Abstract
Social media is more and more dominant in everyday life for people around the world. YouTube content is a resource that may be useful, in social computational science, for understanding key questions about society. Using this resource, we performed web scraping to create a dataset of 644,575 video transcriptions concerning net activism and whistleblowing. We automatically performed linguistic feature extraction to capture a representation of each video using its title, description and transcription (downloaded metadata). The next step was to clean the dataset using automatic clustering with linguistic representation to identify unmatched videos and noisy keywords. Using these keywords to exclude videos, we finally obtained a dataset that was reduced by 95%, i.e., it contained 35,730 video transcriptions. Then, we again automatically clustered the videos using a lexical representation and split the dataset into subsets, leading to hundreds of clusters that we interpreted manually to identify a hierarchy of topics of interest concerning whistleblowing. We used the dataset to learn a lexical representation for a specific topic and to detect unknown whistleblowing videos for this topic; the accuracy of this detection is 57.4%. We also used the dataset to identify interesting context linguistic markers around the names of whistleblowers. From a given list of names, we automatically extracted all 5-g word sequences from the dataset and identified interesting markers in the left and right contexts for each name by manual interpretation. The results of our study are the following: a dataset (raw and cleaned collections) concerning whistleblowing, a hierarchy of topics about whistleblowing, the automatic prediction of whistleblowing and the semi-automatic semantic analysis of markers around whistleblower names. This text mining analysis can be exploited for digital sociology and e-democracy studies.

Keywords Social media · Text mining · Net activism · Computational social science · Natural language processing · Machine learning
1 Introduction

Almost a billion people around the world use social media, and this number will probably increase. Net activism is an important behavior that occurs on social media. Some recent revolutions resulted directly from net activism, like the Orange Revolution in Ukraine in 2014 or the Jasmine Revolution in Tunisia between December 2010 and January 2011 (the “Arab Spring”) [1]. In Ukraine, the Orange Revolution began before social media in 2004, but social media helped the Ukrainian diaspora to continue, through activism on Facebook, for instance. Disruption happens, but social media makes it happen much faster; social media caused, in 2014, the Ukrainian president to leave his country, as the Tunisian president did four years earlier. Net activism plays an important role, not only from a citizen’s point of view but also from a government perspective. For instance, governments must pay attention to the trolling phenomena that aim to destabilize public opinion, which occurred, for example, during the American presidential elections in 2016 on Twitter [25].

Definition A whistleblower is any person, group or institution that, having knowledge of a danger, risk or scandal, sends an alarm signal in the hope of triggering a process of regulation, controversy or collective mobilization. In terms of natural language processing, a “whistleblower” will be a named entity (the name of a person or organization), “knowledge of a danger or risk” will refer to a topic and perhaps a named entity (a product or organization) that fails and finally, an “alarm signal” will refer to some linguistic pattern [15].

Whistleblowers can use social media to spread information, or a third party can reveal a serious event promoting whistleblowers. YouTube is a large-scale platform that has a big impact at the present time. The younger population is more connected to such media than to traditional media sources such as newspapers and TV [28].

So, information is available, and the study of whistleblowing can be divided into two computational tasks: first, predicting where and when the whistleblowing occurred and who it originated from; and second, plotting a topic map that displays which categories of problems whistleblowing focuses on. Indeed, whistleblowing is not a new social phenomenon, and we can try to discover whether whistleblowing on digital social media covers a variety of subjects and which subjects it covers. Since the beginning of the twentieth century (at least), the history of economics has been full of examples of this kind of activism.

More than a century ago, a US law was enacted specifically to protect whistleblowers: the Lloyd-La Follette Act of 1912. This law guaranteed the right of federal employees to furnish information to the US Congress. The word “whistleblower” appeared at beginning of the 1970s. More recently, famous and recurrent examples of whistleblowers have populated the news, such as Wikileaks (an NGO founded by Julian Assange) and Edward Snowden (an ex-NSA employee who revealed the agency’s mass surveillance program). In 1991, French sociologists Luc Boltansky and Laurent Thévenot [11] studied the organization of the city and revealed several types of engagement roles; among them, individual engagement is an emerging model that goes beyond inter-class conflict. Following this framework of city organization, in 2000, two other French sociologists, Francis Chateauraynaud and Didier Torny [15], pointed out the importance of the role of modern prophets, which is played by whistleblowers.

The new face of social worlds within digital worlds strengthens and intensifies whistleblowing. It is well-known that the main factors that are relevant to this idea are the...
popularization of Web 2.0 platforms (starting in 2003–2004, including Myspace, Wikipedia, Facebook, blogging) and the increasing number of social networks (250 in 2016).

Many people share information on social media with some sort of motivation [47]. Our goal is to prove that detecting whistleblowing (a specific motivation) is possible using automatic processing for documents that are posted on YouTube. There are very few works about net activism and protest analysis using social media data analysis [48]. In [41], FrameProv software was used to record a whistleblower’s testimony, which included the whistleblower appearing in front of the camera. In [5], abnormalities were found using image processing in diverse web video categories, including “Comedy,” “People and Blogs,” “Entertainment,” “Education,” “Sports,” “Music,” “News and Politics,” “Non-profit and Activism,” “Films and Animation” and “Gaming.” In [39], the detection of activists on video sharing websites/portals using flagging (YouTube allows users to flag inappropriate content so that their staff can review it) was discussed. However, to our knowledge, we are the first to process a dataset of whistleblowers using a quantitative approach.

Our study relies on making a reference dataset by scraping YouTube using a large time range and also on inferring knowledge from its lexical content with quantitative approaches from text mining. We will proceed to learn terminology using lexical extraction, which will generally be used to extract repeated segments [13, 33, 51] and named entities [40]. As we want to discover topics that may appear in the content, we need to learn an ontology by using a low-resource technique, such as clustering: we use k-means [37] and topic modeling [10]. Our goal is to also use our text mining pipeline to offer evidence of explainability concerning our results.

The working assumptions are the following:

- Whistleblowing can be expressed in any language.
- We exploit only available crawled data to infer new knowledge.
- Natural language processing can help to extract interesting features.
- We work with a reference corpus.
- We do not work with a whistleblower, so the validation of inferred knowledge can be tested only on YouTube with unseen data.
- The lexical context around whistleblower citations can be informative and discriminative.

Next, we will present state-of-the-art studies about YouTube analytics; then, we will present the dataset that we used and its content, followed by the methodology used to extract features and topics. Finally, the last section will be a discussion about the results.

2 YouTube and analytics

YouTube has been used worldwide for many years, and it became popular after its acquisition by Google in 2010. It is so popular that it has made music stars, such as Justin Bieber, billionaires; the Gangnam Style music video has over 1 billion views. It also popularizes different kinds of bloggers that promote products to their followers. According to globalmediainsight.com (2020), the YouTube community is made up of 2.3 billion users; among these users, 245 million are active daily, making YouTube one of the three most active web (and also mobile) platforms. YouTube users consume over 1 billion hours of videos daily (roughly 5 billion videos are viewed per day). The number of unique visitors per day is 560
million (a visitor looks at 8.9 videos per day on average). The three countries with the most YouTube users are the US, India and Japan. Looking at the frequency and volume of uploads, over 500 hours of video uploads happen every minute on YouTube (720,000 hours of video uploads per day). Hence, the platform is a warehouse of videos, with more than 5 billion videos shared. YouTube enables a huge quantity of users to use its services. This has generated considerable interest from researchers performing communication and interaction studies in YouTube itself as a research object, including all possible emerging collective uses, what we find there and who uses this kind of platform [1, 25, 28, 47, 48].

However, YouTube also has the computing platform provision issues that we find in other domains, such as computing machinery. So, we find classical and transversal computing approaches that focus on YouTube data and access, such as cybersecurity [4], machine learning [38], recommendation making [14, 30], streaming properties [36], spam detection [50], ontology and thesaurus work [16], network analysis [31], duplicate detection [46] and natural language processing [52].

Much closer to our topic, some studies have addressed political engagement, although the previous studies are of interest for data processing.

Some specific studies have claimed that Web 2.0 platforms reinforce democracy or can help us to understand Muslim radicalism through the democratization of engagement, for example. In [43], the political engagement of the One Health community was explored using a social media platform. In [32], the authors investigated the role of users in groups, how groups evolve and how the structure of groups organized under a category changes over time. In [42], it was claimed that YouTube is not just subject to the same social forces that limit discursive egalitarianism in the real world; it is subject to economic forces as well. In [20], the effects of social media on citizens’ political participation and the influence of social media on reaching and influencing voters were studied. YouTube’s importance and popularity is such that Barack Obama is being called the first social media president of the United States due to his campaign’s YouTube strategy [22]. According to [21], such sites have given a voice and platform to young people long marginalized by strict regimes. In [27], the grateful comments of users who regularly expressed a sense of empowerment through simple online actions were studied. In [26], the authors compared political discussions and found that Facebook presents a more egalitarian distribution of comments between users and a higher level of politeness in these messages than YouTube. In [35], the authors suggested some new measures of interaction suitable for measuring audience engagement with online media. They showed that among trending videos, users are typically quite active in voting and commenting. However, due to the considerable variation of audience engagement, on average, there is one vote for every 200 views and one comment for every 1600 views.

Activism and extremist activism are important subjects in the literature. In [54], it was claimed that social media platforms have evolved to play a formative role in shaping global public opinion on a broad array of topics, including hacktivism, protests and revolution. The authors of [17] concluded that the vast majority of those who posted a martyr promoting video on YouTube and those who supported this content were in the 18-to-34-year-old age bracket. In [9], the authors stated that we cannot escape machine usage around us, which also implies a heightened need for citizen vigilance. The authors of [24] claimed that the idea of “radicalization via video” offers little insight but merely acts as an easy crutch for those who do not know how to deal with a heinous crime. In [58], the authors showed how online communication enabled diasporic Uyghurs to represent their political and cultural identities, while [29] illustrated the utility of network analysis as a diagnostic tool when dealing with proselytizing
for terrorism on social media platforms; they also showed specific case studies where social network analysis metrics also proved to be efficient for differentiating between al-Muhajiroun-related channels and seemingly similar jihadist propaganda channels. In [44], a codebook with predefined categories was used to study positive and negative videos about Islam on YouTube, showing unsurprisingly that positive videos are exclusively produced by Muslim users. In [2, 3], different machine learning strategies were proposed that were aimed at detecting radicalization efforts, cyber-recruitment, hate promotion and extremist support in a variety of online platforms, including YouTube, Twitter and Tumblr. According to [6], the broadcasting of insurgent atrocities over the Internet on platforms such as YouTube increases the spectacle effect of extreme violence by ensuring that it reaches wider audiences. The authors of [59] compared radical discourse between North Caucasus and Uyghur neo-jihadist narratives on the creation of collective identities and found that the narratives have similar features, which can be identified in a set of sub-narratives.

Environmental activism is a domain-centered activism. In [7, 8], the authors said that activists can be seen as soldiers in a war or perhaps more incisively as freedom fighters waging war on an occupational force. The construction of a series of discursive fields centers occurs around nodal points of war, resistance and injustice. Studying the activist network “Never trust a Cop,” they found that activists used this media to reach and mobilize new audiences that they could not reach through the use of their own alternative activist networks alone. The authors of [55] showed that video sharing also reveals disagreements in online discussions of environmental issues.

It seems difficult to mention political activism without mentioning recent protests and social movements in Egypt [56], Russia [57] and Mexico [53], where users and influencers are often highly educated and feel that they have gone down in the world.

Concerning human rights activism, [12] addressed questions about the danger of activists who incorrectly assume that their social media information is secure.

Regarding anti-war activism, [19] explored rhetorical concerns revealed by the intersections of disability, race, gender and globalization related to the Iraqi situation.

Concerning victimization, [49] analyzed the most active Spanish feminist online communities’ YouTube channels, which were dedicated to fighting gender-based violence; the authors used multimodal analysis and considered the concepts of agency and interpellation as defining practices in the configuration of the victim-subject. According to them, different modes of relationships and different processes of constructing the victim-subject of gender-based violence respond to different logics and models of understanding agency and activism in virtual spaces.

Regarding alter-globalization activism, [48] studied accounts of organizations protesting the G20 event. These accounts were dominated by the violence that accompanied the protests. They showed how the particular dynamic between the overwhelming police presence and violence, and the protestors’ efforts to document the actions of the police, produced accounts that were squarely focused on the violence and spectacle that accompanied and eventually overtook the protests. A topic close to whistleblowing, controversy can be defined, in a simplified way, as occurring when two groups of people disagree about a topic, leading to an interactive and conflictual situation. The authors of [45] introduced a probabilistic framework to model controversy based on a language modeling approach, in which controversy is defined through Wikipedia articles that are similar to a given document. They defined controversy and controversial topics consisting of various features, particularly a mixture model of topic and sentiment. The authors of [23]...
quantified a controversy using a conversation graph about a topic from Twitter, which represents the alignment of opinion among users; then, they partitioned the conversation graph to identify potential sides of the controversy, and finally they measured the amount of controversy from the characteristics of the graph using a random walk, edge betweenness and sentiment polarity.

These studies promote the activism of individual leaders that criticize society and try to sound the alarm, but very few studies use a large dataset to infer knowledge using machine learning, natural language processing or data analysis. Some studies perform an automatic detection of radicalism in social media or conversations. We find no studies that have tried to detect or describe whistleblowing well. Radicalism promotes some specific ideology, unlike whistleblowing, which can sound an alarm about any topic.

3 Datasets

The choice of YouTube for whistleblowing research comes from the fact (seen in section 2) that many activists use this platform to enlarge their audience; additionally, among the list of (about 250) common social media platforms (available from Wikipedia, for instance), we only find one platform for activism, and it is only related to online petitions (i.e., CARE2). Hence, YouTube is a relevant source to use to make a reference dataset.

On the YouTube platform, subtitles correspond to the transcription of the words spoken in a video. Transcription is the process, called speech recognition, of transforming speech into text. The implementation of speech recognition consists not of a single algorithm but of a set of algorithms and practices. Deep neural networks (DNNs), hidden Markov models (HMMs) and other models are used for decoding. For example, algorithms like the perceptual linear prediction (PLP) technique, the Viterbi algorithm for HMM decoding and many others are combined to produce the final result. The following reference from Google Research [34] describes the workflow for the Google YouTube caption system. In our study, we use mainly text from videos, which is available for any video, and focus on the content of the video: we use text from the speech transcription of the video, the title and the description that the user posting the video may provide.

Finding an API to get a dataset from a social network is a key issue, and it can become a pitfall for scientific studies since the available API is often not stable and can be modified (i.e. restricted) by the provider, because the owner of the social network is a big company.

Before 2018, YouTube offered a search API that could be used to retrieve any document from any period (see Fig. 1); we used this ability to create our dataset of videos posted between 2007 and 2018. Several fields are retrievable, such as the transcription of a video (in English), comments, the description and the title.

To use the API, a Google account (like “gmail,” for instance) is required. Then, one goes to Google Cloud to open a new project (https://console.cloud.google.com/). After creating a new project, a project name, a project ID and a project number are created. A client ID, a client secret and a refresh token, which are required to use API functions for video downloading, can then be generated. The first step is to download the video ID. The second step is to download the metadata for each ID. The ability to search the entire YouTube archive from the beginning of YouTube until the present by keyword has not been available since 2019, but timeline analysis is still possible (like the Twitter API).
Following the principle of Occam’s razor (i.e., the simplest solution is often the best) [18], we use the obvious keyword “whistleblower” to query YouTube. From 2007 to 2018, searching for published videos containing the word “whistleblower” leads to a corpus of about 644,575 videos. This corpus is impressive, but of course it includes noise as well as many relevant videos. The data are exploitable and explorable (see Fig. 2).

In our study, we performed automatic clustering on the dataset twice. The first clustering procedure is done on the raw dataset for noise identification. The second clustering procedure is done on the cleaned dataset for topic identification. After the visualization, we apply the first stage of clustering to the data to identify keywords describing noisy content. This clustering is done automatically using topic modeling and k-means. Keywords are identified manually by looking them up in the video clusters. It takes a long time to get the final list of keywords that we use to delete video entries from the raw collection of videos. We see non-relevant information related to cinema, video games, conspiracy and esotericism. With some frequent keywords, we are able to strongly clean the dataset with a high rate of reduction (semi-automatic curation). This is the anti-keywords list: outlast, Scientology, aliens, alien, Star Wars, UFOs, UFO, nibiru, planet, Lucifer, nephilim, Illuminati, Satanic, Nazi, Hitler, extraterrestrial, Life is Strange and The Walking Dead.

Our strategy to clean the dataset involves a triple-level processing procedure: 1—text conversion; 2—keyword selection; 3—keyword suppression. The first step is to convert each video and be sure that we have a text body for further content analysis. Here, there is no technical problem: text is generally available. In this step, we merge the title and transcription. The second step performs filtering with the keyword used to get the videos. As YouTube (like...
many other platforms) uses their own algorithm to retrieve information, some videos do not contain the important keyword. From this second step, we obtain 167,622 documents to use as an effective dataset for content analysis. The third step is tricky, because we apply feature extraction to create clusters (i.e., the stage explained in the next section) and observe whether the clusters are noisy or not. Then, we manually extract a list of noisy keywords that we use to make the final extraction of features. From this step, we obtain a final dataset of 35,730 documents. Figure 3 shows the distribution of the languages used in the dataset. Unsurprisingly, English is the language that is used the most (64% of the collection).

Comments are sometimes useful in making an analysis of importance. In our case, when videos are made by specific users (people or organizations) to sound the alarm, the whole

![Figure 3](https://example.com/image3.png)

**Fig. 3** Distribution of the languages used in video transcriptions from our final dataset (ar: Arabic, cn: Chinese, de: German, en: English, fr: French, ru: Russian)
audience cannot feel responsible for and sensitive about the topic. As we discuss in the results section (section 5, Table 3), many videos have zero comments. It is difficult to determine in some cases whether there are zero comments because comments have been deleted by YouTube tools or just because nobody wanted to write comments. YouTube employs many sophisticated spam control methods to try to remove spam comments (“controversial” comments, or comments that use a swear word); when a YouTube routine has deleted comments for a specific video, no information about those comments is available. In the rest of our analysis, we take into account only three pieces of text information from a video: the title, description and transcription. We are sure that this information will be relevant to the whistleblowing content of the original video, and there is no mixing of languages, whereas comments can be written in any language.

It is also worthwhile to discuss the transcription availability. Many videos do not contain transcriptions for different reasons. First, videos do not contain transcriptions when the language being spoken is not English; the transcription services provided by Google, when we built the dataset in 2016, were only available for videos in English. Second, there are no transcriptions for videos that are free of speech, which may contain video game content, music or just stories without speech. Once a video is posted on the YouTube platform, transcription is not directly available and may take several hours. We also observed that in some cases entire videos and their metadata are completely deleted.

Another point to mention is that some videos are literally deleted from the platform at some later time. We can see this happening because the corpus was made in 2018, and it constitutes a record of what had been published at that time and what is disappearing now. Globally, 22,078 out of 35,730 videos (62%) have no comments, 38% have at least one comment and 5400 videos (15%) have more than five comments. The mean number of comments per video is 11.3.

4 Methodology

4.1 Technical environment

We use a double cloud architecture for job and data parallelization. For data parallelization, we use SLURM. We can run 20 batches at the same time for the following tasks:

- document scraping by year from YouTube (what we call raw collection);
- the generation of sentences from the title, description and transcription metadata (what we call text collection);
- linguistic feature extraction (what we call linguistic collection).

For job parallelization, we use Hadoop with the MapReduce approach for the following task:

- cluster generation using the k-means approach.

The architecture is hosted by the University of Toulouse (IRIT Laboratory) and sponsored by the CNRS (French Centre for Scientific Research). The available memory from the cluster is only 16 GB for one job.
4.2 Feature extraction

A document in itself is not tractable. We need to split the string chain into single units. For us, there are different types of units. For a linguist, a single unit is commonly a word that has a relationship with other words in the same sentence. For us, a unit can be a noun phrase (word) collocation, a relational collocation or an n-gram (with n from 1 to 5 and 1-g being single words). We decompose all the fields from the metadata of a video into single units to perform further analysis. In further analysis, we do not differentiate between units belonging to different fields. Documents with their different fields and different units represent the dataset.

A basic dashboard of the textual content of the corpus consists of the distribution over languages and the simple element distributions. The linguistic representation of features for further analysis is important; they constitute simple elements that can be used to create classes.

By simple elements, we mean words, n-grams, named entities, relational phrases and noun phrases. Simple words were detected using regular expressions and simple splitting was performed with special characters. N-grams were detected using the R package ngram. Relational phrases were detected using the tool MaltParser. Named entities were detected using the Stanford models for named entity recognition. Noun phrases were detected using regular expressions over the syntactic tagging of sentences.

Two kinds of distributions highlight the content: occurrences and unique occurrences. Another good indicator is the average number of simple elements per document (see Table 1 for these distributions for English documents).

4.3 Topic analysis

Our toolbox split the corpus into subsets of about 10,000 videos chosen randomly (documents) from our reference corpus. We repeat this process 10 times, so we obtain 10 subsets to which we apply clustering. This Monte Carlo approach is preferable for achieving an optimal clustering solution that simply splits the dataset rather than splitting the dataset into a partition of three subsets (see section 3). This is necessary because of the computing memory limits of the clustering methods, even using sparse matrix libraries. We use clustering techniques to implement topic detection. We make different samples with a given size in terms of the number of documents, and for each sample, we extract terms for each video ID; hence, we produce the input dataset.

| Feature Type      | Number of Occurrences | Number of Unique Occurrences | Average per Document |
|-------------------|-----------------------|-----------------------------|----------------------|
| 1-g               | 140,084,586           | 768,600                     | 600.73               |
| 2-g               | 130,625,472           | 12,480,788                  | 560.16               |
| 3-g               | 121,835,587           | 45,160,987                  | 522.47               |
| 4-g               | 113,610,531           | 74,399,459                  | 487.20               |
| 5-g               | 105,942,797           | 85,958,733                  | 454.32               |
| Named Entities    | 2,656,663             | 347,951                     | 11.39                |
| Relational Phrases| 19,242,416            | 9,293,178                   | 82.51                |
| Noun Phrases      | 11,417,339            | 4,976,661                   | 48.96                |
To make a document-term matrix, we use the \texttt{tm} package from R. We filter the matrix by removing sparse terms; the terms that occur very infrequently or not at all in documents (i.e., terms occurring 0 times in a document) are removed from the matrix. We set 0.97 as the maximal allowed sparsity. We select features with a frequency greater than 10.

We compute the \texttt{lda} function using the \texttt{topicmodels} package from R, with \texttt{alpha} set to 0.1 and the number of topics set to 100 (clusters). We then export the terms of the clusters with a posterior topic probability greater than 0.99.

We use the \texttt{k-means} function from the \texttt{stats} package from R. We change only one parameter: the number of desired clusters is set to 100. From clustering, we automatically obtain video clusters according to their word distribution. We divide the dataset into different batches of 10,000 documents and we repeat the clustering for each batch.

Then, we perform direct visualization (scanning by eye) and expect to find a high number of distinct topics (see Fig. 4).

### 4.4 Unknown video filtering

Given a category, we used a ranking list approach based on the intersection of n-grams between a set of videos covering a subtopic. Common n-grams are considered patterns and interpreted as a local grammar for whistleblowing related to a specific topic. These patterns are supposed to achieve a good filtering for the discovery of new incoming documents about whistleblowing and the given category. For filtering, we use a joint Jaccard and cosinus value for each unknown video to make a prediction. This value is specific to a particular cluster.

As seen in the filtering procedure below (algorithms 1, 2 and 3), to perform filtering we need the labels of topics and clusters (determined in subsection 4.3), and we need the dataset and feature extraction algorithm seen in section 3 and subsection 4.2, respectively. Algorithm 3

Fig. 4  Discovery of topics made by the clustering of terms based on different samples of our dataset (section 3). From the clustering technique, we automatically obtain more than 100 clusters that we interpret to make a list of categories and subcategories.
returns a Boolean vector that indicates whether an unknown video vector belongs or does not belong to a given topic.

Algorithm 1 Learning step.

**Input:** topic keywords, set of clusters, corpus database.

1: Choose a topic $t_k$ from the set of $n$ topics $\Theta = \{t_k\}$.
2: Choose several clusters $C_k$ (related to $t_k$) from the set of clusters $C$ ($C_k$ is included in $C$) and make a group $G_c$.
3: Get all features from the common dataset (1-grams, 2-grams, 3-grams, 4-grams, 5-grams, noun phrases, named entities).
4: Make a word space $W_s$ from the features.
5: Delete the noise from $W_s$.
6: Make a Boolean vector for each video in $G_c$.
7: Extract the grammar of the topic using the common frequent features $GR_c$.
8: Make a mean vector $V_c$ of $G_c$.

**Output:** $V_c$.

Algorithm 2 Reference step.

**Input:** topic keywords, set of clusters, corpus database, unknown video set.

1: Choose a topic $t_k$ from the set of $n$ topics $\Theta = \{t_k\}$.
2: Get a group of unknown videos related to $t_k$ (topic group), $G_t$ (scrape YouTube by keyword).
3: Get a group of unknown videos related to $t_k$ and whistleblowing (Whistleblowing topic group), $G_w$ (scrape YouTube by keyword).
4: Make Boolean vectors for each video in $G_t$ and $G_w$.

**Output:** $G_t$ and $G_w$.

Algorithm 3 Metrics step.

**Input:** $V_c$, $G_t$ and $G_w$ computed in learning and reference steps;

$T \in (0, 1)$ a threshold.

/* Let $Jac$ be the Jaccard distance and $Cos$ be the cosine distance */

1: for each video Boolean vector $V$ in $G_t$ and $G_w$, do

/* compute the metrics with $V_c$. */
2: $Jac(V, V_c) = \frac{V \cap V_c}{V \cup V_c}$, then normalize by median of all Jaccard measures.
3: $Cos(V, V_c) = \frac{\sum_i V_i V_c}{\sqrt{\sum_i V_i^2} \cdot \sqrt{\sum_i V_c^2}}$, then normalize by median of all cosine measures.
4: $D(V, V_c) = (0.9*Jac(V, V_c) + 0.1*Cos(V, V_c)) > T$.
/* $D$ is Boolean; true means that the unknown video is closed to the cluster vector */
5: end for

**Output:** $D$ values.

We perform a simple evaluation by choosing small datasets with the keyword “whistleblowing” and the keyword of a topic (for instance, “vaccine”) for a period unseen by the crawler. We expect a dataset of about 100 unknown documents on the topic and several documents about both whistleblowing and the topic. From these, we sort out the true positives.
and the true negatives. We apply the Jaccard/cosinus method to output relevant documents that we can compare to the true positive subset for the computation of the accuracy.

### 4.5 Context and attention analysis

The corpus contains several thousand videos about whistleblowing, but we need to find a measure of relevance. First, an indicator to confirm that our corpus is representative of whistleblowing relies on the observation of famous whistleblower names inside the corpus. Let us suppose that our corpus does not contain any famous whistleblower names: this would mean that the corpus (reference text dataset) is badly built. Second, it is possible to extract all left and right context around the citation of the whistleblower names in the corpus. In this way, it is possible to discover which tokens are used and may be typical of the story of the whistleblower or related to whistleblowing in general. This is an analysis of context informative words by semantic interpretation. We perform semantic interpretation in two steps: first, we automatically extract short pieces of context containing a specific name by using regular expressions; second, we manually identify representative lexical markers by looking at the right and left context around the whistleblower name. For instance, let us call \( \{SC_k\} \) the set of short pieces of context, \( \{W_i\} \) the set of words and \( N \) a whistleblower name. Then, a specific \( SC_1 \) consists of a sequence of five words (i.e., a 5-g): \( SC_1 = "W_1 W_2 W_3 W_4 W_5" \). From \( \{SC_k\} \) we select all possible SCs where \( W_1 = N \) to get the right context \( [W_2 W_3 W_4 W_5] \) or where \( W_5 = N \) to get the left context \( [W_1 W_2 W_3 W_4] \). Then, we are sure to get all contexts where \( N \) is cited. For this kind of extraction, we need to be sure that \( N \) is truly the name of a whistleblower. Name entity detection will not be a good solution since many names are cited in documents, and there would be uncertainty about whether a context is related to the whistleblower or not. If we extract the context with a known list of names, we avoid uncertainty about the context. The analysis of the context cannot be done automatically since we do not know what kind of markers we can identify from each context. This is a part of the exploratory analysis of the corpus. Each whistleblower can have their own markers depending on their story and sociological aspects. Here, we proceed to a discovery of possible markers and potential rules about what can be considered a marker around a whistleblower name. In a further step, the typology of the markers we discover should be exploited as rules, but this is beyond the scope of this article.

We get a short list of 30 famous whistleblower names from Wikipedia to see if they occur in the dataset. The names are the following: Snowden, Manning, Maguire, Eckard, Woodford, Newcastle, Percival, Stern, Pandhare, Wilson, Segarre, Deltour, Weber, Crane, Strickland, Tye, Ting, McGill, Doe, Bitterman, Pars, Stevenson, Bond, Nzobonimpa, Siska, Meikar, Do Rego, Vijayakumar, Matobato and Osmund.

We check to see if these names are present in the list of 5-g. From the feature extraction of the 5-g, we got 21,981,438 distinct sequences. If a name occurs in a video, it is present in at least one of the 5-g. Second, we compare the most frequent words at the positions +1, +2, +3, +4 and − 1, −2, −3, −4 from a name to see if the context plays an important role in extracting important words. We list the top 20 words by decreasing order of occurrence.

### 5 Results

In our results, various topics concerning society and institutions are discussed, including technology, life, the economy and the environment, as we can see in Table 2. In this table,
topics and subtopics have been identified by manual interpretation and through studying more than 100 video clusters according to their lexical content. Interpretation is not easy since this content may concern any topic in real life. Performing automatic labeling would require a very large ontology containing all words that could possibly appear in any video transcription and title. Even if such an ontology exists, there is no consensus about common-sense large-scale concept-word ontologies. In our corpus analysis, the choice of a label has been done using either the close context of a video cluster or a word that may appear significantly in titles.

A surprising observation is that whistleblowers do not discuss dangers to the environment only; they are also concerned with various other topics. In general, when clustering returns a group of videos, the videos can concern a subtopic, but often the group includes videos from other topics or noise. We chose as a parameter a maximal size of clusters of between 20 and 50 documents. Below (Table 3) is a list of videos from a cluster concerning the subtopic “Health and Food.” The number of views for each video is not high, but the aggregated views of the whole

| Video ID | Title                                                                 | Number of views | Number of comments | Time posted | Active? |
|----------|------------------------------------------------------------------------|-----------------|--------------------|-------------|---------|
| CAUZKMmAuyc | WWYW #31 Segment 2: Food Safety Whistleblower Kenneth Kendrick         | 1582            | 4                  | 2015–10     | yes     |
| DVx8gRBhTWY | Monsanto FDA and Corruption - TRUTH NOIR ep. 12                       | 730             | 0                  | 2015–09     | yes     |
| qC7dSb-bcBo | Fermented Cod Liver Oil Scandal: Dr. Rudi Moereck                     | 4125            | 6                  | 2015–10     | yes     |
| 4vbX2bTCBo0w | Bovine somatotropin                                                    | 858             | 0                  | 2015–10     | no      |
| fNnmPqLnaro | Eating Disorders & Sexual Abuse | Chatter Chats | Life with Lydia    | 444         | 0       |
| IvCe0XA0laI | The War on Food and Drink Freedom and One Group Thats Fighting Back   | 1665            | 32                 | 2016–03     | no      |
| 9iJQvrEOjiU | How Safe is Atrazine?                                                  | 17,298          | 19                 | 2009–08     | no      |
| owugdp7QTnE | ANP INVESTIGATION: EPAs Troubled Waters                               | 567             | 1                  | 2009–01     | yes     |
| c9M99_BQXkg | Possible Cover-up in China Milk Scandal                                | 16,755          | 46                 | 2008–09     | no      |
| UBMBmIANdk0 | BP Oil Spill: Bible Says Fresh Water Is Being Poisoned                 | 406             | 22                 | 2010–07     | yes     |
| WKTJ47R-hNw | Radioactive Japanese green tea - already in a store near you          | 18,031          | 87                 | 2011–08     | no      |
| qYeUpAodBbo | Fluoride in Drinking water harmful?                                   | 39              | 0                  | 2014–05     | no      |
| nEASlPkuRw0 | Water in Flint Michigan Looks Like Urine                               | 13              | 0                  | 2015–10     | no      |

Total views 62,513
cluster reach a significant level. Some videos are no longer available (i.e., they are set to private), which indicates that a video is politically and personally sensitive.

Keep in mind that whistleblowing is a risk for whistleblowers, and continued whistleblowing causes a recurrence of the danger, i.e., a whistleblower reports a danger and puts himself in a dangerous situation.

To address the second issue (automatic filtering), we tried to isolate common patterns between subsets of videos. In our case, the patterns come from basic features (i.e., noun phrases, n-grams) (Table 4). We selected a subset related to a single topic (about 752 videos), such as vaccines, and we found some typical recurrent features, such as those described in Fig. 5 and the algorithm in section 4.4. The set of 585 videos is unknown from cluster computation and consists of 435 videos related to vaccines only; for example, some people talk about “today I get Covid vaccine.” There are 149 videos related to whistleblowing and vaccines (we used these keywords to be sure, and deleted videos that could be noise or related to conspiracies by checking to see if they used keywords such as “alien,” “nazi” and “illuminati” in the title). We call these features maximal covering features, and they can be used as a “knowledge linguistic base” to make a filtering rule and to find new incoming videos on the same topic. This is why we use the Jaccard distance, which is explained in section 4.4. The combination of the Jaccard and cosinus metrics is robust, although the Jaccard metric captures most of the useful information.

Then, we apply a rule containing one or more strings from the most common multiword (i.e., repeated segments; see Table 4). To test the efficiency of the filtering, we make a test dataset based on a query containing two primary keywords and unknown videos posted at a time outside from time range of our dataset: we use the keywords “whistleblower and vaccine” and get videos from 2020 and 2021, mainly (Fig. 5).

We found 151 videos. After cleaning, we found that two videos were related to conspiracies. So, we found 149 videos related to our given topic and whistleblowing (see Table 5). We investigate how the filtering can differentiate between a standard vaccine video and a video related to whistleblowing and vaccines. The dimension of the common word space computed from the training cluster dataset is 750,782 multiwords (the size of the training dataset is 173 videos). The goal here is to minimize the false positives coming from the vaccine group and maximize the true positives coming from the whistleblowing-vaccine group. Let consider observation of filtered documents, ideally 0% of documents should be filtered in the subset of non-whistleblowing documents, and 100% in the whistleblowing documents. Our algorithm separates the two groups by a difference of 20% as filtered

| Representation                  | Feature Examples                                                   |
|---------------------------------|-------------------------------------------------------------------|
| 5-g                             | centers for disease control of                                    |
| 5-g                             | for disease control and prevention:                               |
| 5-g                             | the centers for disease control                                   |
| Noun Phrase                     | infectious/JJ disease/NN                                           |
| Noun Phrase                     | infectious/JJ diseases/NN                                          |
| 2-g                             | vaccines are                                                      |
| 2-g                             | vaccines and                                                      |
| 2-g                             | vaccines causes                                                   |
| Dependency Tree                 | compound:vaccine_safety                                           |
| Dependency Tree                 | compound:vaccine_schedule                                         |
documents for the threshold $T = 1.1$ (see Table 6); hence, we obtain an accuracy of 57.4%. A baseline would give an accuracy of 50% using random binary classification. This prediction is efficient but can still be improved. Some whistleblowing videos talk about vaccines but also discuss other topics. This can explain why prediction can fail in some cases, and also why the local grammar (frequent common lexical associations) that we use in the common word space is not always spoken by all whistleblowers.

In terms of semantic interpretation using whistleblower names, we found information about 21 out of 30 names (70%) from the list of famous whistleblower names. A majority of the names appear in the dataset. We can study the context of these names. Table 7 shows the 20 words that appear most frequently in the four words before and after “Snowden.” The main words we can interpret as important (other than “whistleblower”) are “Edward” (his first name) and “NSA” (his former employer). We can see that his name also occurs in the same context.

Tables 8 and 9 show the top 20 context words according to the number of times they occur and for each position (the top 20 words at position $-1$, top 20 words at position $-2$, and so on).

![Diagram](image.png)

**Fig. 5** As shown in the whistleblowing detection algorithm presented in section 4.3, we base the filtering and whistleblowing classification on the extraction of features from whistleblowing clusters. We then compute the distances (which we combine) to decide whether an unknown video is about whistleblowing or not.

**Table 5** Cluster of videos from the vaccine topic group

| ID            | Title                                                                 | Date      |
|---------------|----------------------------------------------------------------------|-----------|
| 1ZnyXkLWSIQ   | Amazing Polly BARDA Whistleblower Links zur Gesundheitsmafia          | 2020-05-19|
| QDfgXa5g6KM   | Can Trump keep this whistleblower quiet? Covid-19 whistleblower tells all | 2020-05-07|
| 5F0cd-w5fZk   | Chatter #170 - Ryan Hartwig on the Censorship Inside Facebook & How Big Tech Is Corrupting Our World | 2021-07-22|
so we get eight ranking lists rather than one. Thus, the context is more precise in terms of the position around the target word (“snowden”). In Tables 8 and 9, in red and yellow, we can see interesting words that also occurred in the global top 20 ranking. Words in green are new words and enhance the list of interesting words. For instance, we see the verb “to say” and the nouns “revelations” and “government.” This information is an important part of the positional embedding around a target word.

### Table 6
Filtering of videos concerning the vaccine topic. We have two different groups: one is related to whistleblowing and vaccines ($G_w$) and the other is related only to vaccines ($G_t$).

| Video Group | Number of Documents | % of documents seen as positives | Accuracy |
|-------------|---------------------|----------------------------------|----------|
| $G_t$       | 603                 | 45%                              |          |
| $G_w$       | 149                 | 64.7%                            |          |
| **Total**   | **752**             |                                  | **57.4%**|

### Table 7
Top 20 words in the same context as “snowden”.

| Word: occurrences | Word: occurrences |
|-------------------|-------------------|
| edward: 31758     | with: 4712        |
| the: 10170         | in: 4224          |
| s: 9421            | that: 4124        |
| is: 6530           | of: 3385          |
| whistleblower: 6103| for: 2862        |
| and: 6003          | has: 2846         |
| to: 5722           | team: 2558        |
| nsa: 5402          | shir: 2538        |
| a: 5158            | on: 2523          |
| snowden: 4945      |                   |

### Table 8
Top 20 words at positions ($−4, −3, −2, −1$) to the left of “snowden”.

| Word: occurrences | Word: occurrences |
|-------------------|-------------------|
| edward: 31758     | with: 4712        |
| the: 10170         | in: 4224          |
| s: 9421            | that: 4124        |
| is: 6530           | of: 3385          |
| whistleblower: 6103| for: 2862        |
| and: 6003          | has: 2846         |
| to: 5722           | team: 2558        |
| nsa: 5402          | shir: 2538        |
| a: 5158            | on: 2523          |
| snowden: 4945      |                   |
6 Conclusion

The Internet involves the daily contributions of a large number of users through various platforms, including social media platforms. Social media reveals the activism of specific people that we call whistleblowers; whistleblowing is an important action, as a counterpower, for the well-being of our societies. Using the YouTube platform, we were able to make a corpus of 644,575 video transcriptions from 2007 to 2018, which were cleaned and reduced to 35,730 documents. The raw dataset contains many rumors, conspiracies and various other documents due to the ambiguity of the query word. Our goal was to explore the content of the cleaned dataset using natural language processing and machine learning techniques. We have extracted different linguistic representations from this dataset, including different lengths of n-grams, pattern noun phrases, named entities and dependency noun phrases (in total, 646 million features). We applied clustering techniques to these features to enable topic detection; we found 18 subtopics. The topic-aggregated number of views can indicate a large audience (several thousand users). We were able to isolate some linguistic patterns that we can interpret as a topic-oriented whistleblowing local grammar. These patterns can be used to generate rules to perform filtering to find new incoming video documents concerning the same whistleblowing topic. When we performed testing on 585 video documents that are not in the dataset used for rule generation, we obtained a 57.4% accuracy for classifying new whistleblowing videos for a specific topic. Based on interpretative semantic analysis, we studied famous whistleblower names and discovered that positional ranking lists around a target name enable the retrieval of a greater number of context-informative words. This analytical framework can be useful for the explainability of the attention mechanism in positional embedding. YouTube’s API has stronger restrictions due to privacy concerns, but we believe it remains a good source of information for studying net activism.

Acknowledgments We wish to thank all anonymous reviewers of this special issue of Multimedia Tools and Applications (MTAD), directed by professors T.D. Singh, S. Thokchom, L.D. Singh and B.K. Balabantaray, for their relevant comments. Our thanks also go to Jacques Thomazeau for cluster assistance and to Saber Nouioua for software engineering.
Author contribution conception – programming – writing.

Funding The CNRS (French National Institute of Scientific Research) provided a PEPS (exploratory project) grant entitled EXIA (information extraction and applications) (grant number UPEM-PEPS-2015).

Data availability The dataset is freely available on the open-science Zenodo platform (funded by the European Union) with the following doi: https://doi.org/10.5281/zenodo.5824627.

Code availability The code is available on the open-science Zenodo platform (funded by the European Union) with the following doi: https://doi.org/10.5281/zenodo.5825530. Most of the codes run on a big data platform for parallelization using the SLURM framework and Hadoop framework.

Declarations

Conflict of interest None.

References

1. Aal K, Krüger M, Rohde M, Tadic B, Wulf V (2019) Social media and ICT usage in conflicts areas. In: Reuter C (ed) Information technology for peace and security. Springer Vieweg, Wiesbaden. https://doi.org/10.1007/978-3-658-25652-4_18
2. Agarwal S, Sureka A (2014) A focused crawler for mining hate and extremism promoting videos on YouTube. HT 294–296
3. Agarwal S, Sureka A (2015) Topic-specific YouTube crawling to detect online radicalization. In: Chu W, Kikuchi S, Bhalja S (eds) 10th international workshop on databases in networked information systems (DNIS). Univ. Aizu, Aizuwakamatsu, Japan
4. Aggarwal N, Agrawal S, Sureka A (2014) Mining YouTube metadata for detecting privacy invading harassment and misdemeanor videos. In: Miri A, Hengartner U, Huang NF et al (eds) 12th Annual International Conference on Privacy, Security and Trust (PST). Ryerson Univ., Toronto, Canada
5. Algur SP, Bhat P (2016) Abnormal web video detection using density based LOF method. Int J Comput Sci Eng 4(4):6–14 E-ISSN 2347–2693
6. Andre V (2014) The Janus face of new media propaganda: the case of Patani Neojihadist YouTube warfare and its Islamophobic effect on cyber-actors. Islam Christ-Muslim Relat 25(3):335–356
7. Askanius T, Uldam J (2011) Online social media for radical politics: climate change activism on YouTube. Int J Electron Gov 4(1/2):69–84
8. Askanius T, Uldam J (2012) Political identity and radical politics: mediating COP15 activism on YouTube. In: Mediation and protest movements. Intellect books, Bristol & Wilmington
9. Baratz M (2009) Twitter, YouTube among the new propaganda tools of the Battle in Gaza. DigitalMedia
10. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
11. Boltanski L, Thévenot L (1991) De la justification. Les économies de la grandeur [On justification. Economy of excellence]. Gallimard, Paris
12. Bosch T (2012) YouTube now lets you blur faces in videos: will it help keep human rights activists safe? Slate. http://www.slate.com/blogs/future_tense/2012/07/18/youtube_announces_tool_for_blurring_faces_in_videos_to_help_human_rights_activists.html. Accessed Sept 2022
13. Bourigault D, Jacquemin C (1999) Term extraction + term clustering: an integrated platform for computer-aided terminology. In: Proceedings of conference of the European chapter of the Association for Computational Linguistics (EACL)
14. Bbric M, Rozic E, Zarko IP (2012) Recommendation of YouTube videos. In: Biljanović P et al (eds) Proceedings of the 35th international convention. Croatian Society for Information and Communication Technology, Electronics and Microelectronics (MIPRO), Rijeka, pp. 1775–1779
15. Chatenaynaud F, Tomny D (1999) Les sombres précurseurs: Une sociologie de l’alerte et du risque [Dark forerunners: a sociology of alert and risk]. Éditions de l’EHESS, Paris
16. Chowdury R, Adnan Md, Monsur N, Mahmud GAN et al (2013) Data mining based spam detection system for YouTube. In: Bajwa IS, Naeem A, Pichappan P (eds) 8th international conference on digital information management (ICDIM). Islamabad, Pakistan
17. Conway M, McInerney L (2008) Jihadi video and auto-radicalisation: evidence from an exploratory YouTube study. In: Ortiz-Arroyo D, Larsen HL, Zeng DD et al (eds) 1st European conference on intelligence and security informatics. Esbjerg, Denmark
18. Domingos P (1999) The role of Occam's razor in knowledge discovery. Data Min Knowl Disc 3:409–425
19. Dubisar AM (2015) Embodying and disabling antiwar activism: disrupting YouTube’s "Mother’s day for peace". Rhetor Rev 34(1):56–73
20. Efling R, van Hillegersberg J, Huibers TWC (2011) Social media and political participation: are Facebook, twitter and YouTube democratizing our political systems? ePart 25–35
21. Eltahawy M (2010) Facebook, YouTube and twitter are the new tools of protest in the Arab world. Washington Post
22. Fernando A (2009) The revolution will be mashed up (and uploaded to YouTube). Communication World
23. Garimella K, De Francisci Morales G, Gionis A, Mathioudakis M (2017) Quantifying controversy in social media. In: Proceedings of the 10th ACM international conference on web search and data mining (WSDM)
24. Githens-Mazer J (2010) Radicalisation via YouTube? It’s not so simple. The Guardian
25. Hall W, Tinati R, Jennings W (2018) From Brexit to Trump: social media’s role in democracy. Computer 51(1):18–27. https://doi.org/10.1109/MC.2018.1151005
26. Halpern D, Gibbs J (2013) Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. Comput Hum Behav 29(3):1159–1168
27. Kavada A (2012) Engagement, bonding, and identity across multiple platforms: Avaaz on Facebook, YouTube, and MySpace. MedieKultur J Media Commun Res
28. Khan ML (2017) Social media engagement: what motivates user participation and consumption on YouTube? Comput Hum Behav 66:236–247 ISSN 0747-5632
29. Klausen J, Barbieri ET (2012) The YouTube jihadists: a social network analysis of Al-Muhajiroun’s propaganda campaign. Perspect Terror 6(1):36–53
30. Konapure RC, Lobo LMRJ (2021) Video content-based advertisement recommendation system using classification technique of machine learning. J Phys Conf Ser 2021:012025
31. Konstantinidis S, Fernandez-Luque L, Barmidis P, Karlsten R (2013) The role of taxonomies in social media and the semantic web for health education: a study of SNOMED CT terms in YouTube health video tags. Methods Inf Med 52(2):168–179
32. Laine SMS, Ercai G, Luo B (2011) User groups in social networks: an experimental study on YouTube. In: 44th Hawaii international conference on system sciences (HICSS), pp 1–10
33. Lebart L, Salem A, Berry L (1998) Exploring textual data. Springer, Netherlands. https://doi.org/10.1007/978-94-017-1525-6
34. Liao H, McDermott E, Senior A (2013) Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. In: IEEE workshop on automatic speech recognition and understanding (ASRU), pp 368–373. https://doi.org/10.1109/ASRU.2013.6707758
35. Liikkanen LA (2013) Three metrics for measuring user engagement with online media and a YouTube case study. CoRR abs/1312.5547
36. Liu J, Wu C (2017) Deep learning based recommendation: a survey. In: Kim K, Joukov N (eds) Information Science and Applications (ICISA) 2017. Lecture notes in electrical engineering, vol 424. Springer, Singapore. https://doi.org/10.1007/978-981-10-4154-9_52
37. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, vol 1, pp 281–297
38. Magdy W, Sajja HD, El-Ganainy T, Sebastiani F (2015) Distant supervision for tweet classification using YouTube labels. In: International AAAI conference on web and social media (ICWSM), pp 638–641
39. Mahmood S (2012) Online social networks: the overt and covert communication channels for terrorists and beyond. In: 2012 IEEE conference on Technologies for Homeland Security (HST), pp 574–579
40. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the Association for Computational Linguistics: system demonstrations, Baltimore, Maryland
41. Mansoor A-R (2019) FrameProv: towards end-to-end video provenance. In: NSPW ’19: proceedings of the new security paradigms workshop, pp 68–77. https://doi.org/10.1145/3368860.3368866
42. Marwick AE (2007) The People’s republic of YouTube? Interrogating rhetorics of internet democracy. Rhetorics of Internet Democracy
43. Mekaru SR, Brownstein J (2014) One health in social networks and social media. Revue Scientifique et Technique–Office International des Epizooties 33(2):629–637
44. Mosernghdvishvili L, Jansz J (2013) Framing and praising Allah on YouTube: exploring user-created videos about Islam and the motivations for producing them. New Media Soc 15(4):482–500
45. Myungha J, Foley J, Dori-Hacohen S, Allan J (2016) Probabilistic approaches to controversy detection. In: Proceedings of the 25th ACM international on conference on information and knowledge management (CIKM), pp 2069–2072. https://doi.org/10.1145/2983323.2983911
46. O’Callaghan D, Harrigan M, Carthy J, Cunningham P (2012) Network analysis of recurring YouTube spam campaigns. International AAAI conference on web and social media (ICWSM)
47. Oh S, Syn SY (2015) Motivations for sharing information and social support in social media: a comparative analysis of Facebook, twitter, delicious, YouTube, and Flickr. JASIST 66(10):2045–2060
48. Poell T, Borra E (2012) Twitter, YouTube, and Flickr as platforms of alternative journalism: the social media account of the 2010 Toronto G20 protests. Journalism 13(6):695–713. https://doi.org/10.1177/1464884911431533
49. Puente SN, Romero DF, Garcia RR (2015) Online activism and subject construction of the victim of gender-based violence on Spanish YouTube channels: multimodal analysis and performativity. Eur J Women Stud 22(3):319–333
50. Ramos-Munoz JJ, Prados-Garzon J, Ameigeiras P et al (2014) Characteristics of mobile YouTube traffic. IEEE Wirel Commun 21(1):18–25
51. Riloff E (1996) An empirical study of automated dictionary construction for information extraction in three domains. Artif Intell 85(1–2):101–134. https://doi.org/10.1016/0004-3702(95)00123-9
52. Rodriguez-Fuentes L, Penagarikano M, Varona A et al (2014) KALAKA-3: a database for the recognition of spoken European languages on YouTube audios. In: Calzolari N, Choukri K, Declerck T et al (eds) 9th International Conference on Language Resources and Evaluation (LREC). Reykjavik, Iceland
53. Sandova-Almazán R, Gil-García JR (2013) Cyberactivism through social media: twitter, YouTube, and the Mexican political movement “I’m number 132”. HICSS 1704–1713
54. Trotter D, Fuchs C (2015) Social media, politics and the state: protests, revolutions, riots, crime and policing in the age of Facebook, Twitter and YouTube Routledge ISBN 9781138798243.
55. Uldam J, Askanius T (2013) Online civic cultures: debating climate change activism on YouTube. Int J Commun 7:1185–1204
56. Ungerfeider N (2011) Massive Egyptian protests powered by YouTube, twitter, Facebook, Twitpic. Fast Company
57. Ushkin SG (2014) User comments on the protest actions in the Russian segment of YouTube. Sotsiologicheskie Issledovaniya:127–133
58. Vergani M, Zuev D (2011) Analysis of YouTube videos used by activists in the Uyghur nationalist movement: combining quantitative and qualitative methods. J Contemp China 20(69):205–229 PII 933184845
59. Vergani M, Zuev D (2015) Neojihadist visual politics: comparing YouTube videos of North Caucasus and Uyghur militants. Asian Stud Rev 39(1):1–22

Publisher’s note  Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.