Social media monitoring tools for air quality accounts

Polychronis Charitidis
ITI-CERTH
Thessaloniki, Greece
charitidis@iti.gr

Symeon Papadopoulos
ITI-CERTH
Thessaloniki, Greece
papadop@iti.gr

Lazaros Apostolidis
ITI-CERTH
Thessaloniki, Greece
laaposto@iti.gr

Ioannis Kompatsiaris
ITI-CERTH
Thessaloniki, Greece
ikom@iti.gr

ABSTRACT

In recent years, social media have become a powerful medium of public discourse, and in particular microblogging services (e.g., Twitter) play a crucial role in information dissemination. Being able to efficiently monitor and glean insights from streams of social media interactions could be valuable for better mapping and engaging with a community of interest. In this paper, we focus on the domain of air quality and present the outcomes of our work in the context of a research project, called hackAIR. To fulfill the project's communication needs we provide novel tools for effectively discovering key social media accounts, keeping up-to-date with trending news in the area of air quality, and better understanding the social media audience of the project Twitter account. We showcase our tools through the use case of hackAIR's social media communication activities.

CCS CONCEPTS

• **Information systems** → **Web searching and information discovery**; **Web mining**; **Social networks**.

KEYWORDS

social media, account recommendation, discussion mining, news search, audience analytics, community detection

ACM Reference Format:
Polychronis Charitidis, Symeon Papadopoulos, Lazaros Apostolidis, and Ioannis Kompatsiaris. 2019. Social media monitoring tools for air quality accounts. In *WI '19 Companion: IEEE/WIC/ACM International Conference on Web Intelligence (WI '19 Companion), October 14–17, 2019, Thessaloniki, Greece.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3358695.3361106

1 INTRODUCTION

The advent of social media led to the establishment of new channels for communication, self-expression, information gathering and sharing for billions of Internet users worldwide. The penetration of social media has consistently increased throughout the last decade and it has nowadays reached a point where the large majority of people worldwide regularly use one or more of numerous popular platforms for a wide variety of purposes. As a result, social media is often regarded as a sensor of real-world trends and events [1] and as an indispensable tool for performing social science research [17] and consumer intelligence gathering and marketing [6].

In this paper, we present the outcomes of our work in the context of a Horizon 2020 research and innovation project, called hackAIR[1]. hackAIR is an open technology platform that can be used to access, collect and improve Air Quality (AQ) information in Europe. Given the increasing importance of social media for information sharing and intelligence gathering, the hackAIR project leveraged social media monitoring in order to support and reinforce its key research and innovation activities. In particular, the communication team of the project was in need of effective ways of discovering key social media accounts to engage, keeping up-to-date with trending news in the area of air quality, and to better understand their social media audience. To this end, an existing open-source social media monitoring framework [21] was configured, deployed and extended to serve these communication needs.

Given the project communication needs, three tools were designed: One tool automates the process of identifying AQ-related

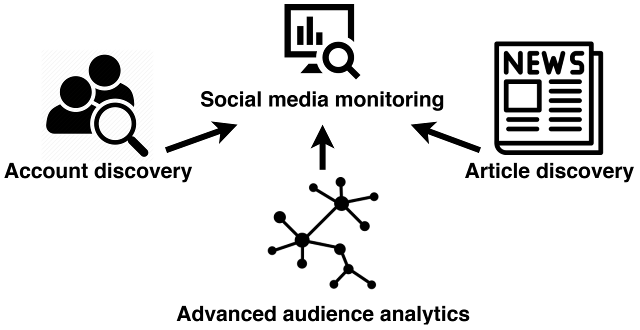


Figure 1: We developed three social media mining tools to support the engagement and communication strategy of the H2020 hackAIR project focusing on Air Quality (AQ) topics: a) relevant account discovery, b) news article discovery, and c) advanced audience analytics.

[1] http://www.hackair.eu/

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WI '19 Companion, October 14–17, 2019, Thessaloniki, Greece
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6988-6/19/10…$15.00
https://doi.org/10.1145/3358695.3361106

users from social media and provides analytics on their activities. This means that it offers an easy way to inspect, communicate and interact with such accounts which is crucial for disseminating the project outcomes to a wide community of relevant stakeholders. A second tool provides social media account managers with AQ-related newsworthy content. This is particularly useful because it helps managers stay up-to-date with the latest developments and trends in the area of AQ research, technology, activism and policy. Also, this task provides insights about the news people share in social media and discovers content that is worth sharing, potentially leading to an increased follower base. The final tool provides advanced audience analysis through employing network analysis algorithms to discover communities and creating a more structured view over the project social media audience. The visualizations provided by this extension are valuable for understanding the different segments of the project audience and measuring the reach and impact of the project dissemination activities. Figure 1 presents a schematic representation of the three tools.

Although the approaches discussed in this paper often refer specifically to the hackAIR project and its Twitter account, they are possible to apply to other research projects and Twitter accounts, as long as the topics of interest include air quality and pollution, for which some specific keyword sets and classifiers have been built, as will be described in the following.

In Section 2, we summarize a few past works that are related to the three developed tools. Section 3 presents the proposed methodology for discovering air quality related accounts on different social media channels. Next, Section 4 outlines the methodology for discovering relevant and trending AQ-related newsworthy content. Section 5 presents the methodology for analyzing and visualizing the social media audience of a Twitter account of interest. Section 6 showcases some of the results on the hackAIR Twitter account and Section 7 concludes the paper.

2 RELATED WORK

Similarity-based social media user discovery: There are many works focused on the discovery of related or similar users in a social media context. Researchers from Twitter share their proposed framework for finding similar Twitter accounts in [8], where they propose an approach to discover similar Twitter users using different signals like cosine follow score, number of suggestions' followers, page rank score, and historic follow-through rate. Authors in [20] present a methodology for discovering and suggesting similar Twitter accounts, based entirely on their disseminated content in terms of Twitter entities used. Work in [16] propose a method to detect similar groups across multiple social networks, including detecting group structure based on random walks, extracting similarity features, and inferring group similarity using probabilistic model. Another study [28] estimates the similarity between users according to their physical location histories. In [25], a methodology is proposed for identifying user communities on Twitter, by defining a number of similarity metrics based on their shared content, following relationships and interactions.

News and topic detection in social media: In recent years, progress has been made in the field of news discovery and topic detection in social media. Social media is often seen as a sensor of trending topics and streams of tweets are analysed with the goal of extracting keywords, phrases or clusters of similar tweets that exhibit "bursty" behaviour and correspond to topics of increasing news interest [1]. Work in [26] propose a hierarchical entity-aware event discovery model to learn news events from multiple aspects in Twitter. Moreover, work in [27] utilizes a geotagging procedure to find Twitter users in a given geographical area. They track the post updates of such Twitter users and in second step they perform online clustering to the group tweets of to report potential news in the area. To identify hot topics, the approach presented in [14] demonstrates a clustering algorithm that identifies geographic communities by correlating the time series for a set of topic words. Authors in [24] present a topic tracking system that can provide temporal topic connections in online news feed.

Social media audience analysis: In terms of audience analytics and community detection [18], a social media analytics platform is presented in [22] that is oriented towards political context. In this work, a structural approach of social media audience analysis is examined for the identification of influential individuals or communities. Authors in [29] utilize Twitter analytics as well as the behavior of each user's audience to present an influence score for each user. In [7], a social multimedia analytics system is presented based on a community detection approach, using latent representation learning and multi-modal latent topic modeling and personal user profiling techniques. Work in [11] develops a approach to discover communities and influential users in Twitter by applying spectral clustering algorithm. Evaluation of community structure and topic modeling methods as a process for characterizing the clustering of opinions about human papillomavirus (HPV) vaccines on Twitter is presented in [23]. To deal with the detection of geolocated communities in Twitter in disaster situations authors in [2] propose a fast-greedy optimization of modularity clustering algorithm with semantic similarity so that it can deal with the complex social graphs extracted from Twitter.

3 AIR QUALITY ACCOUNT DISCOVERY

An important feature of several social media monitoring tools is the identification of accounts that are related with an account of interest. To this end, we developed an account discovery module that requires only the provision of an initial set of air quality-related keywords and then automatically generates and continuously refines a list of accounts related with air quality and air pollution. As relevant accounts, we define social media accounts that regularly post about air quality (or air pollution) topics. In addition, the module generates a number of statistics with respect to the activity, popularity and geographical distribution of the discovered accounts. This procedure is generally applicable and can identify air quality-related nodes in social media. In Section 5 we also provide a methodology that identifies influential and possibly relevant accounts based on the community structure around a selected account of interest.

The account discovery methodology comprises several steps that are illustrated in Figure 2. The first two steps deal with the creation of a seed list of multi-lingual keywords related to the topic of interest (in this case air quality) and their use as query terms

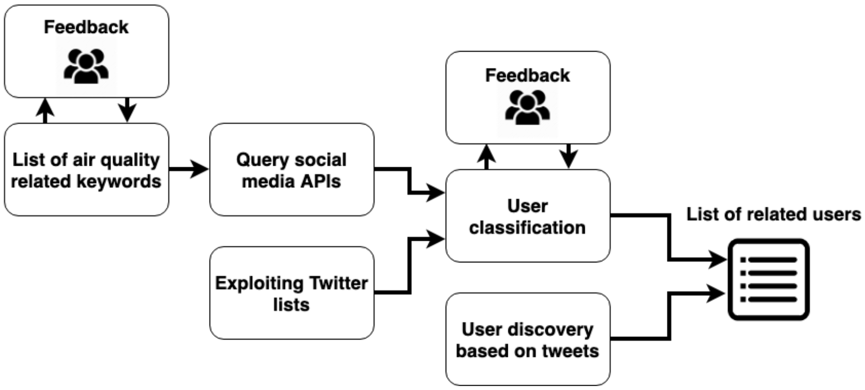


Figure 2: A schematic representation of the account discovery methodology

to the APIs of social media platforms. In these steps, experts on air quality issues provided feedback on the selected keywords and suggested additional ones that could be relevant to the topic. After that, a set of candidate accounts is generated by querying the APIs of multiple social media platforms. Then, the initial set of potentially relevant accounts is further expanded by employing a method that exploits account co-occurrence in Twitter lists to identify Twitter accounts that are similar with those retrieved in the first step.

Having a set of candidate accounts from the previous steps, a filtering step takes place that aims at discarding accounts that are not relevant with air pollution topics despite having been retrieved using air quality-related keywords (we observed that this is a common issue). To this end, a classifier is built and employed to classify accounts as relevant/irrelevant with air quality based on the text contained in the name and description fields of the accounts. To build the classifier, we used a set of training accounts that were manually annotated as relevant/irrelevant.

A limitation of the above account discovery pipeline is that it considers only accounts that contain air quality-related terms in their name or description fields while ignoring accounts that often post about air quality issues without having air quality-related terms in their profile (name and description). To this end, an additional step was realized, specifically for Twitter, with the aim of detecting accounts that frequently post air quality-related content. This is achieved by collecting posts that contain air quality-related terms over a long period of time and identifying accounts that consistently post about air quality. This set of accounts is combined with the accounts discovered based on their description and name fields to generate the final list of interesting accounts.

The output of this step is a set of accounts that are very likely to be relevant with air quality topics. As a final step, several statistics are calculated about the retained accounts such as popularity, activity and geographical distribution.

3.1 Profile-based social media account retrieval

The selection of appropriate keyword terms for querying the APIs of social media platforms is crucial for the overall effectiveness of the account discovery module as it determines the initial pool of candidate accounts. To generate the list of keywords we worked closely with domain experts from Greece, Germany, Norway and Belgium, and especially with engagement and communication managers of the hackAIR project. As a result of this collaborative effort, a list[2] of 464 multi-lingual (90 English, 78 German, 101 Norwegian, 98 French and 97 Dutch) air quality-related keywords was composed (e.g. air pollution, luftforurensing, luftverschmutzung, la pollution de l'air, luchtvervuiling).

Having composed the list of search keywords, the next step of the pipeline is the retrieval of candidate accounts by querying the APIs of social media platforms. Our initial intention was to use the APIs of Twitter, Facebook, YouTube and Google+. We excluded Google+ API because it produces very limited relevant information in preliminary experiments, and Facebook due to its recent changes[3] on April 2018, which introduced restrictions to some endpoints that were critical for the application. The following paragraphs provide the details of how each API was queried. The collected results are then processed by the subsequent steps of the pipeline of Figure 2.

Twitter. The Twitter REST API provides programmatic access to read and write operations such as creating new Tweets, reading user profile and follower data, etc. In this case we use it to retrieve Twitter account information in JSON format using keywords. Specifically, we use the `users/search` endpoint[4] that provides a simple, relevance-based search interface to public Twitter accounts. Documentation of the API also states that only the first 1,000 matching results are available per keyword query. Information about each retrieved account includes the account's name, location, language, description, number of followers, number of tweets, etc.

YouTube. There are different types of resources that can be retrieved using the YouTube Data API. To search for YouTube channels with air quality-related keywords, we use the `search/list` endpoint[5] which returns a collection of search results that match the query parameters specified in the API request. By default, the set of results contains all matching video, channel and playlist resources but can be configured to return only channels. For each channel, information such as name, description, number of subscribers/videos/views, etc. is provided.

[2] https://bit.ly/2Q7DJ37
[3] https://developers.facebook.com/docs/graph-api/changelog/breaking-changes/
[4] https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/api-reference/get-users-search
[5] https://developers.google.com/youtube/v3/docs/search/list

3.2 Twitter list mining

We try to further extend the collected set of candidate Twitter accounts by exploiting Twitter Lists[6] and leveraging the method presented in [13]. A Twitter List is a group of Twitter accounts that allows users to see tweets from a group of people in individual tweet timelines, without having to put them all together into their own timeline of people they follow. Twitter users can create their own lists or subscribe to lists created by others. Thus, one may expect that accounts appearing in the same list will likely post similar content and will therefore be similar to each other.

Our implementation of this algorithm works as follows: Let T be the target account for which we want to find similar accounts. In the first step, we retrieve all Twitter lists L_i in which T is a member using the `lists/memberships` endpoint of the Twitter API. Then, the `lists/members` endpoint is used to retrieve all other members (accounts) in the lists. Let U denote the set of unique accounts that were retrieved from all L_i Lists. The similarity of each account $u \in U$ with the target account T is computed by counting the number of times u appears in the same list with T, and then normalizing by dividing with N, the number of lists that the target account belongs to. To tell whether two accounts are similar enough, a similarity threshold θ must be defined. Taking into account the fact that the account discovery pipeline includes a sophisticated filtering step that will remove most of the irrelevant accounts, we opt for using a low threshold ($\theta = 0.3$), in order to allow for more potentially relevant accounts to be discovered. The algorithm described above is first executed using as seeds 10 Twitter accounts that are manually selected to be related to air pollution and also highly popular and active.

3.3 Account classification

The previous steps of the account discovery pipeline lead to a collection of accounts with air quality-related terms in their name and description fields. However, we found that a significant portion of these accounts are not relevant with air quality, e.g. because they concern more general environmental issues, indoor air quality commercial products, etc.

Table 1 provides some indicative examples of relevant and irrelevant accounts. The first and the third account in the table are directly related with AQ as they represent organization that monitor air pollution. On the other hand, the second account is related with electric car news and the forth is a commercial smoke detector product which are not related with AQ despite the fact that they contain AQ related terms in their description.

As a result, it was considered necessary to create an intelligent filtering mechanism for detecting and discarding irrelevant accounts. To this end, we trained a classifier to distinguish between relevant and irrelevant accounts based on textual metadata such as their name and description.

To build the classifier (`AQ_Account`) we created a training set of 1000 examples that were manually labeled as relevant or irrelevant by reading the textual fields of the account profile (name and description). The text was pre-processed by applying stemming and stop-word removal and represented using a tf-idf bag-of-words vector. As classification algorithm, we used L2-regularized L2-loss Support Vector Machines (`sklearn` [19] implementation based on [4]) with default parameters. Note that the classifier was trained using accounts with textual metadata in different languages and is therefore applicable to all accounts of the collection. Since the accuracy of the classifier depends heavily on the size and the quality of the annotations in the training set, to further improve the classifier we asked domain experts from the hackAIR project to help us improve the composition of the training set. In particular, we provided them with a large list of accounts along with the classifier's decisions and asked them to manually check the correctness of each classification. This way, a new training set of 1150 examples was created. Table 2 provides details about the `Air_Quality_Account` classifier before and after feedback. We notice that after feedback the training set was slightly expanded and that classifier performance improved (measured by Precision (P) and Recall (R) that were estimated using 10-fold cross-validation).

Name	Description	Label
Breathing London	Monitoring the Air Pollution in London since 2010. - data from http://t.co/YK6JzIRq	Relevant
Drive Electric	We promote electric vehicles as a fun, practical transportation alternative that reduces energy consumption and air pollution.	Irrelevant
Plume Labs	Proud builders of Flow, the first smart mobile air quality tracker, and the Plume Air Report, the free app to track live air pollution forecasts around the world.	Relevant
Birdi	Birdi Smart Detector tracks indoor air quality	Irrelevant

Table 1: Indicative examples of relevant/irrelevant accounts.

Classifier	samples	relevant	irrelevant	P	R
Before feedback	1000	400	600	85.0	85.4
After feedback	1150	450	700	87.5	86.8

Table 2: AQ_Account classifier details

3.4 Tweet-based account discovery

So far, we focused on discovering accounts that contain air quality-related text in their name or description. Here, we describe an approach for detecting Twitter accounts that post air quality-related information but were not possible to retrieve with the previously discussed methods. In short, the approach consists of collecting air quality-related tweets for a long period of time, using a classifier similar to the one presented above for filtering based on relevance and identifying accounts that regularly post such tweets.

To this end, we leveraged a large dataset of tweets that was initially collected to facilitate the needs for work presented in [5]. The dataset contains a large amount of tweets in English that contain air quality-related terms and were collected over a period of almost one year (8/2/2017-18/1/2018) around five cities in the UK (London, Liverpool, Manchester, Birmingham and Leeds) and five cities in US (New York, Boston, Pittsburgh, Philadelphia and Baltimore).

[6] https://developer.twitter.com/en/docs/accounts-and-users/create-manage-lists/overview

As in the case of social media accounts, not all tweets collected with air quality-related keywords are actually related to air quality. Thus, to distinguish relevant from irrelevant tweets we build a classifier `AQ_Tweet` using a training set of 1800 English tweets, manually labeled with respect to whether they provide information about air quality (relevant) or not (irrelevant). The same type of pre-processing is applied to the tweets as in the case of accounts (i.e. stemming and stop-word removal) and a tf-idf bag-of-words representation is used. We also use the same classification algorithm we used in the previous subsection, i.e. L2-regularized L2-loss Support Vector Machine with default parameters. Details of the classifier and its performance are shown in Table 3.

Classifier	samples	relevant	irrelevant	P	R
AQ_Tweet	1800	800	1000	90.3	89.4

Table 3: AQ_Tweet classifier details

After all tweets are classified, irrelevant tweets are removed and accounts are sorted based on the number of air quality-related tweets they post per day. We empirically found that using a threshold of 0.4 relevant tweets per day returns accounts that were relevant with air quality.

4 DISCOVERY OF AQ-RELATED ARTICLES

To keep track of the latest developments and trends in the area of air quality research, technology, activism and policy there is a need of a tool that provides latest information from various news sources. This tool works in a complementary manner with the methodologies discussed in previous sections, as it provides insights about the news people share in social media. The list of news related articles is provided in the form of an RSS feed.

To discover AQ-related content, we devised the following methodology. Initially, we create a list of Twitter users that are relevant with air quality as discussed in Section 3. This user list is updated in frequent intervals (two weeks) in order to continuously enrich the seed list with potentially newly appearing relevant accounts. The next step involves the retrieval of every tweet that is posted in the previous 24 hours from each account in the seed list.

To filter tweets that are irrelevant with air quality, we use an updated version of the classifier that was proposed in section 3.4. After the classification process we retrieve a list of tweets that are related to air quality. We then apply a method for estimating the geographic location of a tweet based on its text [15]. This method works by dividing the earth surface into rectangular cells, and then computing the probabilities of each term occurring in each cell, using a very large training corpus of geotagged items. Given an item with unknown location, a probability is computed for each cell based on the item's terms and the center of the most likely cell is used as the item's estimated location. Using this method we then distinguish between EU and non-EU referring tweets. We utilize the pool of AQ-related tweets to extract the shared links (URLs). In the next step, we filter these URLs using a curated list[7] of trustworthy web domains from news sites and air quality related resources to retrieve related articles. This list was compiled manually after careful inspection of the unfiltered URLs for a short period of time, discarding irrelevant or unrelated ones. We also employ some basic regular expressions to get articles from urls that do not appear in the list above The final step includes the title and metadata extraction from each collected article using the Newspaper3k[8] Python library. Metadata include author, publication date and content summary. The resulting RSS feed from the collected articles is updated daily.

Title	Summary
Halting deforestation is "just as urgent" as reducing emissions	By protecting and restoring forests, the world would avoid climate change consequences.
The satellite that can clean up space rubbish from Earth's orbit	RemoveDebris has been successfully fired into space as part of a plan to clean up the millions of pieces of rubbish floating in Earth's orbit.

Table 4: Irrelevant article examples

In some rare cases, the results happen to be irrelevant with air quality. Most of the times, such irrelevant articles are associated with climate change topics. This can be explained due to the fact that tweet classification cannot fully distinguish between air quality and climate change topics, since similar terms appear in both contexts. Table 4 presents a list of irrelevant examples that have been retrieved using the article discovery service. We demonstrate relevant article content in Section 6.

5 SOCIAL MEDIA AUDIENCE ANALYSIS

Even though the analytics offered by social media platforms provide a good overall idea about the popularity of an account/channel of interest and about the composition of its audience (in the form of aggregated demographics), they provide little or no guidance or support for implementing an engagement strategy. In particular, they neglect valuable information that is hidden in the structure of the network formed by the account's audience.

Following the typical structure of social networks, the follower network of a target Twitter account exhibits community structure [18], i.e. the nodes of the network (followers) tend to form groups that are densely connected to each other (i.e. between nodes of the same group) and sparsely connected to nodes of other groups. Uncovering community structure is of importance because it facilitates building an understanding regarding the dynamics of the network (e.g. information is transmitted faster within a community) and the roles that different nodes (social media accounts) have with respect to information dissemination. To this end, we devised a methodology for analyzing the structure of the network formed by a target Twitter account and its followers, providing useful insights about the resulting communities and influential nodes.

5.1 Follower graph construction

A graph is a mathematical structure used to represent sets of objects that exhibit pairwise relations between them. More formally, it is an ordered pair $G = (V, E)$ comprising a set of vertices or nodes V together with a set of edges N, which are pairs of vertices (i.e., an

[7] http://bit.ly/2JtT6SY

[8] https://github.com/codelucas/newspaper

edge associates two vertices). In our case, the vertices of the graph correspond to the followers of the target account and their followers, and the edges of the graph correspond to the follow relationships between all these Twitter accounts (i.e., the graph corresponds to the 2-hop neighborhood of the target account). To build such a graph, the necessary data were retrieved from the Twitter API. More specifically, the followers/ids endpoint[9] was used, which returns a collection of user ids for every user following the specified user. The endpoint was first queried using the id of the target account as input, returning the ids of its immediate followers. Then, to retrieve the followers of target's followers, the same endpoint was used treating each of the returned accounts as input to a new query.

Note that data collection is a slow procedure because the Twitter API allows only 60 requests per hour to the followers/ids endpoint. Moreover, to facilitate a more informative insight of the graph, additional queries must be performed to retrieve the account names (e.g. @hack_air) and other Twitter information corresponding to the retrieved account ids (e.g. 4525271542). In particular, the users/lookup endpoint[10] was used, which returns user objects for up to 100 users per request, as specified by comma-separated values passed to the user_id parameter.

One limitation of this approach is that each request in followers/ids endpoint returns up to 5000 random follower ids. Due to Twitter API rate limits we were unable to perform more than one request per account as this would make the procedure even slower. This means that the graph is a sample of the full set of relationships among users; however, it is expected to retain the basic structure of the follower network.

5.2 Community detection

To perform community detection in the graph of target Twitter account followers, we used the Louvain algorithm [3]. This is one of the most widely used algorithms and has been successfully applied to networks of many different types, including social networks sampled from Twitter [9] and LinkedIn [10]). The method is a greedy optimizer of modularity, a measure of network structure that quantifies the quality of a community assignment by measuring how much denser the connections are within communities compared to what they would be in a random network.

Before applying the Louvain algorithm on the target's account follower graph, we first removed leaf-nodes (i.e. Twitter accounts that are connected to only a single account) to simplify the graph and to generate an easier to read visual representation. To perform community detection, we used a Python implementation[11] of the Louvain algorithm.

After detecting communities, we named them after the five most influential nodes (those with higher degree) within each community. In many cases just by inspecting the most influential nodes of each community, it is apparent that nodes in a community have semantic relevance with each other. For example they might represent accounts from a certain country, business and industry, science and research, technological accounts etc.

5.3 Related and influential user discovery

Using information from the graph, we could locate the most influential accounts, i.e. accounts that are followed by many others (and therefore are likely to exert influence on them). To this end, we first filtered nodes based on their in-degree, i.e. the number of incoming edges, keeping only nodes that were followed by at least 100 accounts. Notice that due to the way that the input graph was constructed, all discovered influential users are followers of the target account (as they have higher number of incoming edges compared to their followers, for which no followers were collected). Although discovering influential users and their corresponding communities, out of the existing follower base of target account is useful for tailoring the social media strategies, we were further interested in relevant influential accounts that do not follow the target account (and hence could be the targets of communication efforts with the hope of growing the project follower base).

To find influential users that appear in the graph but do not follow the target account we need information about the number of followers of these users. This information was already retrieved when we queried the user/lookup endpoint in section 5.1. Due to the fact that a large amount of these users may be irrelevant with air pollution topics, we make the following assumption regarding relevance: the relevance of an account is highly dependent on the number of target account followers that this account follows (since these accounts are highly likely to regularly post about air quality issues and news). With this in mind, we first filtered accounts with less than 1000 followers (in that case 1000 was used instead of 100, since the initial set of accounts was much larger and contained much more influential accounts), and then also filtered accounts that followed less than 10% of target account followers. The list of influential users is regularly updated. This list can serve as a valuable resource for communication managers and people who want to target new audiences or plan user engagement strategies. Additionally, can get useful insights about the most dominant communities and create customized engagement campaigns tailored for certain community accounts.

5.4 Visualization

After detecting the communities, visualizations provided by the monitoring tool than can be used to extract useful information. Tool provides visualization of the graph where the layout was determined by the Python implementation[12] of ForceAtlas2 [12] graph layout algorithm. Graph nodes are colored according to community assignment and their size is proportional to their edge degree (i.e. total number of incoming edges). For the visualization of the graph sigma.js[13] library is used. This visualization service can enable users to locate key accounts in each of the communities, get Twitter insights and check the graph progress through time. Users can click on the nodes of the graph and inspect the account information, its community, and its connections to other communities. There is also the possibility to locate nodes with high incoming degrees (they appear larger) and discover influential accounts for every community. Another valuable functionality is that the user can inspect the progress of the graph in time. This can provide strong

[9] https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/api-reference/get-followers-ids

[10] https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/api-reference/get-users-lookup

[11] https://github.com/taynaud/python-louvain

[12] https://github.com/bhargavchippada/forceatlas2

[13] http://sigmajs.org/

insights about how communities behave and what are the impacts on them as the network of target account is evolving.

6 EMPIRICAL STUDY

In this section we demonstrate some of the results and visualizations of the monitoring tool extensions we described in the previous sections, using as target account the Twitter account of the hackAIR project (@hack_air)

6.1 @hack_air relevant account discovery

Using the methodology we described in section 3 we create a list with air quality/pollution related accounts. The service[14] that provides the list of hackAIR relevant accounts is shown in hackAIR-relevant tab and is updated on weekly basis. This list provides useful information about the popularity and activity of the related accounts and additional information about the language, locations and region (EU/non EU) of each account. By the time this paper was written, the most recent list that was retrieved by the service was at 8 of May 2019. The service had discovered 389 active air quality/pollution related users. 340 were extracted from Twitter and 49 from YouTube. 101 of them were located in EU region, 144 out of EU and 144 were lacking location information. Among the most prominent languages between users were the English language with 325, french with 20, German with 14, dutch with 8 and Norwegian with 5 users.

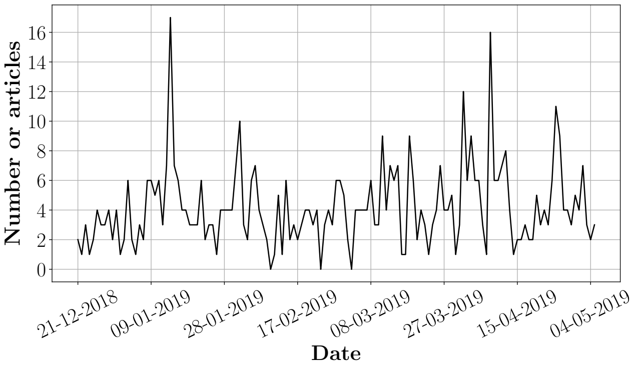


Figure 3: Daily related article distribution

6.2 Discovery of newsworthy articles

Following the methodology of Section 4, we create an RSS feed[15] with valuable articles for the hackAIR project which contain the latest news, developments and trends in the area of air quality, research, technology, activism and policy. The service is running on daily basis. Figure 3 shows the daily distribution of related articles retrieved by the hackAIR-newsworthy article discovery service. The daily average number or related news is five articles.

Table 5 presents a list of relevant examples that demonstrate a sample of useful and worth sharing articles that have been retrieved using the article discovery service.

Title	Summary
Air pollution particles found in mothers' placentas.	Scientists have found that particles of air pollution travel through pregnant women's lungs and lodge in their placentas.
Illegal pollution is worse than we thought.	The government has admitted that pollution is much worse in some areas than previously believed.
The school where the children can taste pollution.	Greater Manchester pollution levels are so bad that teachers can no longer open windows in the classroom.
20,000 Belgians became air pollution scientists.	The biggest ever citizen investigation into air pollution has produced some interesting results about air quality in Europe.

Table 5: Related article examples

6.3 @hack_air audience analysis

To provide additional useful insights for the communication managers of hackAIR project, that offer support for designing and executing their dissemination and engagement strategy we employ the methodology discussed in Section 5.

Initially we construct the follower graph. To give an indication about the size of the resulting network, when this procedure was first applied in mid-July 2018, it amounted to a total of 331,691 follow relationships (graph edges) between 268,123 distinct accounts (graph vertices), 374 of which are immediate followers of hackAIR Twitter account (@hack_air). Then we perform community detection in the graph. As we described in the previous subsection discovered communities are named using the names of their most influential users. For the hackAIR case, we manually name the discovered communities by inspecting the most influential nodes. These nodes exhibit semantic relationships in each community. The following 10 relevant communities of Twitter accounts were identified: 1) science-innovation, 2) smart-green cities, 3) technology, open data and sensors A, 4) technology, open data and sensors B, 5) UK and general air pollution accounts, 6) German accounts, 7) Belgian and Norwegian accounts, 8) French accounts, 9) Irish accounts, 10) Greek accounts.

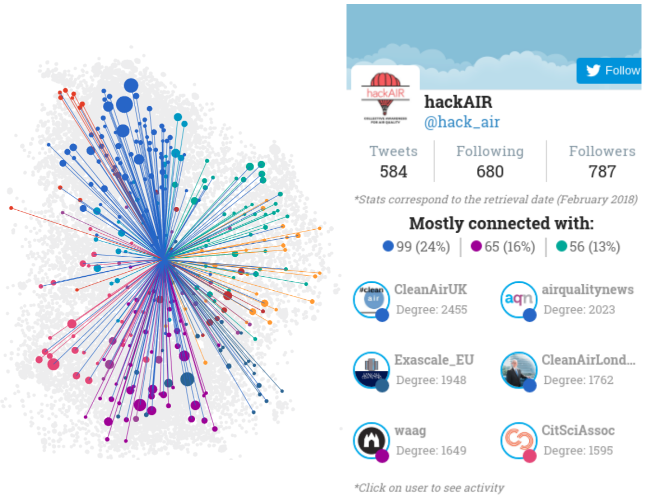
This service[16] also that provides the list of influential users discovered with procedure described in Section 5.3 in Twitter influential tab. The user can sort the accounts based on their number of followers (influential) or based on how many @hack_air followers it follows (relevant). There is also information about whether this account already follows the @hack_air account or not. This process is executed weekly and can further assist pilot coordinators and local communication managers to identify ambassadors on a regular basis.

Furthermore, the visualization service can enable users to locate key accounts in each of the communities, get Twitter insights and check the graph progress through time. Figure 4 illustrates the visualization service. Users can click on the nodes of the graph and inspect the account information, its community, and its connections to other communities. There is also the possibility to locate nodes with high incoming degrees (they appear larger) and discover influential accounts for every community.

[14] http://hackair-mklab.iti.gr/table/

[15] http://hackair-mklab.iti.gr/feed/rss-aq.xml

[16] http://hackair-mklab.iti.gr/table/



584 680 787

*Stats correspond to the retrieval date (February 2018)

Mostly connected with:

● 99 (24%) | ● 65 (16%) | ● 56 (13%)

CleanAirUK Degree: 2455

airqualitynews Degree: 2023

Exascale_EU Degree: 1948

CleanAirLond... Degree: 1762

waag Degree: 1649

CitSciAssoc Degree: 1595

*Click on user to see activity

Figure 4: Network and community visualization.

7 CONCLUSIONS

We presented three novel tools that could be used either as extensions on an existing open-source social media monitoring framework, or as standalone services. These tools aim to assist the communication and dissemination needs of a certain Twitter account that is related with the air quality domain. Specifically these tools are focused on effectively discovering key social media accounts, keeping up-to-date with trending news in the area of air quality, and provide visualizations of the social media audience and communities and influential accounts. Using as target account the Twitter account of the hackAIR project (@hack_air) we showed that these three tools provide valuable insights for social media account managers. These insights could potentially contribute to planning a dissemination strategy, open communication channels and engagement with related accounts, spread newsworthy articles and increase the follower base of the target account. Furthermore, these tools are possible to apply to other air quality related research projects and Twitter accounts to utilize the power of social media.

ACKNOWLEDGMENTS

This work is partially funded by the European Commission under the contract number H2020-688363 hackAIR.

REFERENCES

[1] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, and Alejandro Jaimes. 2013. Sensing trending topics in Twitter. *IEEE Transactions on Multimedia* 15, 6 (2013), 1268–1282.

[2] Mohamed Bakillah, Ren-Yu Li, and Steve HL Liang. 2015. Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: the case study of typhoon Haiyan. *International Journal of Geographical Information Science* 29, 2 (2015), 258–279.

[3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[4] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 27.

[5] Polychronis Charitidis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Twitter-Based Sensing of City-Level Air Quality. In *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 1–5.

[6] Weiguo Fan and Michael D Gordon. 2014. The Power of Social Media Analytics. *Commun. ACM* 57, 6 (2014), 74–81.

[7] Aleksandr Farseev, Ivan Samborskii, and Tat-Seng Chua. 2016. bbridge: A big data platform for social multimedia analytics. In *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 759–761.

[8] Ashish Goel, Aneesh Sharma, Dong Wang, and Zhijun Yin. 2013. Discovering similar users on twitter.

[9] Przemyslaw A Grabowicz, José J Ramasco, Esteban Moro, Josep M Pujol, and Victor M Eguiluz. 2012. Social features of online networks: The strength of intermediary ties in online social media. *PloS one* 7, 1 (2012), e29358.

[10] Jonathan Haynes and Igor Perisic. 2009. Mapping search relevance to social networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*. ACM, 2.

[11] Yin Huang, Han Dong, Yelena Yesha, and Shujia Zhou. 2014. A scalable system for community discovery in twitter during hurricane sandy. In *2014 14th IEEE/Acm International Symposium on Cluster, Cloud and Grid Computing*. IEEE, 893–899.

[12] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one* 9, 6 (2014), e98679.

[13] Nont Kanungsukkasem and Teerapong Leelanupab. 2016. Power of crowdsourcing in Twitter to find similar/related users. In *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE, 1–6.

[14] Hwi-Gang Kim, Seongjoo Lee, and Sunghyon Kyeong. 2013. Discovering hot topics using Twitter streaming data social topic detection and geographic clustering. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. IEEE, 1215–1220.

[15] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2017. Geotagging text content with language models and feature mining. *Proc. IEEE* 105, 10 (2017), 1971–1986.

[16] Xiaoming Liu, Chao Shen, Xiaohong Guan, and Yadong Zhou. 2018. We know who you are: Discovering similar groups across multiple social networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 99 (2018), 1–12.

[17] Y. Mejova, I. Weber, and M.W. Macy. 2015. *Twitter: A Digital Socioscope*. Cambridge University Press. https://books.google.gr/books?id=ODfBjwEACAAJ

[18] Symeon Papadopoulos, Yiannis Kompatsiaris, Athena Vakali, and Ploutarchos Spyridonos. 2012. Community detection in social media. *Data Mining and Knowledge Discovery* 24, 3 (2012), 515–554.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[20] Gerasimos Razis and Ioannis Anagnostopoulos. 2016. Discovering similar Twitter accounts using semantics. *Engineering Applications of Artificial Intelligence* 51 (2016), 37–49.

[21] Manos Schinas, Symeon Papadopoulos, Lazaros Apostolidis, Yiannis Kompatsiaris, and Pericles A Mitkas. 2017. Open-Source Monitoring, Search and Analytics Over Social Media. In *International Conference on Internet Science*. Springer, 361–369.

[22] Stefan Stieglitz and Linh Dang-Xuan. 2013. Social media and political communication: a social media analytics framework. *Social network analysis and mining* 3, 4 (2013), 1277–1291.

[23] Didi Surian, Dat Quoc Nguyen, Georgina Kennedy, Mark Johnson, Enrico Coiera, and Adam G Dunn. 2016. Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. *Journal of medical Internet research* 18, 8 (2016), e232.

[24] Carmen K Vaca, Amin Mantrach, Alejandro Jaimes, and Marco Saerens. 2014. A time-based collective factorization for topic discovery and monitoring in news. In *Proceedings of the 23rd international conference on World wide web*. ACM, 527–538.

[25] Eleni Vathi, Georgios Siolas, and Andreas Stafylopatis. 2015. Mining interesting topics in Twitter communities. In *Computational Collective Intelligence*. Springer, 123–132.

[26] Jingjing Wang, Wenzhu Tong, Hongkun Yu, Min Li, Xiuli Ma, Haoyan Cai, Tim Hanratty, and Jiawei Han. 2015. Mining multi-aspect reflection of news events in twitter: Discovery, linking and presentation. In *2015 IEEE International Conference on Data Mining*. IEEE, 429–438.

[27] Hong Wei, Jagan Sankaranarayanan, and Hanan Samet. 2017. Finding and tracking local Twitter users for news detection. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 64.

[28] Xiangye Xiao, Yu Zheng, Qiong Luo, and Xing Xie. 2014. Inferring social ties between users with human location history. *Journal of Ambient Intelligence and Humanized Computing* 5, 1 (2014), 3–19.

[29] Velissarios Zamparas, Andreas Kanavos, and Christos Makris. 2015. Real time analytics for measuring user influence on twitter. In *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 591–597.