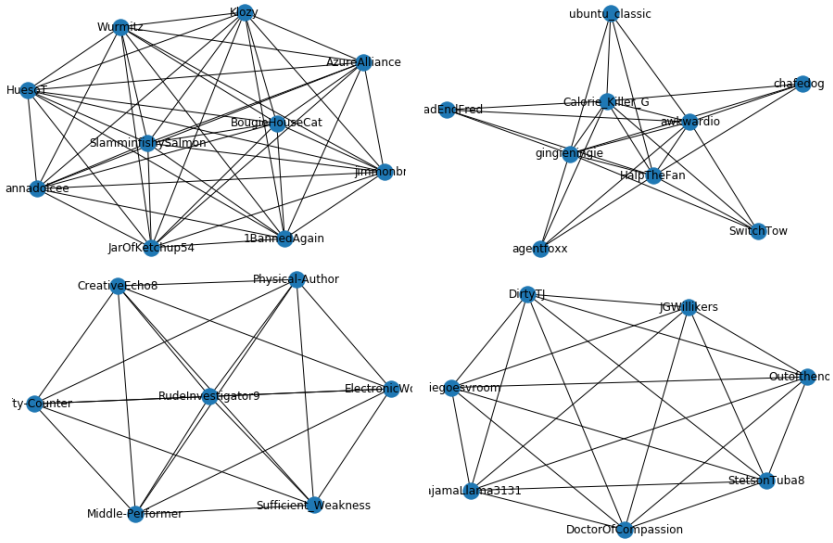

Figure 18: Four communities of authors with homogeneous interests derived from $\mathcal{S}_2''$

virtual subreddit. The third approach is automatic and allows the construction of virtual communities of users having the same interests. It can be useful to define a recommender system suggesting to a user other ones with similar interests or to allow Reddit to propose a new functionality aiming at creating communities of users with common interests. We applied the three approaches on the posts on COVID-19 published on Reddit between January and April 2020 and also reported the information discovered. Finally, we highlighted that the proposed approaches can be applied to analyze the posts about other emergencies published on Reddit.

In the future, we plan to extend the research proposed in this paper along various directions.

First, we would like to generalize the proposed approaches so that they can also operate on other social networks, such as Quora, 4Chan and Digg, just to cite a few. Moreover, we plan to define collaborative filtering recommender systems exploiting the results of the three approaches discussed in this paper to suggest to users other users and subreddits with similar interests. Last, but not the least, we plan to apply sentiment analysis techniques to identify new forms of classification of Reddit users, which consider not only the content of their posts but also the sentiments used to express them.

References

[1] J. Aggarwal, E. Rabinovich, and S. Stevenson. Exploration of gender differences in COVID-19 discourse on Reddit. *arXiv preprint arXiv:2008.05713*, 2020.

[2] E. Akbas and P. Zhao. Attributed graph clustering: An attribute-aware graph embedding approach. In *Proc. of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM'17)*, pages 305–308, Sydney, Autralia, 2017.

[3] A. Akkaya and N. İlhan. Sentiment Analysis of the Coronavirus Vaccine on Social Media. In *Proc. of the International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT'21)*, pages 295–299, Ankara, Turkey, 2021. IEEE.

[4] A. Alambo, M. Gaur, U. Lokala, U. Kursuncu, K. Thirunarayan, A. Gyrard, A. Sheth, R.S. Welton, and J. Pathak. Question answering for suicide risk assessment using Reddit. In *Proc. of the International Conference on Semantic Computing (ICSC'19)*, pages 468–473, Newport Beach, CA, USA, 2019. IEEE.

[5] M.S. Anbarasi, V. Iswarya, M. Sindhuja, and S. Yogabindiya. Ontology oriented concept based clustering. *International Journal of Research in Engineering and Technology*, 3(2), 2014.

[6] R. Baghel and R. Dhir. A frequent concepts based document clustering algorithm. *International Journal of Computer Applications*, 4(5):6–12, 2010.

[7] V. Basile, F. Cauteruccio, and G. Terracina. How Dramatic Events Can Affect Emotionality in Social Posting: The Impact of COVID-19 on Reddit. *Future Internet*, 13(2),29:1–32, 2021. MDPI, Basel, Switzerland.

[8] M. E. Basiri, S. Nemati, M. Abdar, S. Asadi, and U.R. Acharrya. A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets. *Knowledge-Based Systems*, 228:107242, 2021. Elsevier.

[9] Z. Batooli and M. Sayyah. Measuring social media attention of scientific research on novel coronavirus disease 2019 (COVID-19): An investigation on article-level metrics data of dimensions. *Preprint from Research Square*, 2020.

[10] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. The pushshift Reddit dataset. In *Proc. of the International AAAI Conference on Web and Social Media (ICWSM'20)*, volume 14, pages 830–839, Atlanta, GA, USA, 2020. AAAI Press.

[11] S. Bhatt, S. Padhee, A. Sheth, K. Chen, V. Shalin, D. Doran, and B. Minnery. Knowledge graph enhanced community detection and characterization. In *Proc. of the International Conference on Web Search and Data Mining (WSDM'19)*, pages 51–59, Melbourne, Australia, 2019.

[12] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. MIT Press.

[13] C. Bothorel, J.D. Cruz, M. Magnani, and B. Micenkova. Clustering attributed graphs: models, measures and methods. *Network Science*, 3(3):408–444, 2015. Cambridge University Press.

[14] F. Cauteruccio, E. Corradini, G. Terracina, D. Ursino, and L. Virgili. Investigating Reddit to detect subreddit and author stereotypes and to evaluate author assortativity. *Journal of Information Science*, 2021. SAGE.

[15] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A.E. Hassanien. Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. *Applied Soft Computing*, 97:106754, 2020. Elsevier.

[16] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceeding of the ACM on Human-Computer Interaction*, 1(CSCW):31:1–31:22, 2017. ACM.

[17] E. Chen, K. Lerman, and E. Ferrara. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance*, 6(2):e19273, 2020. JMIR Publications.

[18] C. Chew and G. Eysenbach. Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 outbreak. *PloS one*, 5(11), 2010. Public Library of Science.

[19] P. Chunaev. Community detection in node-attributed social networks: a survey. *Computer Science Review*, 37:100286, 2020. Elsevier.

[20] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C.M. Valensise, E. Brugnoli, A.L. Schmidt, P. Zola, F. Zollo, and A. Scala. The COVID-19 social media infodemic. *arXiv preprint arXiv:2003.05004*, 2020.

[21] E. Corradini, A. Nocera, D. Ursino, and L. Virgili. Investigating the phenomenon of NSFW posts in Reddit. *Information Sciences*, 566:140–164, 2021. Elsevier.

[22] T.O. Cunha, I. Weber, H. Haddadi, and G.L. Pappa. The effect of social feedback in a Reddit weight loss community. In *Proc. of the International Conference on Digital Health Conference (ICDHT'16)*, pages 99–103, Bordeaux, France, 2016. Springer.

[23] P. De Meo, G. Quattrone, G. Terracina, and D. Ursino. Integration of XML Schemas at various "severity" levels. *Information Systems*, 31(6):397–434, 2006.

[24] A.D. Dubey. Twitter Sentiment Analysis during COVID-19 Outbreak. *Available at SSRN 3572023*, 2020.

[25] Z. Ertem, A. Veremyev, and S. Butenko. Detecting large cohesive subgroups with high clustering coefficients in social networks. *Social Networks*, 46:1–10, 2016. Elsevier.

[26] H. Fang, H. Cheng, and M. Ostendorf. Learning latent local conversation modes for predicting comment endorsement in online discussions. In *Proc. of the International Workshop on Natural Language Processing for Social Media (SocialNLP'16)*, pages 55–64, Austin, TX, USA, 2016.

[27] N. Gozzi, M. Tizzani, M. Starnini, F. Ciulla, D. Paolotti, A. Panisson, and N. Perra. Collective response to media coverage of the covid-19 pandemic on reddit and wikipedia: Mixed-methods analysis. *Journal of Medical Internet Research*, 22(10):e21597, 2020. JMIR.

[28] J. Gu, W. Feng, J. Zeng, H. Mamitsuka, and S. Zhu. Efficient semisupervised MEDLINE document clustering with MeSH-semantic and global-content constraints. *IEEE Transactions on Cybernetics*, 43(4):1265–1276, 2012. IEEE.

[29] A. Gupta, J. Gautam, and A. Kumar. A survey on methodologies used for semantic document clustering. In *Proc. of the International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS'17)*, pages 671–675, Chennai, India, 2017. IEEE.

[30] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques - Third Edition*. 2011. Morgan Kaufmann notes.

[31] B.D. Horne and S. Adali. The impact of crowds on news engagement: A reddit case study. In *Proc. of the International AAAI Conference on Web and Social Media (ICWSM'17)*, page 11(1), Montréal, Québec, Canada, 2017.

[32] B.D. Horne, S. Adali, and S. Sikdar. Identifying the social signals that drive online discussions: A case study of Reddit communities. In *Proc. of the International Conference on Computer Communication and Networks (ICCCN'17)*, pages 1–9, Vancouver, BC, Canada, 2017. IEEE.

[33] M.S. Hossain and R.A. Angryk. GDClust: A graph-based document clustering technique. In *Proc. of the International Conference on Data Mining Workshops (ICDMW'07)*, pages 417–422, Washington, DC, USA, 2007. IEEE.

[34] M. Hossin and M.N. Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015. Academy & Industry Research Collaboration Center (AIRCC).

[35] L. Jin, Y. Chen, T. Wang, P. Hui, and A.V. Vasilakos. Understanding user behavior in online social networks: A survey. *IEEE Communications Magazine*, 51(9):144–150, 2013. IEEE.

[36] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[37] J. Kim and M. Hastak. Social network analysis: Characteristics of online social networks after a disaster. *International Journal of Information Management*, 38(1):86–96, 2018.

[38] B. Kleinberg, I. van der Vegt, and M. Mozes. Measuring emotions in the Covid-19 real world worry dataset. *arXiv preprint arXiv:2004.04225*, 2020.

[39] G. Kou and Y. Peng. An application of latent semantic analysis for text categorization. *International Journal of Computers Communications & Control*, 10(3):357–369, 2015. CCC Publications.

[40] G. Kou, Y. Peng, and G. Wang. Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information Sciences*, 275:1–12, 2014. Elsevier.

[41] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F.E. Alsaadi. Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing*, 86:105836, 2020. Elsevier.

[42] S. Lahiri, S.R. Choudhury, and C. Caragea. Keyword and keyphrase extraction using centrality measures on collocation networks. *arXiv preprint arXiv:1401.6571*, 2014.

[43] Y. Lama, D. Hu, A. Jamison, S.C. Quinn, and D.A. Broniatowski. Characterizing trends in Human Papillomavirus Vaccine discourse on Reddit (2007-2015): an observational study. *JMIR Public Health and Surveillance*, 5(1):e12480, 2019. JMIR Publications.

[44] A. Leavitt and J.A. Clark. Upvoting hurricane Sandy: event-based news production processes on a social news site. In *Proc. of the International Conference on Human Factors in Computing Systems (SIGCHI'14)*, pages 1495–1504, Toronto, Canada, 2014. ACM.

[45] A. Leavitt and J.J. Robinson. The role of information visibility in network gatekeeping: Information aggregation on reddit during crisis events. In *Proc. of the International Conference on Computer Supported Cooperative Work and Social Computing (CSCW'17)*, pages 1246–1261, Portland, OR, USA, 2017.

[46] T. Li, G. Kou, Y. Peng, and S.Y. Philip. An integrated cluster detection, optimization, and interpretation approach for financial data. *IEEE Transactions on Cybernetics*, 2021. IEEE.

[47] T. Li, G. Kou, Y. Peng, and Y. Shi. Classifying with adaptive hyper-spheres: An incremental classifier based on competitive learning. *IEEE transactions on systems, man, and cybernetics: systems*, 50(4):1218–1229, 2017. IEEE.

[48] D.M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, and S.S. Ghosh. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during Covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635, 2020. JMIR.

[49] J. Lu, S. Sridhar, R. Pandey, M.A. Hasan, and G. Mohler. Redditors in recovery: text mining Reddit to investigate transitions into drug addiction. *arXiv preprint arXiv:1903.04081*, 2019.

[50] P.D. Mahendhiran and S. Kannimuthu. Deep learning techniques for polarity classification in multimodal sentiment analysis. *International Journal of Information Technology & Decision Making*, 17(03):883–910, 2018. World Scientific.

[51] J.N. Matias. Going dark: Social factors in collective action against platform operators in the Reddit blackout. In *Proc. of the International Conference on Human Factors in Computing Systems (ACM CHI 2016)*, pages 1138–1151, San Jose, CA, USA, 2016. ACM.

[52] A.N. Medvedev, R. Lambiotte, and J.C. Delvenne. The anatomy of Reddit: An overview of academic research. In *Dynamics on and of Complex Networks*, pages 183–204. Indianapolis, IN, USA, 2017.

[53] R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *Proc. of the International Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 404–411, Qatar, Qatar, 2004. Association for Computational Linguistics.

[54] A.G. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[55] M. Miller, T. Banerjee, R. Muppalla, W. Romine, and A. Sheth. What are people tweeting about Zika? An exploratory study concerning its symptoms, treatment, transmission, and prevention. *JMIR Public Health and Surveillance*, 3(2):e38, 2017. JMIR Publications.

[56] L. Mitchell. *A Phenomenological study of social media: boredom and interest on Facebook, Reddit, and 4chan.* 2012. University of Victoria, British Columbia, Canada.

[57] C. Murray, L. Mitchell, J. Tuke, and M. Mackay. Symptom extraction from the narratives of personal experiences with COVID-19 on Reddit. *arXiv preprint arXiv:2005.10454*, 2020.

[58] U. Naseem, I. Razzak, M. Khushi, P.W. Eklund, and J. Kim. Covidsenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis. *IEEE Transactions on Computational Social Systems*, 8(4):1003–1015, 2021. IEEE.

[59] R. Navigli and S.P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012. Elsevier.

[60] L. Nemes and A. Kiss. Social media sentiment analysis based on COVID-19. *Journal of Information and Telecommunication*, 5(1):1–15, 2021. Taylor & Francis.

[61] E. Newell, D. Jurgens, H.M. Saleem, H. Vala, J. Sassine, C. Armstrong, and D. Ruths. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. In *Proc. of the International Conference on Web and Social Media (ICWSM 2016)*, pages 279–288, Cologne, Germany, 2016. AAAI.

[62] M.E.J. Newman. Properties of highly clustered networks. *Physical Review E*, 68(2):026121, 2003. APS.

[63] M.E.J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5):323–351, 2005. Taylor & Francis.

[64] K. Nishimoto and K. Matsuda. Informal communication support media for encouraging knowledge-sharing and creation in a community. *International Journal of Information Technology & Decision Making*, 6(03):411–426, 2007. World Scientific.

[65] T. O'Neill. 'Today I Speak': Exploring How Victim-Survivors Use Reddit. *International Journal for Crime, Justice and Social Democracy*, 7(1):44, 2018. Queensland University of Technology.

[66] L. Palopoli, D. Saccà, G. Terracina, and D. Ursino. Uniform techniques for deriving similarities of objects and subschemes in heterogeneous databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):271–294, 2003.

[67] T. Pay. Totally automated keyword extraction. In *Proc. of the International Conference on Big Data (Big Data 2016)*, pages 3859–3863, Washington, D.C., USA, 2016. IEEE.

[68] R. Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001. JSTOR.

[69] D. Qiu, H. Li, and Y. Li. Identification of active valuable nodes in temporal online social network with attributes. *International Journal of Information Technology & Decision Making*, 13(04):839–864, 2014. World Scientific.

[70] E. Rahm and P.A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350, 2001.

[71] T.B.A. Rakib and L.K. Soon. Using the Reddit corpus for cyberbully detection. In *Proc. of the Asian Conference on Intelligent Information and Database Systems (ACIIDS'18)*, pages 180–189, Dong Hoi City, Vietnam, 2018. Springer.

[72] F. Role and M. Nadif. Beyond cluster labeling: Semantic interpretation of clusters' contents using a graph representation. *Knowledge-Based Systems*, 56:141–155, 2014. Elsevier.

[73] S. Romeo, A. Tagarelli, and D. Ienco. Semantic-based multilingual document clustering via tensor modeling. *Available at* `https://hal.archives-ouvertes.fr/hal-01130094/`, 2014.

[74] S. Rose, D. Engel, N. Cramer, and W. Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1:1–20, 2010. Wiley, New York.

[75] M. Roy, N. Moreau, C. Rousseau, A. Mercier, A. Wilson, and L. Atlani-Duault. Ebola and localized blame on social media: analysis of Twitter and Facebook conversations during the 2014–2015 Ebola epidemic. *Culture, Medicine, and Psychiatry*, 44(1):56–79, 2020. Springer.

[76] N.Y. Saiyad, H.B. Prajapati, and V.K. Dabhi. A survey of document clustering using semantic approach. In *Proc. of the International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT'16)*, pages 2555–2562, Chennai, India, 2016. IEEE.

[77] N. Schrading, C.O. Alm, R. Ptucha, and C. Homan. An analysis of domestic abuse discourse on Reddit. In *Proc. of the International Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*, pages 2577–2583, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[78] J. Sedding and D. Kazakov. WordNet-based text document clustering. In *Proc. of the International Workshop on RObust Methods in Analysis of Natural Language Data (ROMAND 2004)*, pages 104–113, Geneva, Switzerland, 2004.

[79] M. Sharma, K. Yadav, N. Yadav, and K.C. Ferdinand. Zika virus pandemic-analysis of Facebook as a social media health information platform. *American Journal of Infection Control*, 45(3):301–302, 2017. Elsevier.

[80] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, and Y. Wang. A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv preprint arXiv:2003.13907*, 2020.

[81] G. Stoddard. Popularity dynamics and intrinsic quality in Reddit and hacker news. In *Proc. of the International AAAI Conference on Web and Social Media (ICWSM'15)*, page 9(1), Oxford, UK, 2015.

[82] Y.A. Strekalova. Health risk information engagement and amplification on social media: News about an emerging pandemic on Facebook. *Health Education & Behavior*, 44(2):332–339, 2017. SAGE Publication.

[83] M. Suran and D.K. Kilgo. Freedom from the press? How anonymous gatekeepers on Reddit covered the Boston Marathon bombing. *Journalism Studies*, 18(8):1035–1051, 2017. Taylor & Francis.

[84] G. Tang, Y. Xia, E. Cambria, P. Jin, and T.F. Zheng. Document representation with statistical word senses in cross-lingual document clustering. *International Journal of Pattern Recognition and Artificial Intelligence*, 29(02):1559003, 2015. World Scientific.

[85] M. Tsvetovat and A. Kouznetsov. *Social Network Analysis for Startups: Finding connections on the social web*. Sebastopol, CA, USA, 2011. O'Reilly Media, Inc.

[86] S. Wold, K. Esbensen, and P. Geladi. Principal Component Analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. Elsevier.

[87] A. Yadav and D.K. Vishwakarma. A Language-independent Network to Analyze the Impact of COVID-19 on the World via Sentiment Analysis. *ACM Transactions on Internet Technology (TOIT)*, 22(1):1–30, 2021. ACM.

[88] C. Yan, M. Law, S. Nguyen, J. Cheung, and J. Kong. Comparing Public Sentiment Toward COVID-19 Vaccines Across Canadian Cities: Analysis of Comments on Reddit. *Journal of medical Internet research*, 23(9):e32685, 2021. JMIR.

[89] M. Zhan, H. Liang, G. Kou, Y. Dong, and S. Yu. Impact of social network structures on uncertain opinion formation. *IEEE Transactions on Computational Social Systems*, 6(4):670–679, 2019. IEEE.

[90] J.S. Zhang, B.C. Keegan, Q. Lv, and C. Tan. A tale of two communities: Characterizing reddit response to covid-19 through/r/china_flu and/r/coronavirus. *arXiv preprint arXiv:2006.04816*, 2020.

[91] B. Zhu, X. Zheng, H. Liu, J. Li, and P. Wang. Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics. *Chaos, Solitons & Fractals*, 140:110123, 2020. Elsevier.

[92] C. Zhu, J. Ma, D. Zhang, X.Han, and X. Niu. Hierarchical document classification based on a backtracking algorithm. In *Proc. of the International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'08)*, volume 2, pages 467–471, Jinan, China, 2008. IEEE.