Dictionary-based Classification of Tweets About Environment

Michela Cameletti
University of Bergamo
Silvia Fabris
University of Bergamo
Stephan Schlosser
Georg-August-Universitat Gottingen Universitatsmedizin
Daniele Toninelli (daniele.toninelli@unibg.it)
University of Bergamo https://orcid.org/0000-0002-3158-1982

Research

Keywords: tweet filtering, big data analysis, dictionary-based selection, dictionary-based search, unsupervised algorithm

Posted Date: August 27th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-51088/v2

Dictionary-based Classification of Tweets About Environment

Michela Cameletti[1], Silvia Fabris[1], Stephan Schlosser[2], Daniele Toninelli[1]

[1] *University of Bergamo, Dept. of Management, Economics and Quantitative Methods, Bergamo, Italy.*

[2] *University of Göttingen, Center of Methods in Social Sciences, Göttingen, Germany.*

Corresponding author: Daniele Toninelli (daniele.toninelli@unibg.it).

Abstract

In the era of social media, the huge availability of digital data (e.g. posts sent through social networks or unstructured data scraped from websites) allows to develop new types of research in a wide range of fields. These types of data are characterized by some advantages such as reduced collection costs, short retrieval times and production of almost real-time outputs. Nevertheless, their collection and analysis can be challenging. For example, particular approaches are required for the selection of posts related to specific topics; moreover, retrieving the information we are interested in inside Twitter posts can be a difficult task.

The main aim of this paper is to propose an unsupervised dictionary-based method to filter tweets related to a specific topic, i.e. environment. We start from the tweets sent by a selection of Official Social Accounts clearly linked with the subject of interest. Then, a list of keywords is identified in order to set a topic-oriented dictionary. We test the performance of our method by applying the dictionary to more than 54 million geolocated tweets posted in Great Britain between January and May 2019.

Keywords: *tweet filtering; big data analysis; dictionary-based selection; dictionary-based search; unsupervised algorithm.*

Declarations

Availability of data and material: The datasets generated and analyzed and the R code used are available in the github repository, https://github.com/silviafabris/Twitter_dictionary_based_classification.

Competing interests: The authors declare that they have no competing interests.

Funding: Work supported by the University of Bergamo (60% University Funds and project "STaRs - Azione 3: Outgoing Visiting Professor 2019").

Corresponding author: Daniele Toninelli (daniele.toninelli@unibg.it; ORCID: 0000-0002-3158-1982).

Other authors' ORCID: Michela Cameletti: 0000-0002-6502-7779; Stephan Schlosser: 0000-0001-9805-4485.

Authors' contributions: This is a joint work made by the contribution of all authors. All authors read and approved the final manuscript.
Acknowledgements: Not applicable.

List of abbreviations

GB: Great Britain

OSA: Official Social Accounts

SVM: Support Vector Machine

TP: number of true positive

FP: number of false positive

TN: number of true negative

FN: number of false negative

AC: accuracy

SE: sensitivity (or recall)

SP: specificity

PR: precision

1 Introduction

Despite being born as communication tools, nowadays social media and microblogging platforms have become a common source of information. These sources nowadays can be used to develop and/or to support scientific research in a wide range of fields. The collection and the analysis of such data is still an evolving and very promising research field. There are clearly some advantages: for example, data obtained from the Internet are available at lower costs, in shorter times and are easier to be collected than data obtained by means of traditional surveys. Nevertheless, researchers cannot control the data collection phase: if questionnaire-based surveys ask specific answers connected to particular research purposes, microblogging texts or web scraped data rely on "listening" and "measuring" what it is available. Moreover, since no sampling strategy drives the data collection, the information obtained from microblogging platforms or from the web could be non-representative of the population of interest.
Twitter is one of the most relevant and used microblogging platform worldwide, with 321 million active users in February 2019[3]. Using this platform, it is possible to send short messages (up to 280 characters), to resend (i.e. to retweet) them or to like or comment posts shared by other users. Thanks to the wide and still increasing spread of Twitter, a huge amount of data is produced daily. This information can be used for developing various types of research on a wide range of topics, such as epidemiology (Ahmed et al. 2019), politics (Budiharto and Meiliana 2019) or well-being and happiness (Mitchell et al. 2013; Baylis et.al 2018). However, retrieving and processing Twitter data can be extremely

[3] Source: https://en.wikipedia.org/wiki/Twitter (latest access on September 3rd, 2019).

challenging. Shared posts can be about personal opinions, ideas, goals and events, but they also include advertisements and news. Consequently, the identification and the selection of tweets regarding a specific topic is a difficult task. This is mainly due to the completely unstructured nature and to the limited length of posts. There are no precise rules about what to post and how to share information: e.g., users can write plain text or use hashtags to refer to specific topics, to express opinions about an event or to highlight a theme or a fact. Furthermore, news media, government organizations, ONGs, no profit associations, industries and so forth share posts containing information related to their activity, news or advertising.

The purpose of this paper is to propose and test a dictionary-based method to filter tweets concerning a specific topic. We focus particularly on environment, given its societal importance and its relevance in driving future governmental and cross-national policies. Moreover, as shown in Toninelli et al. (2018) environment is one of the most emerging and important latent dimensions linked to the citizens' personal well-being. This topic is also very challenging, from our point of view, because in all its facets (e.g. climate change, recycling, renewable energy, global warming), it is associated to a language that changes constantly and evolves quickly[4]. For this reason, it is necessary to constantly update the dictionary used to retrieve or select tweets regarding this theme.

The approach we propose sets up a dictionary starting from a list of keywords obtained by analyzing tweets published by a selected list of Official Social Accounts (OSA) whose activity is strictly related to environment. We test the performance of this method using 54,135,006 tweets posted in Great Britain (GB) between 2019/01/14 and 2019/05/13. These tweets are fully geolocated, given the adopted collection method, based on the "theory of circles" (Schlosser et al. 2019). As a consequence, it would be possible to study, for example, the spatial variability of the sentiment or the inclination for a certain topic across the country's sub-areas.

In order to evaluate the capability of our method in filtering tweets, we compute some indicators (accuracy, sensitivity, specificity, precision, F1 score; for further details, see sect. 5). All of these measures confirm the good performance obtained by retrieving environment tweets using our algorithm. This represents an important contribution in enhancing the literature about how to select/filter messages sent through social media about a specific topic. First, we propose a flexible method that can be easily applied to any topic of interest. Second, instead of using pre-set and already-available dictionaries, which may have issues of adaptability and/or do not fit perfectly with the studied topic, the user can create an ad-hoc and personalized dictionary for filtering and selecting tweets linked to a specific theme. Third, we do not relay on a single or on a short list of predefined keywords (that can be too general or not completely focused on the topic), but we rather allow the researcher to expand as necessary the list of keywords and hashtags that can be updated and renewed any time is needed, facing the issue of the frequent changes that characterize the language of a topic such as the environment. Finally, our method relies on a dictionary that is not based on just single words, but on combinations of words (i.e. on bigrams and trigrams), thus

[4] Source: https://www.theguardian.com/environment/2019/may/17/why-the-guardian-is-changing-the-language-it-uses-about-the-environment (latest access on September 3rd, 2019).

reducing the inclusion of non-pertinent tweets. Moreover, the inclusion of hashtags enriches considerably the dictionary.
This paper is structured as follows: in Sect. 2 we frame our paper within the existing literature and we review the main works related to the classification and the tweets filtering; Sect. 3 describes the analyzed data; Sect. 4 introduces the algorithm set to build the dictionary and to filter tweets; Sect. 5 discusses the results of the quality metrics we used for testing our algorithm using the GB tweets; Sect. 6 discuss our main findings and provides ideas for further studies.

2 Related Works

In order to classify and filter tweets by their content, two main methodological approaches have been proposed in the literature when categories are known: dictionary-based and supervised methods (Grimmer and Stewart 2013).
The first approach filters tweets about a specific topic by using a set of keywords defining a dictionary. For example, Cody et al. (2015) explore climate change sentiment by selecting tweets containing at first the word "climate", and related expressions such as "global warming", "climate realist", "climate change" and "anthropogenic global warming". It is also possible to perform the selection of tweets by considering specific hashtags related to the topic of interest. In this regards, Reyes-Menendez et al. (2018) study the opinion about environment by selecting posts containing the hashtag #WorldEnvironmentDay. A similar approach has been used by Pruss et al. (2019). For filtering tweets regarding the Zika infection, they first use the keywords "zika" and "ZIKAV" to find posts; then, they apply a topic model like the Latent Dirichlet Allocation to find the most popular sub-topics (e.g. environmental concerns, vaccination, emergency declaration). These strategies based on keywords and hashtags are very simple and easy to be applied; nevertheless, they do not guarantee to identify all the tweets connected to the specific argument. At the same time a large set of keywords or hashtags can lead to the inclusion of posts not strictly related to the topic of interest.
Supervised methods, instead, require the human intervention to create a (large) dataset of labeled tweets (i.e. classified into a predefined set of categories) that will be used for training (or supervising) a statistical model or a machine learning algorithm. This estimated classifier is then adopted for predicting the category of a new tweet. This approach is used for example by Frenda et al. (2019), who adopt a common method in machine learning as Support Vector Machine (SVM) to automatically detect sexist and misogyny on Twitter. In particular, this study uses, as training dataset, the freely available corpora known as "Automatic Misogyny Identification – IBEREVAL 2018"[5] and considers, as input variables, some lexical and stylistic features of the post. The accuracy of such a classifier is equal to 76%. Similarly, Foucault et al. (2016) combine SVM and a Naïve Bayes method to classify tweets sent from French institutions into four communication categories (sharing experience, promoting participation, interacting with the community, and promoting-informing about the institution). Each tweet is represented by 18 features derived from metadata information, punctuation marks, tweet-specific characteristics (e.g. use of hashtags and emoticons) and lexical features. By means of cross-validation, the classification performance is evaluated

[5] Source: https://amiibereval2018.wordpress.com/ (latest access on 2019, September 3rd).

obtaining a value of the F1 index[6] equal to 72%. This kind of approach is also known as "feature-based modeling", because it requires to extract textual features from the tweets to be successively provided as input to the machine learning algorithms. An extension of this supervised approach is represented by deep learning methods. They employ neural networks and usually outperform feature-based models. Nizzoli et al. (2019), for example, analyze extremist propaganda and try to identify pro-ISIS tweets. In particular, they show that a Recurrent-Convolutional Neural Network with pre-trained word embedding is able to reach a F1 score of 0.9. Stowe et al. (2019) adopt SVM and linguistic features to identify tweets related to hurricane events. As a comparison, they implement also two deep learning methods (Multi-layered Perceptron and Convolutional Neural Network); they find that the Multi-layered Perceptron performs better with a F1 score equal to 0.83. They also run feature-based algorithms (logistic regression, SVM and Naïve Bayes algorithm) with linguistic, temporal and geospatial features to predict people behavior during hurricane events. In this case SVM provides the best performance with F1 values included between 0.47 and 0.79 according to the considered features.
The method we propose in this paper is unsupervised and dictionary-based. Differently from Cody et al. (2015) and Pruss et al. (2019), it builds a dictionary that includes the most common bigrams[7], trigrams[8] and hashtags about environment without needing a starting set of keywords. Only a list of OSA has to be provided in advance and a limited number of human checks are needed in order to avoid the inclusion of too general keywords or acronyms that would lead to the selection of tweets not strictly pertaining to environment (or to the chosen topic). Thus, our approach minimizes the amount of required human work, because it doesn't need the set of labeled tweets for training (as required by supervised approaches) or a predefined set of keywords (that could be too much general or not completely focused on the studied topic). At the same time, thanks to the arbitrary selection of the OSA and to the possibility of reviewing step by step the dictionary creation, it is very flexible and could be applied to and personalized for any topic of interest.

3 Data Collection and Preliminary Cleaning

In the following subsections we describe the two datasets used for the analysis. The first one (Sect. 3.1) is composed by a sample of tweets posted by OSA related to the analyzed topic, environment. Starting from these data, the algorithm sets up the dictionary. The latter is then applied to the second dataset (Sect. 3.2), composed by the tweets posted in GB between 2019/01/14 and 2019/05/13. The algorithm has been implemented using the R software[9].

[6] The F1 score is a performance index depending on precision and recall (See Sect. 5 for its definition).
[7] A bigram is a pair of consecutive words commonly associated each other.
[8] A trigram is a group of three consecutive words commonly associated one to another.
[9] The code and the data are available at the following link: https://github.com/silviafabris/Twitter_dictionary_based_classification

3.1 Tweets from OSA

The general idea behind the tweet selection is that posts speaking about the same topic should be similar and different from tweets related to other themes. As a consequence, tweets pertaining to a certain topic should generally include similar words or combination of words. Our work aims at detecting and studying posts about environment. For this purpose, our preliminary objective is to set up a dictionary including the most common and relevant keywords related to such a topic. As first step, we identified 12 OSA linked to environment. In particular, we chose verified accounts[10] (or profiles that have at least 10,000 followers) belonging to no-profit associations, research institutes and intergovernmental organizations whose activity is related to environment[11]. The OSA selection is an arbitrary phase of the algorithm. The chosen accounts are selected because of their popularity and with the aim of covering all the possible aspects of environment (e.g. climate change, plastic pollution, nature protection). Note that, as it will be described in Sect. 4.2, the OSA choice can cause effects on the final dictionary. For each account, we retrieved all the most recent posted or retweeted tweets that Twitter leads us to download up to 2019/05/10. Among the obtained 38,611 tweets, we kept exclusively posts written in English. Then, we cleaned their corpus by removing url links, html codes, non-ascii and special characters, but we kept hashtags. This list of cleaned tweets is our first dataset. In Sect. 4.1 we analyze this dataset, in order to detect the most recurrent expressions (i.e. bigrams, trigrams) and hashtags used by the considered OSA.

3.2 Tweets from GB

A second dataset is used in order to test and apply the dictionary built with the selected keywords starting from the first dataset introduced in sect. 3.1. This second database includes all the tweets sent in GB from 2019/01/14 to 2019/05/13 (i.e. for a total of 120 days). The tweets are collected through the "theory of circles" method, described in Schlosser et al. (2019) and further tested in ###CARMA2020####. Generally, just the 1-2% of tweets contains GPS coordinates[12]. Nonetheless, the "circle approach" allows us to geolocate all tweets directly in the collection phase, associating each post to one of the NUTS-1[13] sub-areas covering GB (see Figure 3, left).
After having preliminary removed messages sent by bots[14], we obtained 54,135,006 tweets, that corresponds to an average of 4,921,364 tweets for each NUTS area.

---

[10] https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts (latest access on September 3rd, 2019).
[11] @climateprogress (Climate Progress), @ClimateReality (Climate Reality), @friends_earth (Friends of the Earth), @Greenpeace (Greenpeace), @GreenpeaceUK (Greenpeace UK), @LessPlasticUK (Less Plastic), @PlasticPollutes (Plastic Pollutes), @UNEnvironment (UN Environment Programme), @UNFCCC (UN Climate Change), @World_Wildlife (World Wildlife Fund), @WWF (WWF), @WWFScotland (WWF Scotland).
[12] https://developer.twitter.com/en/docs/tutorials/tweet-geo-metadata.html (last access on August 2nd, 2019)
[13] Source: https://ec.europa.eu/eurostat/web/nuts/background (latest access on September 3rd, 2019).
[14] A bot is an automated program which interacts automatically on the social network.

The next step consisted in cleaning the tweets' corpus. In doing so, we tried to keep as much information as possible by replacing htmls, emojis and slangs with equivalent-meaning expressions. For example, the Unicode character "\U0001f602", corresponding to the emoji  , was translated with "face with tears of joy". Note that we kept hashtags in the text because they are crucial in detecting tweets related to the specific topic. The cleaned data of this second dataset are then used in order to compute some performance indexes for our dictionary (see Sect. 5.1).

4 Methods

After having cleaned the posts sent by the selected OSA (Sect. 3.1), we use them in order to set up the dictionary (the first dataset, introduced in 3.1). The steps of the algorithm for the dictionary definition are explained in Sect. 4.1 and summarized in the flow charts of Figure 1 and 2. The final dictionary is composed by a set of selected bigrams, trigrams and hashtags and is applied to the full set of GB tweets described in Sect. 3.2 (the second dataset) in order to select tweets regarding the chosen topic.

4.1 Selection and cleaning of bigrams, trigrams and hashtags from OSA tweets

Given the tweets collected from the selected OSA and preprocessed (see Sect. 3.1), we produce the list of all bigrams and trigrams (also named *expressions* in the following) with the corresponding frequencies (see Table 1). This represents the starting point of the dictionary creation (step *a* in Figure 1). Expressions which do not appear frequently are usually not related to the topic or are too general to be included in the final dictionary. For this reason, in order to select the most pertaining bigrams and trigrams (that will be later used to define the dictionary), some additional steps are required. After having looked at the full list of bigrams and trigrams not containing stop words[15], it is possible to proceed directly with the definition of some thresholds for the frequencies in order to exclude expressions which do not occur often (step *c* in Figure 1). In our case, we exclude bigrams which appear less than 65 times and trigrams posted less than 35 times (see the non-grey expressions in Table 1). However, the resulting list of expressions could still include bigrams and trigrams not related to environment (e.g. common or general expressions, country and state names). Consequently, an additional cleaning stage is required (step *b* in Figure 1). In particular, the algorithm proposes the following possibilities:
a *standard* cleaning (step *d* in Figure 1) to be performed after the choice of the frequency thresholds (step *c* in Figure 1): the user reviews the expressions selected after applying the thresholds and remove any expressions not strictly related to the topic;
an optional *extra* cleaning stage before the choice of the thresholds (step *b* in Figure 1). The aim is to remove, from the list of selected OSA expressions, some common terms which are widely used in Twitter and very likely not related to environment. Even if this step is optional, we highly suggest to use it, because it reduces the standard review process performed at step *d* in Figure 1. The extra

[15] Stop words are the most common words used in a language (such as for example "the", "a", "an", "in"). In this case the list of stop words is given by three different lexicons ("onix", "SMART" and "snowball").

cleaning considers the full set of GB tweets described in Sect. 3.2 to identify the list of *general expressions*, i.e. popular bigrams and trigrams (step *a* in Figure 2). These recurrent expressions are used to remove from the OSA bigrams and trigrams list general expressions such as "trump administration", "taking action", "million people". It is important to note that this procedure can be performed by using the full set of GB tweets or a smaller sample, in order to reduce the computational time. Our empirical experience with our case study demonstrates that the final dictionary does not change considerably by using different samples or the complete dataset of GB tweets. For this reason, we decided to use a random sample of 3.5 million tweets collected between March 10th and May 13th, 2019. We arbitrary decided for a very low threshold: in the list of general expressions we take into account just bigrams and trigrams tweeted at least 20 times. This way, we obtained 30,656 general expressions (step *a* in Figure 2). However, this vector of recurrent bigrams and trigrams may contain expressions linked to environment, such as "climate change", which we do not obviously want to be part of the list (otherwise they will be not included in the dictionary). Thus, a review of the list of general expressions is necessary. This can be done by adopting one of the following two approaches:
user-based approach (step *b* in Figure 2): the user examines all the general expressions one by one and remove the ones related to environment;
list-based approach (step *c* in Figure 2)*:* in this case we assume that a set of expressions related to environment is available (prepared ad hoc by the researcher or taken from existing dictionaries). The two lists will be matched and the environment-related expressions will be removed from the set of general expressions.

In this research we adopted the list-based approach by considering a list of 49 expressions[16] prepared specifically for our case study. After removing from the list of general expressions the terms related to environment, the resulting vector is composed by 30,632 expressions. It is then possible to proceed with the extra cleaning of the OSA bigrams and trigrams by removing all the terms included in the set of 30,632 common expressions (step *b* in Figure 1). Moreover, in this same step, all the bigrams and trigrams that contain country names and USA state names are removed. The extra cleaning step removed 7 bigrams and no trigrams (see the red expressions in Table 1). Finally, after the *extra* cleaning, the *standard* cleaning (step *d* in Figure 1) is used to review the new list of OSA expressions in order to exclude other terms not related to the studied topic, such as "start donating" or "coral reefs" (see blue expressions in Table 1). For our application

[16] air clean, air pollution, air quality, carbon emissions, carbon pollution, clean air, clean energy, climate action, climate change, climate conference, climate crisis, climate reality, climate science, climate solutions, coal ash, coal plants, coalfired power, conference cop, environmental laws, extreme weather, food waste, fossil fuel, fossil fuels, fuel industry, gas drilling, gas emissions, gas industry, global climate, global temperatures, global warming, greenhouse gas, healthy environment, offshore drilling, palm oil, paris agreement, plastic bags, plastic bottles, plastic packaging, plastic pollution, plastic straws, plastic waste, renewable energy, singleuse plastic, singleuse plastics, tar sands, toxic chemicals, toxic pesticides, warming world, weather events.

this standard review step removed 10 bigrams and 1 trigram. As result, we obtain the final list of bigrams and trigrams related to the topic.

FIGURE 1 and FIGURE 2 here (now at the end of the file)

After the previous phase, the algorithm gives the possibility to enrich the dictionary with hashtags (step *e* of Figure 1), which are normally used by Twitter users to identify and categorize tweets. In our case we decided to include hashtags because they can be extremely useful in filtering a tweet linked to environment. First of all, we analyze the hashtags used by the selected OSA accounts (see Table 2). They represent either popular themes on Twitter (i.e. trends), slogans used in the OSA description (e.g. #UseLessPlastic for the @LessPlasticUK account) or they are created by OSA for particular international events (e.g. #PlasticFreeFriday). As done previously for OSA expressions, we need to remove hashtags too general (such as "#nature") and referring to countries or to states. Even in this case, it is possible to adopt the *standard* review only (step *h* of Figure 1), or to also use an *extra* cleaning (step *f* of Figure 1). For this purpose, we create a list of general (i.e. popular) hashtags by using a sample of the GB tweets (as described in Sect. 3.2). After removing the general hashtags (see red terms in Table 2), we selected the 60 most popular OSA hashtags (step *g* of Figure 1). Finally, a last standard review (step *h* of Figure 1) is implemented with the aim of excluding the hashtags that are too generic to be part of the dictionary (such as "#climate" or acronyms like "#dyk", i.e. "Do You Know"). These excluded hashtags are reported in blue in Table 2.

The final dictionary is composed by 35 OSA expressions (listed in Table 1) and by 52 hashtags (see Table 2). We apply this dictionary to over 54 million tweets collected in GB in order to select only tweets that contain at least one expression included in the final dictionary. As a result, we obtained 107,176 tweets related to environment.

4.2 The effect of the number of considered OSA

The number of OSA considered for defining the dictionary can be arbitrary chosen by the user, as described in Sect. 4.1. In order to assess whether there is an effect of the number of selected OSA on the final list of expressions defining the dictionary, we compare the list of bigrams/trigrams obtained using the 12 OSA described in Sect. 4.1 (named LIST 1, in the following) with the one obtained by using 22 OSA (the previous 12 OSA plus 10 new ones17, named LIST 2). With LIST 2 the number of OSA tweets increases from 38,611 to 57,029. Moreover, by applying the extra cleaning and keeping the same thresholds set in Sect. 4.1, we obtain a larger final list of expressions composed by 74 bigrams/trigrams (instead

[17]@climateprogress, @ClimateReality, @friends_earth, @Greenpeace, @GreenpeaceUK, @LessPlasticUK, @PlasticPollutes, @UNEnvironment, @UNFCCC, @World_Wildlife, @WWF, @WWFScotland, @NRDC, @nature_org, @EnvDefenseFund, @Earthjustice, @foe_us, @guardianeco, @HuffPostGreen, @insideclimate, @PlanetGreen, @ClimateCentral.

of 35). This means that increasing the numbers of OSA leads to a larger set of expressions but also to a more demanding standard cleaning step. As a consequence, we suggest to keep the number of OSA between 10 and 15 in order to avoid unnecessary cleaning to remove expressions which are too generic and not strictly related to the topic of interest and are highly time consuming. Moreover, by comparing the final expression list obtained with LIST 1 and LIST 2, it can be observed that: *i)* there is just 1 expression in LIST 1 which does not appear in LIST 2; *ii)* there are 40 expressions contained in LIST 2 which are not included in LIST 1; *iii)* considering the 20 most recurrent expressions, the two lists differ by just 6 terms (3 are contained in LIST1 and not in LIST2 and, contrarily, 3 are included in LIST2 and not in LIST 1). On the basis of these results, we can conclude that, even if LIST2 gives rise to a larger set of expressions, by looking at the most frequent terms the two lists are almost equal.

TABLE 1 and TABLE 2 here (now at the end of the file)

5 Results: dictionary performance

In order to evaluate the performance of the dictionary-based filtering, we randomly choose 600 tweets selected and 600 not selected by the algorithm (i.e. classified as not linked to "environment"). Then, we manually classify these posts into two categories: "related" and "non-related" to environment. This allows us to compute the following relevant quantities, which can be collected in the confusion matrix reported in Table 3:
number of **true positive** (TP), i.e. number of tweets correctly classified by the algorithm as related to environment;
number of **false positive** (FP), i.e. number of tweets wrongly classified by the algorithm as related to environment;
number of **true negative** (TN), i.e. number of tweets correctly excluded by the algorithm because not linked to environment;
number of **false negative** (FN), i.e. number of tweets wrongly classified by the algorithm as not pertaining the environment.

TABLE 3 here (now at the end of the file)

The algorithm performance has been evaluated through the following indexes, based on the confusion matrix shown in Table 3:

a. **accuracy** (AC): proportion of tweets correctly classified by the algorithm on the total number of tweets processed:

$$AC = \frac{TP + TN}{TP + TN + FP + FN}; \qquad 1$$

b. **sensitivity** (SE): proportion of tweets pertaining to the argument that are correctly filtered by the algorithm:

$$SE = \frac{TP}{TP + FN}; 2$$

c. **specificity** (SP): proportion of tweets not related to environment which are correctly excluded by the algorithm:

$$SP = \frac{TN}{TN + FP}; 3$$

d. **precision** (PR), i.e. the proportion of tweets truly related to the topic among tweets classified by the algorithm as pertaining the chosen topic:

$$PR = \frac{TP}{TP + FP}; \qquad 4$$

e. **F_1 score** (F1) defined as a function of PR and SE and given by:

$$F_1 = 2\frac{PR \cdot SE}{PR + SE}. \qquad 5$$

For all the indicators the range is between 0 and 1 and the "the higher the better" rule holds. Results, here, are expressed in percentage terms.
AC can be used as a first overall measure to evaluate the classification algorithm performance, taking into account tweets correctly classified on the total number of posts. In particular, following this criterion, our method is able to classify correctly 98.42% of the total number of tweets. This measure of performance is quite high, if compared, for example, with the accuracy obtained by the Automatic Misogyny Identification – IBEREVAL 2018 classifier (equal to 76%). However, AC, used by itself, can be misleading, especially when there is a severe class imbalance in the classification problem.
A more informative evaluation is obtained by using AC together with SE (also known as recall), which represents the ability of the algorithm in correctly selecting all the tweets concerning environment on the total number of tweets linked to the topic. The SE value is even higher than the AC measure: 99.32% of the tweets truly related to environment are identified as relevant by the algorithm. Just 0.68% of the tweets not linked to environment are wrongly selected by our algorithm, that represents a very small percentage. However, even a high value of SE could hide a problematic situation. This happens when, for example, the number of false positives is high.
In such cases, it is necessary to consider an additional measure as PR. The latter expresses the proportion of tweets truly talking about environment (TP) among the tweets classified by the algorithm as related to the topic (FP+TP). Our algorithm of classification leads to excellent performances also from this point of view: 97.5% of tweets classified as linked to environment are actually speaking about this topic.
Both SE and PR are independent of the number of true negatives. Thus, the evaluation can be improved by taking into account SP in order to evaluate the percentage of tweet not linked to the environment that are correctly excluded by the algorithm: this percentage is equal to 97.5%. A low value of SP would mean that the algorithm has a high rate of false positive. Nevertheless, both considering tweets linked and not-linked to the environment, the algorithm we propose is able to find out a quite high percentage of correctly classified tweets. This means that the capability of our method is very well balanced for both the categories of tweets (linked and not linked with the topic, i.e. TP and TN), at least in our case study.
Given the usual trade-off that exists between SE and PR (i.e. when SE increases, PR decreases and vice versa), it is suggested to combine the two measures in an

overall index represented by the F1 score. This is defined as the harmonic mean of SE and PR. As a result, we obtain a value of F_1 score equal to 98.40%. This is very satisfactory and shows that the algorithm has a low rate of false positives and false negatives; this means that we are able to correctly identify relevant messages and, at the same time, do not include in our analysis non-pertinent tweets.
Table 4 summarizes all the performance measures computed for our method.

TABLE 4 here (now at the end of the file)

All the scores reported in Table 4 are very close to 100 and denote, generally (i.e. from any of the considered point of view), very good performances of the algorithm.
Concluding, evaluating all the indexes together, we notice that our dictionary-based algorithm performs very well. In particular, it outperforms the supervised feature-based models of Frenda et al. (2019) and of Foucalt et al. (2016) which report an AC of 76% and a F_1 score equal to 72%, respectively. At the same time our method performs more similarly (despite anyway outperforming such methods of more than 8 percentage points) to the computationally intensive deep learning methods implemented in Stowe et al. (2019) and Nizzoli et al. (2019), which obtain values of the F_1 score equal to 83% and 90%, respectively. In the final balance our method is able to outperform both the measures obtained by Stewe et al. (2019) and by Nizzoli et al. (2019). Additionally, the computational complexity and the load for the researcher, using our method, is noticeably reduced, in comparison to such intensive deep learning based approaches.

6 Conclusions

In this paper we propose an unsupervised dictionary-based algorithm for filtering tweets concerning a specific topic: environment. As discussed in Sect. 4.1 and depicted in Figure 1, the algorithm follows different potentially iterative steps. In particular, we consider two review phases (standard and extra) both applied to the expressions and to the hashtags as well. The extra cleaning step is useful to reduce the final standard review of expressions, but it is not mandatory and users can proceed with the standard cleaning only. This is a first advantage of our method, in terms of flexibility: the researcher can decide whether (or how frequently) to implement a deeper cleaning phase according to several factor (e.g., how frequently the base language for the topic under study changes, how much time passed after the previous analysis, how much time and resources are available for the processing, and so forth). However, generally we suggest to adopt the standard approach only the first time the user creates a dictionary and to prefer the double review (standard and extra) otherwise, in order to simplify the dictionary creation procedure. In particular, if a list of environment related expressions is already available (i.e. we can implement the list-based approach of Figure 2), our suggestion is to use this big set of general expressions as a base for the extra cleaning. On the contrary, if such a list is not available or is outdated (i.e. the user-based approach of Figure 2), it is preferable to limit the number of general terms that later should be manually checked in order to avoid including environment-related expressions. In any case these two different strategies (list-based vs user-

based approach) will not change significantly the final dictionary results, as proved by our results (see sect. 4.2).
Moreover, the flexibility of our methods is even more enhanced by the fact that the thresholds set to select the bigrams and trigrams keywords and the list of hashtags to be included in the dictionary can be increased or lowered, if needed. These settings can be defined according to the study background, circumstances, resources and can be changed or adapted taking into account, for example, if we are developing a first run rather than further waves of the same analyses. The approach we propose for setting the thresholds is even less prone to the arbitrary of a human decision than it could seem. By visualizing the entire list of bigrams, trigrams and hashtags, it should become clear that at some point the expressions start to be unrelated to the argument of interest. This makes the identification of the correct threshold a relatively easy and a not-so-arbitrary task. Some keywords, such as "coral reefs", which can appear into the final dictionary, can be relevant to the phenomenon object of study, but at the same time can lead to the inclusion of misleading tweets. We suggest to keep exclusively expressions strictly related to the environment topic, trying, generally, to be more restrictive than inclusive.
As shown in Sect. 5, the performance indexes obtained using our method highlight a very high quality, even if we compare them with the main literature findings. Both AC and F_1 scores are higher than 98%; but also evaluating the capability of identifying TP and TN the performances are excellent (97.5%). Our method seems to outperform the main algorithms recently proposed in the literature, while being at the same time extremely convenient from the computational point of view (i.e. the running time is extremely lower if compared with deep learning models).
This is not the only advantage of our algorithm. It is extremely flexible, since it can be applied in any type of field and studying any type of topic that includes a textual analysis of messages sent by using a social media.
Moreover, we propose a method that allows the researcher to include a pre-set dictionary, in order to create an own personalized one or allows to integrate both approaches. Being able to update or "increase the size" of a dictionary, when desired, is a valuable advantage, mostly if the topic studied is linked to a language quickly changing and/or to events that drive and characterize the citizen perception of the topic itself.
Finally, instead of proposing as starting point a list of single keywords or terms, we propose the use of bigrams and trigrams; this choice reduces the error of misclassification related to the use of single words. At the same time, the inclusion of hashtags and the "translation" of emoticons or web links allow to keep connected the dictionary creation with the most recent updates and events linked to the topic under study.

Our research is currently affected by some limits, that can suggest ideas for further research. The unsupervised method we propose is still not fully automatic, because it requires to set the thresholds for selecting the hashtags and bigram/trigrams and a final manual removal of keywords not strictly related to the topic of interest is needed. Nevertheless, this could also be seen as an advantage, because it makes the algorithm flexible and modifiable according to the users' requirements or preferences. Moreover, the list of OSA used as base for our algorithm affects the obtained dictionary and this phenomenon could be more relevant in other field or studying other topics. Thus, other less arbitrary criteria

about which and how many OSA to select as starting point can be proposed and tested.
Other ideas for further research include deeper and more general studies aimed, for example, to check the degree of generalization of our methods in other contexts (i.e. varying topics and/or countries/languages).
As future research, we intend to increase the number of keywords included in the dictionary by means of a periodic analysis of the OSA accounts. In fact, environment is a really sensitive, discussed and extremely trending topic, which can vary frequently. For these reasons, it is crucial to update the dictionary on a regular basis, in order to capture new trends, impact of events, and the consequent changes of people opinions. Our algorithm perfectly fits with these needs. And our algorithm can also be used to study how quickly and how much a dictionary regarding our topic could change over time.
In addition, our method allows to filter tweets by topic, thus it can be applied as starting point to develop a wide variety of analysis regarding other topic or can be used to go deeper into the study of our same topic. Further studies could be focused, for example, on sub-arguments of environment. For example, it would be interesting to filter tweets related to local problematics (i.e. air pollution) rather than to global issues (i.e. global warming) for more detailed longitudinal and spatial studies of the sentiment. This extremely detailed information could be used to study the sentiment on a small scale and, at the same time, to explore how much people care about big themes such as earth health. In this way, we are able to capture the population feelings, to link this to national and/or international policies and events and to identify the main drivers of the inclination and sentiment trends.
Finally, the flexibility of our method can be finalized to create several dictionaries for all the sub-topics connected to a more general phenomenon, such as the well-being (that includes, by nature, different dominion, e.g.: social involvement, health, work status, discrimination; see Toninelli and Cameletti 2018). In this case, selected tweets can be used to study the single dominions and to estimate the subjective well-being and/or how much a single dominion is able to affect the subjective well-being of a population. This will represent an improvement with respect to standard questionnaire-based surveys, such as the European Social Survey[18]. Better, the two types of sources can be integrated. In fact, thanks to the real-time collection of tweets, it will be possible to obtain timely information about a multidimensional phenomenon such as the well-being with a very high temporal and spatial resolution. These results can be of high value for evaluating the interventions of policy makers, for measuring the effectiveness of advertising campaigns, for studying a lot of other socio-demographic phenomena.

References

- Ahmed W, Bath P A, Sbaffi L, Demartini G (2019) Novel insights into views towards H1N1 during the 2009 Pandemic: a thematic analysis of Twitter data. Health Information and Libraries Journal, 36(1): 60–72
- Baylis P, Obradovich N, Kryvasheyeu Y, Chen H, Coviello L, Moro E, Cebrian M, Fowler JH (2018) Weather impacts expressed sentiment. PloS one, 13(4)
- Bing L (2015). Sentiment analysis and opinion mining. New York: Cambridge University Press.

[18] For further information, see: https://www.europeansocialsurvey.org/.

- Budiharto W, Meiliana M (2018) Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. Journal of Big Data 5(1): 1–10
- Cody E M, Reagan A J, Mitchell L, Dodds P S, Danforth C M (2015) Climate change sentiment on Twitter: An unsolicited public opinion poll. PLoS ONE, 10(8): 1–18
- Foucault N, Courtin A (2016) Automatic Classification of Tweets for Analyzing Communication Behavior of Museums, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 3006-3013, https://www.aclweb.org/anthology/L16-1480
- Frenda S, Ghanem B, Montes-y-Gómez M, Rosso P, Pinto D, Singh V (2019) Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. Journal of Intelligent & Fuzzy Systems, 36(5): 4743–4752
- Grimmer J, Stewart B (2013) Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, 21(3): 267–297
- Mitchell L, Frank M R, Harris K D, Dodds P S, Danforth C M (2013) The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. PLoS ONE, 8(5)
- Moran P A (1950) Notes on continuous stochastic phenomena. Biometrika, 37: 17–23
- Mouginot J, Rignot E, Bjørk A A, van den Broeke M, Millan R, Morlighem M, Noel B, Scheuchl B, Wood M (2019) Forty-six years of Greenland Ice Sheet mass balance from 1972 to 2018. Proceedings of the National Academy of Sciences of the United States of America, 116(19): 9239–9244
- Nielsen F Å (2011) A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages, Keraklion, Crete, Greece, 93-98
- Nizzoli L, Avvenuti M, Cresci S, Tesconi M (2019) Extremist Propaganda Tweet Classification with Deep Learning in Realistic Scenarios. Proceedings of the 11th International ACM Conference on Web Science (WebSci'19), Boston, USA, 203–204, https://doi.org/10.1145/3292522.3326050
- Pruss D, Fujinuma Y, Daughton A R, Paul M J, Arnot B, Szafir D A, Boyd-Graber J (2019) Zika discourse in the Americas: A multilingual topic analysis of Twitter. PLoS ONE, 14(5): 1–23
- Reyes-Menendez A, Saura J R, Alvarez-Alonso C (2018) Understanding #worldenvironmentday user opinions in twitter: A topic-based sentiment analysis approach. International Journal of Environmental Research and Public Health, 15: 2537
- Schlosser S, Toninelli D, Fabris S (2019). Looking for Efficient Methods to Collect and Geolocalise Tweets, in Smart Statistics for Smart Applications – Book of Short Papers SIS2019. Milan (IT), 18-21 June 2019, 2019, 1057–1062
- (https://it.pearson.com/docenti/universita/partnership/sis.html)
- Spratt D, Dunlop I (2019) Existential climate-related security risk: A scenario approach. Policy Paper from Breakthrough-National Centre for Climate Restoration, Breakthrough - National Centre for Climate Restoration, Melbourne, Australia https://www.preventionweb.net/go/65812
- Stowe K, Anderson J, Palmer M, Palen L, Anderson K (2018) Improving Classification of Twitter Behavior During Hurricane Events. Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, 67–75 (https://www.aclweb.org/anthology/W18-3512.pdf)
- Toninelli D, Cameletti M (2018) Is Structural Equation Modelling Able to Predict Well-being?, in A. Abbruzzo, E. Brentari, M. Chiodi and D. Piacentino (Eds), Book of short Papers SIS 2018, Proceedings of the 49th Scientific Meeting of the Italian Statistical Society, Palermo (IT), 20–22 June 2018. Pearson, ISBN: 9788891910233 (https://bit.ly/382HPRA)

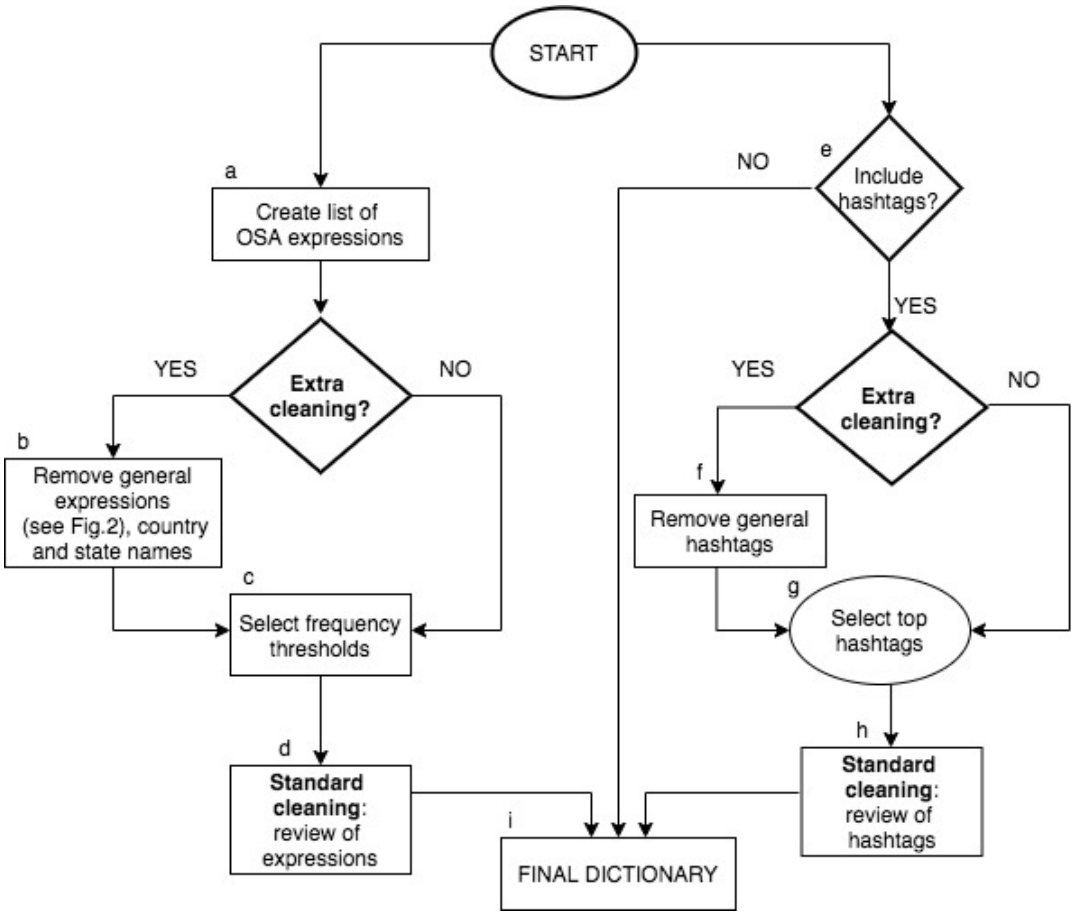

Figure 1 Algorithm for creating the final dictionary

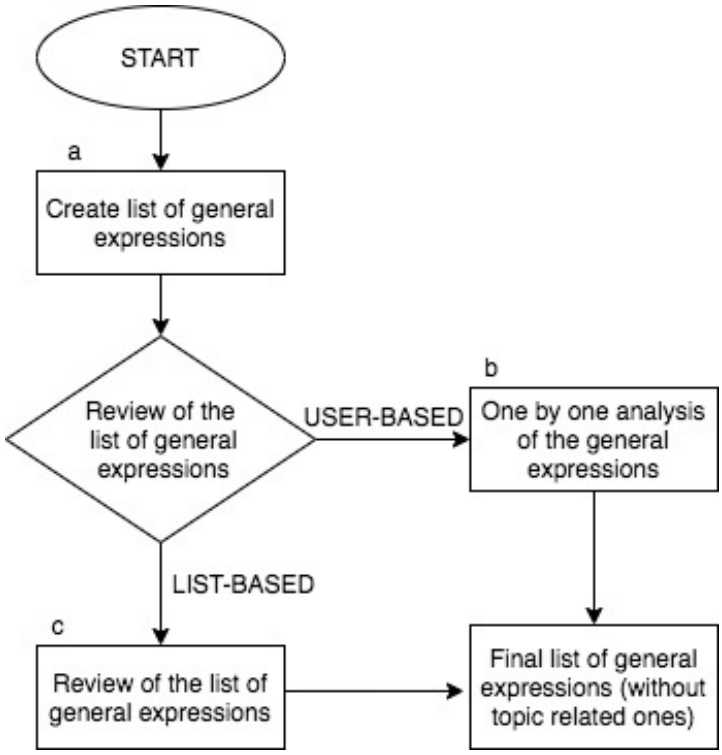

Figure 2 Extra cleaning phase (step *b* of Figure 1)

Table 1 OSA bigrams and trigrams and corresponding frequencies.

Bigram / Trigram	Freq.	Bigram / Trigram	Freq.
climate change	1357	leadership corps	81
donating tweet	547	million people	81
start donating	547	reality leadership	80
tweet unsubscribe	547	climate conference	79
plastic pollution	460	singleuse plastics	79
climate crisis	405	climatechange conference	77
climate action	302	extreme weather	77
air pollution	283	air quality	74
palm oil	253	offshore drilling	70
climate reality	246	coral reefs	68
singleuse plastic	245	paris agreement	68
renewable energy	213	tar sands	68
plastic waste	202	food waste	66
fossil fuel	183	marine life	66
clean energy	180	world leaders	66
fossil fuels	176	antarctic ocean	65
greenhouse gas	147	natural gas	65
plastic packaging	128	deposit return	63
uselessplastic lessoceanplastic	125	ocean plastic	63
plastic straws	124	uk government	63

connect earth	119	david attenborough	62
trump administration	118	barrier reef	61
gas emissions	117	sea ice	61
global warming	107	raise awareness	60
carbon emissions	99	[more excluded bigrams]	
conference cop	97	antarctic ocean sanctuary	43
global climate	94	support balloon releases	43
taking action	90	rising global temperatures	29
plastic bags	89	arctic sea ice	28
scott pruitt	89	conference sb bonn	26
send thinkprogress	86	drasticonplastic timer challenge	24
human health	84	action summit gcas	22
national park	84	exposing white nationalism	22
clean air	82	fashioned po box	22
plastic bottles	82	mobile calendar wallpaper	22

***Legend**. Gray: excluded (frequencies lower than 65/35 for bigrams/trigrams). Red: removed by the extra cleaning (step a, Figure 1); blue: removed by the manual cleaning (step d, Figure 1).*

Table 2 OSA hashtags and corresponding frequencies.

Hashtag	Freq.	Hashtag	Freq.
#climatechange	1530	#youngchamps	86
#climateaction	861	#renewables	85
#cop	759	#earthday	83
#plasticpollutes	629	#refusesingleuse	82
#endangeredemoji	566	#renewableenergy	80
#parisagreement	461	#reuse	79
#plasticpollution	389	#oceanplastic	78
#earthhour	375	#passonplastic	77
#uselessplastic	337	#biodiversity	75
#climate	301	#beatairpollution	74
#beatplasticpollution	277	#climateambition	72
#earthhourscotland	244	#nature	70
#breakfreefromplastic	237	#bethechange	69
#plastic	207	#talanoa	65
#fracking	193	#nothirdrunway	63
#cleanseas	191	#fridaysforfuture	62
#beatpollution	178	#zerowaste	61
#globalgoals	174	#africaclimateweek	61
#pandahugs	161	#cleanenergy	60
#lessoceanplastic	154	#beachclean	60
#solvedifferent	150	#wildforlife	60
#plasticfree	149	#breathelife	59

#protectantarctic	136	#fightforyourworld	58
#lessplastic	132	#singleuseplastic	57
#connect	129	#climatebreakdown	56
#airpollution	126	#climatechangebill	56
#climateemergency	125	#renewable	56
#worldenvironmentday	125	#solar	55
#dyk	120	#climatestrike	55
#endoceanplastics	111	#atlanta	53
#gcas	97	#oneplanet	52
#actonclimate	95	#blueplanet	51
#promisefortheplanet	94	#climatehope	50
#sb	89	#worldwildlifeday	50
#dropdirtypalmoil	88	#bees	50
#drasticonplastic	87	#reusable	50

***Legend**. Gray: excluded (not belonging to the top 60). Red: removed by the extra common cleaning (step f, Figure 1); blue: removed by the standard review (step h, Figure 1).*

Table 3 Confusion matrix.

		True category	
		Not related	Related
Predicted category	Not related	TN	FN
	Related	FP	TP

Table 4 Performance indexes values (in percentages) for the environment dictionary-based algorithm.

AC	SE	SP	PR	F1
98.42	99.32	97.55	97.50	98.40

Figures

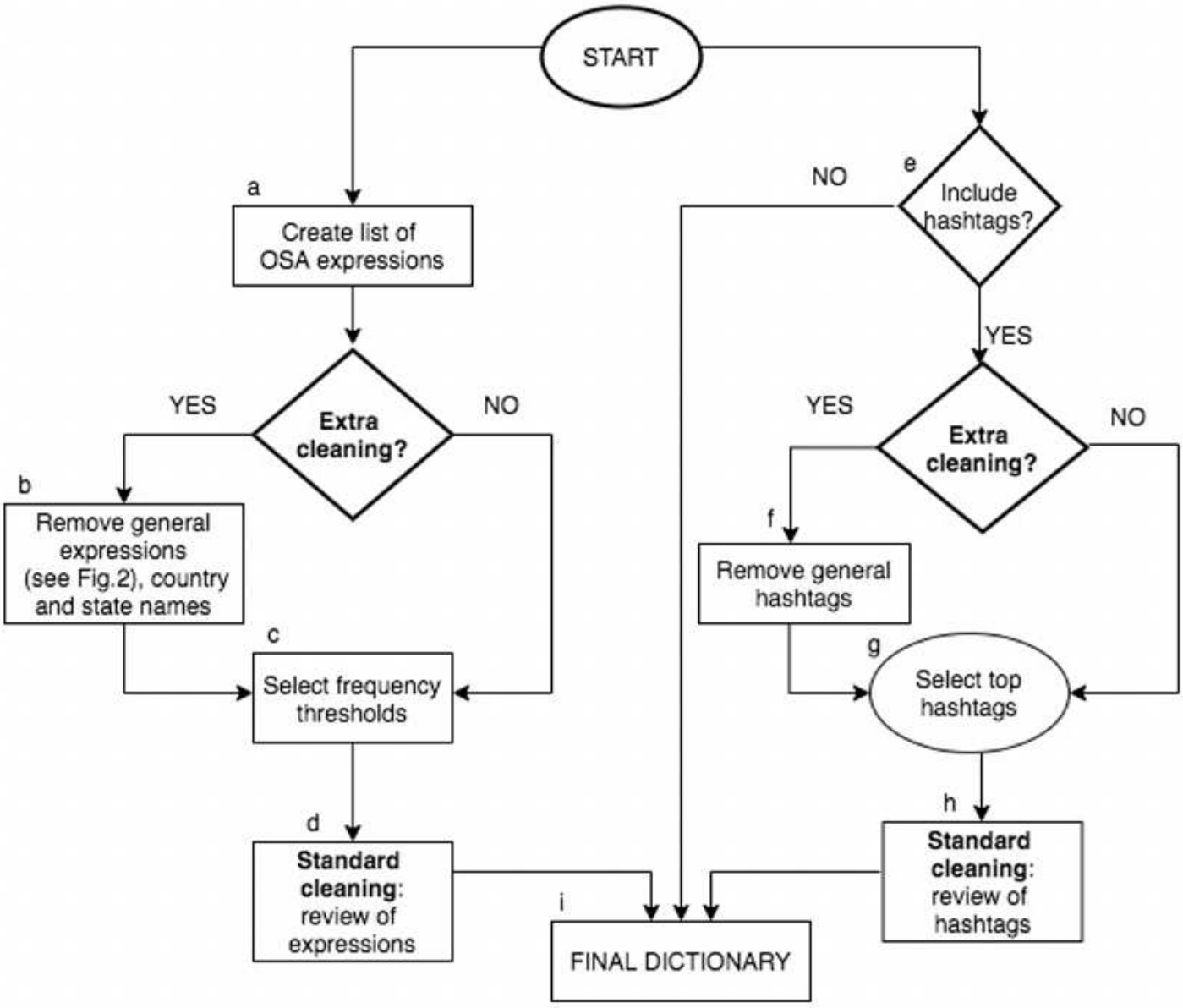

Figure 1

Algorithm for creating the final dictionary

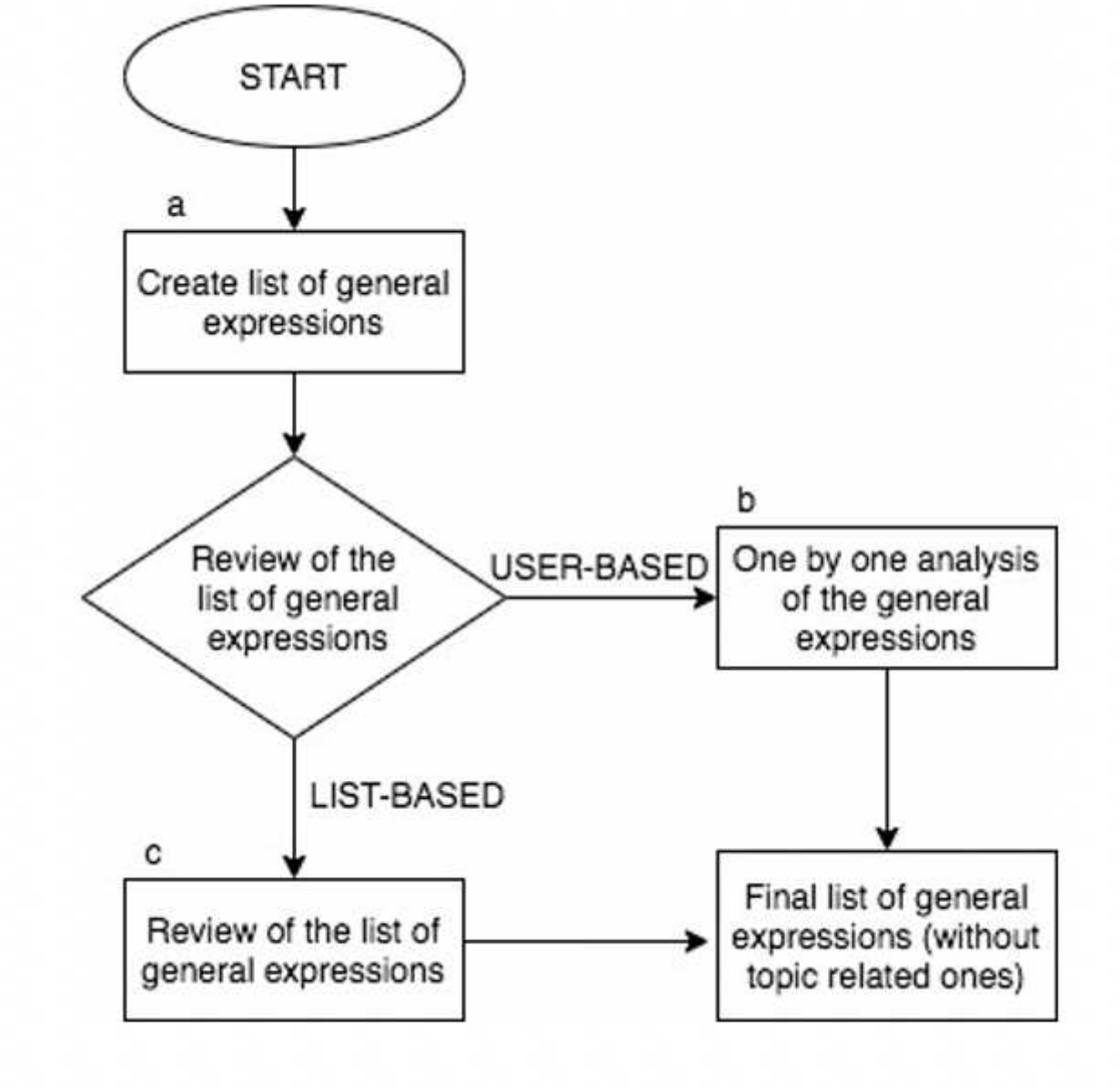

Figure 2

Extra cleaning phase (step b of Figure 1)