Bozarth and Budak *EPJ Data Science* (2022) 11:30
https://doi.org/10.1140/epjds/s13688-022-00343-9

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Keyword expansion techniques for mining social movement data on social media

Lia Bozarth[1]* and Ceren Budak[1]

*Correspondence:
lbozarth@umich.edu
[1] School of Information, University of Michigan, Ann Arbor, US

Abstract

Political and social scientists have been relying extensively on keywords such as hashtags to mine social movement data from social media sites, particularly Twitter. Yet, prior work demonstrates that unrepresentative keyword sets can lead to flawed research conclusions. Numerous keyword expansion methods have been proposed to increase the comprehensiveness of keywords, but systematic evaluations of these methods have been lacking. Our paper fills this gap. We evaluate five diverse keyword expansion techniques (or pipelines) on five representative social movements across two distinct activity levels. Our results guide researchers who aim to use social media keyword searches to mine data. For instance, we show that word embedding-based methods significantly outperform other even more complex and newer approaches when movements are in normal activity periods. These methods are also less computationally intensive. More importantly, we also observe that no single pipeline can identify little more than half of all movement-related tweets when these movements are at their peak mobilization period offline. However, coverage can increase significantly when more than one pipeline is used. This is true even when the pipelines are selected at random.

Keywords: Keyword expansion; Query expansion; Data mining; Twitter; Social movement

1 Introduction

Twitter is one of the most popular online platforms for studying social and political movements [1–3]. Researchers routinely leveraged Twitter data to analyze a single high-profile movement [4, 5], compare and contrast multiple movements (e.g., movement vs. countermovement) [6, 7], and evaluate different types of participants (e.g., organizations, slacktivists) in social movements [8, 9]. Twitter's unmeasured popularity for social movement research is, in no small part, due to the easy access to its data. This popularity is likely to continue and grow since Twitter has made its data freely available for academic research [10]. This trend underlines the importance of systematically evaluating the effectiveness of automated approaches for identifying social movement-related content on Twitter. This is the challenge we take on in this paper.

Thus far, researchers have predominantly relied on keywords (e.g., hashtags) to mine relevant Twitter datasets [11–13]. However, prior studies [1, 14] have demonstrated comprehensiveness and representativeness issues with this approach, and the risks of reaching flawed research conclusions [1, 15–17]. For instance, [16] shows that keyword-based text mining may oversample Twitter users with extreme political identities. Various methods have been proposed to improve data mining through *keyword expansion*. Most notably, researchers have proposed term and document frequency-based methods [18] and their variations, such as Chi-square with document frequency [19]. Others have proposed vector similarity-based approaches including co-occurrence vectors [20] and word embedding approaches [21]. Additionally, scholars have also suggested more sophisticated classifier-based pipelines for keyword expansion [22, 23]. Most notably, King et al. [22] use an ensemble of models to capture a more diverse range of potential keywords. Similarly, Linder [23] uses two separate classifiers to maximize their pipeline's recall score and improve keyword diversification.

These are all noteworthy contributions. However, these methods have been evaluated using limited data in terms of diversity and size, and few comprehensive comparisons have been made to rank these diverse methodologies. Thus, there is a distinct lack of benchmarks. Our work aims to fill this gap by asking the following question: given a large search dataset (e.g., 1 month of Twitter Decahose tweets), and a set of initial keywords related to a movement, how well can existing pipelines retrieve movement-related content from the search dataset by identifying additional keywords?

To answer this question, we select several distinct keyword expansion pipelines. We then assess each pipeline's performance across a representative list of contemporary social movements. We use evaluation metrics and criteria informed by related work. To elaborate, we identify five distinct pipelines that collectively cover broad categories of past literature proposed in this space as well as extensions that incorporate recent advances in natural language processing. Specifically, we include one pipeline that uses term and document frequency weigh-in, another that incorporates word embedding, and finally, three classifier-based approaches proposed by related work [22, 23]. The classifier-based approaches consist of both deep learning and non-deep learning models (see detailed description in Sect. 2.2).

Next, we identify a list of representative social movements. We use the popular dataset from [24]. It contains tens of thousands of offline event records for 20 ongoing social movements. We cluster these movements using their offline attributes (e.g., the number and size of offline events), and then select a representative subset. Specifically, we select the following five movements: *lgbtq* (e.g., support/against lgbtq rights, pride parades), *government* (e.g., support/against government actions, domestic and foreign policies, or specific institutions and officials), *guns* (e.g., support/against gun-control), *immigration* (e.g., support/against DACA, support/against building the border wall) and *healthcare* (e.g., supporting/against expanding healthcare). By using these distinct movements, we can deliver a more comprehensive assessment of each pipeline's performance. Additionally, prior literature shows that social movements rise and fall, and their languages change accordingly [6]. As such, we conduct separate evaluations for each movement using two offline activity levels: high activity (i.e., a movement at its peak) and base activity (i.e., a movement at its normal activity level). This allows us to determine whether a pipeline's performance varies across time periods when the movement is in maintenance or growth.

Finally, we evaluate each pipeline's performance using the following metrics (i) tweet and user coverage; (ii) content similarity and topic coverage; (iii) temporal similarity and bursts; and (vi) resource requirements. The first three metrics are informed by the most popular types of analyses in social movement research (see details in Sect. 2.1), and the last metric addresses real-world resources constraints.

Since there is no gold standard ground truth set that contains all related tweets for any given movement, in our work we define ground truth as the union of all relevant tweets returned by the pipelines for a fixed movement and activity level. Under this definition, a tweet is movement-related if it contains at least one movement-related keyword. In other words, ground truth is generated at the keyword level. While this definition is inherently noisy, it is interpretable and the errors are transparent [22, 23]. For example, it's easy to tell which keywords are noisier than others. Researchers can further clean up the collected data using different strategies (e.g., building supervised classifiers to filter out irrelevant tweets). These valuable efforts are outside the scope of our study.

Here, we measure pipeline performance against the aforementioned ground truth. Our results demonstrate that the word embedding-based approach significantly outperforms other pipelines for movements in their base activity periods. In fact, this technique on average can identify 75% of tweets identified by all pipelines (i.e, the ground truth set) for movements in base activity level. The next highest performing pipeline is only able to identify an average of 46% of the ground truth set. Additionally, the word embedding approach also has the broadest coverage of coherent topics extracted from movement tweets. For instance, for *lgbtq* in base-activity, at least 25% of tweets for 94% of all topics are identified by this approach. This is true for only 75% of all topics for the next best methodology. Furthermore, the word embedding approach requires significantly fewer computing resources than other more advanced pipelines.

We also observe that no single pipeline significantly outperforms others when movements are in their peak offline activity level. The highest coverage by any one pipeline on average is only 62%. This is likely because offline protests bring a more diverse group of participants into movements, which results in the content being diverse and, thus, hard to capture by a single approach. Results here suggest that researchers should select multiple pipelines to gather a more comprehensive dataset. Our similarity analysis identifies clusters of pipelines as a function of the overlap and similarity of the content they capture. These clusters help identify a strategy for choosing subsets of data mining pipelines to maximize coverage. For instance, combining the word embedding-based method with a classifier-based method has better coverage than using two classifier-based methods, though even randomly selecting at least two pipelines significantly boosts data coverage.

Furthermore, pipelines are generally better at identifying heavy participants who frequently post movement-related tweets than light participants (users who only post 1 or 2 movement-related tweets). This is especially true when a movement is growing in popularity (e.g., an average pipeline is able to identify 75% of heavy participants, but only half of light participants). This result is similar to related work [16], and suggests that keyword-based data mining is more likely to miss peripheral participants.

Finally, we show that the time-series of data extracted by these pipelines is largely similar. Still, they significantly differ from a movement's offline activity – irrespective of the data mining pipeline used. In other words, the choice of the keyword expansion method does not significantly impact temporal analysis. However, researchers need to keep in mind that these temporal patterns are not necessarily reflective of the dynamics of the offline movement.

2 Related work

In this section, we first review Twitter data-based studies on a wide range of social and political movements. We focus on identifying these studies' data mining techniques and types of downstream analyses. Next, we focus on the most popular data collection approach – keywords and keyword expansion – in greater detail. We describe the five pipelines we have chosen for evaluation.

2.1 A condensed review of social movement studies using Twitter data

To understand researchers' reliance on different data mining techniques, we first obtain lists of the most popular movements of the 2010s from [25, 26]. We then search Google Scholar[1] for published academic manuscripts using the terms "twitter" + x where x is a given movement. We limit search results to the first two pages and also filter out qualitative research (e.g., studies based on interviews). Using this approach, we obtain a total of 53 social movement research papers using Twitter data. These papers were published between 2012 and 2021 (the median is 2018). We manually review the data collection technique used in each paper. While this process does not generate a comprehensive set of social movement literature, it provides valuable insights into the range of data collection methods most commonly used in social movement research. We observe that 44 (or 83%) of the papers relied on keywords to obtain Twitter data. Only 5 (or 11.3%) of the 44 papers also used keyword expansion methods. Next, 8 (or 15%) relied on user accounts (e.g., identifying prominent Twitter accounts within a movement and extracting their tweets [27]), and finally 1 (or 2%) relied on enterprise tools with unspecified mining scope. Results here illustrate the popularity of the keyword-based data collection approach.

Next, we examine and identify types of popular downstream analyses. We observe that the most popular type of analysis – performed in 33 (or 62%) of the 53 papers – was tweet classification-based analysis. In these papers, researchers manually coded tweets according to specific theoretical frameworks, or they leveraged automated approaches such as topic modeling to identify tweet topics [2, 28]. The second most popular is time-series analysis (36% papers). It is followed by network-based analysis (30% papers), user categories and demographics-based analysis (28% papers), and finally word frequency/word distribution-based analysis (17% papers). While this list of most common analyses is not comprehensive, it helps us identify the most important inquiries a data collection system ought to support. As such, we use results here to determine three of the most useful metrics for evaluating the performance of different keyword expansion techniques. The analyses we present in this paper evaluate pipelines based on measures related to all aforementioned analyses, with the exception of network analysis.

2.2 Keywords and keyword expansion techniques

The previous section shows that social movement studies commonly use predetermined keywords to mine Twitter data. More importantly, past work also suggests that data collection (e.g., data comprehensiveness) can be significantly improved through keyword expansion [22, 23]. As such, our paper primarily relates to keyword expansion methods. Here we provide a summary of research in this area.

[1] https://scholar.google.com/

The most common keyword expansion approach is to simply expand the list by identifying other keywords (also called candidate keywords) that commonly co-occur with the original set of keywords (e.g. [20, 29–31]). The set of candidate keywords is manually inspected to filter out the ones that are unrelated. This process is commonly applied once, identifying a second set of related keywords given an initial set. While it is possible to generalize the framework to an arbitrary k-step process where the seed keywords are expanded in each step, this is not common in the literature.

Researchers have also explored more sophisticated approaches to perform keyword expansion using three common methodological paradigms: (i) term and document frequency based approaches [18, 19, 32, 33]; (ii) word embedding based methods [21, 34]; and (iii) classification models [22, 23, 35]. Some studies combine multiple paradigms. For instance, Req-rec [36] uses Rocchio for query generation and classifiers for filtering. Our work incorporates at least one existing method from each paradigm for performance evaluation. Specifically, we include five approaches as described below. We also fine-tune these approaches to fit our experiment setup (see details in Sect. 3.2).

TFIDF The term frequency–inverse document frequency (TFIDF) is a weighting scheme that measures the importance of a word within a given set of documents [18]. It is commonly used in information retrieval, text summary, and text mining. Term frequency (TF) indicates the number of occurrences of a word in a document. Inverse document frequency (IDF) is an inverse function of the number of documents that contain a word. As an example, the word "so" is one of the most common words in English. As such, it commonly has a high TF score and a very low IDF score across different documents and corpora. Conceptually, higher TF and IDF scores indicate greater word importance. It's one of the simplest methods and is often used as a baseline in information retrieval and keyword extraction tasks [20].

Word2Vec Word embedding models typically use a simple 2-layer neural network to convert each word into a vector representation. For instance, the word "data" can be represented with the vector $[-0.0122, -1.267, 0.004, \ldots]$. Additionally, words that share similar semantic meanings have a higher cosine similarity between their vectors. To illustrate, the vector representation for "blacklivesmatter" has a significantly higher cosine similarity to the vector for "blacklivematters" than "university". Word2Vec essentially uses word co-occurrences to determine semantic similarity. For example, the words "apple" and "orange" are frequently followed by the same words (e.g., "juice"). That is, both "apple" and "orange" frequently co-occur with the word "juice". Word2Vec models rely on such co-occurrences to generate vector representations such that the vectors for "apple" and "orange" end up being similar. Researchers often rely on Word2Vec to generate similarity measures for text documents [21, 37]. Thus far, the most commonly used word embedding methods include GloVe [38] and FastText [39]. Additionally, FastText was shown to have better performance than GloVe [39]. For our work, we use a modified version of [21]. Specifically, we use FastText word embedding to obtain additional keywords that are semantically similar to the initial set of seed keywords.

King This approach was originally proposed by King et al. [22] to identify keywords in unstructured text data. It incorporates both human input, and labels from an ensemble of

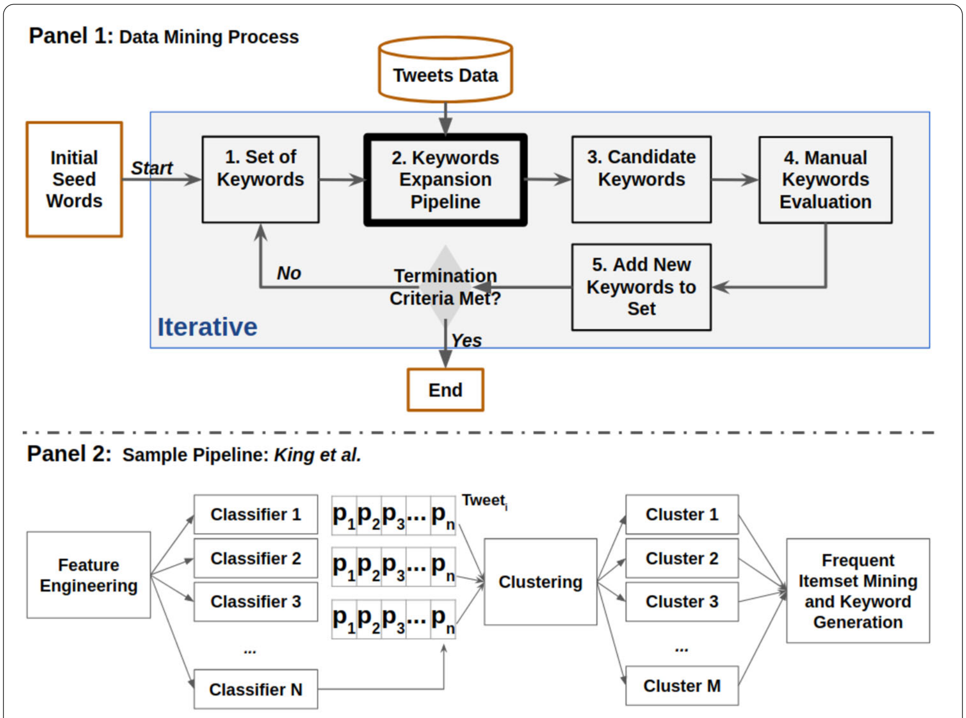


Figure 1 Overview of the Data Mining Process (Panel 1) and Sample Keyword Expansion Pipeline (Panel 2). Note, Sect. 3.1 describes the data collection process for *tweets data* and *initial seed words*. Section 3.2 outlines the setup for *keyword expansion pipelines*. Finally, Sect. 3.3 contains *keywords evaluation and selection* steps

classifiers. It was presented as a better alternative to fully manual (i.e., researchers selecting keywords in ad hoc ways) or fully automated approaches for running keyword expansion based data mining. This method consists of many steps. A flowchart of this complex pipeline is shown in Fig. 1 (Panel 2). We briefly describe it here and refer readers to the original paper for details. Given a set of relevant documents D and a search dataset S (for example, D is a set of tweets that contain the "blacklivesmatter" keyword, and S is all the tweets from June 2020), the pipeline first trains a set of classifiers to label (with probability) each tweet $i \in S$ as relevant or irrelevant. Next, the tweets are algorithmically assigned into different clusters according to classifiers' labels. For each cluster, King et al. then apply frequent itemset mining using the Apriori algorithm to generate potential keywords. They then rank the keywords and return the top keywords that best characterize the cluster. Finally, they manually examine these potential keywords to determine their relevance.

Linder This approach was originally proposed by Linder [23]. Similar to King et al. [22], this method also incorporates both human input and labels from automated models. The authors argued that it has better empirical results, notably higher recall scores, than the one presented by King et al. [22]. Particularly, Linder [23] noted that the original likelihood function that ranks potential keywords proposed in [22] contributed to lower performance. We briefly describe this approach here and refer readers to the original paper for details. Given the initial set of relevant documents D and search dataset S, this approach first trains a stochastic gradient descent (SGD) model using unigram features. The original paper chooses to use the SGD algorithm over other approaches, such as Random Forest and Gradient Boost, because SGD is the highest-performing model as measured using F1 score. The SGD model is used to label each tweet $i \in S$ as relevant or irrelevant. Let D' and S' be the updated results after the classification step. That is, D' is the union of D and the additional documents labeled by the SGD model as relevant. The author then uses Lasso regression to determine unigrams that have the highest odds of appearing in D' than S'. The author uses Lasso because it's able to diversify keywords. That is, if multiple words are highly correlated, Lasso only suggests one of them. This set of potential keywords is then manually reviewed by the author to determine their relevance.

Linder-BERT This approach is the same as *Linder*, except this pipeline replaces the SGD Classifier with DistilBERT [40], a fast and light-weight deep-learning based model. We hypothesize that a more advanced, deep-learning based classifier can potentially boost the performance of the original method proposed by Linder [41].

In sum, many existing social movement studies using Twitter data are heavily reliant on keywords to collect movement-related content. Additionally, researchers have proposed many keyword expansion pipelines to improve data mining [22, 23] to identify movement-related content. However, the authors who proposed these methods used different datasets and metrics for performance assessment. The absence of systematic evaluation makes it difficult to know which pipelines have better empirical results. In other words, there is a lack of benchmark studies. In this work, we evaluate the performance of five distinct pipelines across a diverse set of social movements, providing insights for social movement researchers interested in collecting better quality data from Twitter.

3 Experiment setup

An overview of the keyword expansion-based data mining process is shown in Fig. 1 (Panel 1). Keyword expansion pipelines (see example pipeline in Panel 2 of Fig. 1) use an initial set of seed keywords and a search dataset containing all available documents to generate new keywords. The subset of these keywords deemed to be relevant to the movement will then be supplied back into the pipeline. This repetition (or loop) terminates when the stopping criteria are met. The researcher can choose stopping criteria based on aspects such as incurred costs (e.g. labeling costs), marginal gain in newly identified content or a combination of the two.

In this section, we first describe our data and the process of generating initial seed keywords (see orange components in Panel 1 of Fig. 1). We then describe our setup for the 5 keyword expansion pipelines (see Components 1, 2, and 3 in Panel 1 of Fig. 1). Finally, we illustrate our procedure of using Amazon Mturk to manually identify true positive keywords. (see Component 4 in Panel 1 of Fig. 1)

3.1 Data

We use 3 types of data in our work: (i) movements and their offline protests, (ii) Twitter Data, and (iii) seed keywords collected using the first 2 datasets.

3.1.1 Movements offline data

We use data provided by Crowd Counting Consortium (CCC) [24]. This dataset has been used by many prior social movement studies [25, 42–44]. The researchers at CCC compile social and political crowd events in the U.S. including protests, rallies, strikes, etc. Each event has a corresponding record including date, location, the number of participants, organizers, issue area (e.g., climate change, migration, guns), source of information, etc. Between February, 2018 and April, 2019, the CCC aggregated a total of 18.7K offline events. Similar to [45], we assign these events into 20 distinct social movements using each event's primary objective or issue area. For instance, CCC recorded that over a hundred protesters had rallied against the deportation of a local man in St. Petersburg, Florida on Feb 1st, 2018. We use the issue area recorded by the CCC to assign this event to the *immigration* movement. Similarly, over 300 people in Boise, Idaho protested in support of expanding healthcare coverage on Feb 4th, 2019. We assign this event to the *healthcare* movement.

Next, we use Principal Component Analysis (PCA) with varimax rotation to cluster these 20 movements according to the attributes of their offline events. These attributes include (i) the total number of events; (ii) total number of unique days with at least 1 crowd event; the (iii) average and (iv) max number of participants of these events; and finally (v) the number of unique claims. The resulting 5 clusters are plotted in Fig. 2. The x-axis and y-axis are the 1st and 2nd principal components respectively. As shown, the movements *lgbtq* (such as supporting/against lgbtq rights, pride parades), *government* (such as supporting/against government actions, domestic and foreign policies, or specific institutions and officials; against government corruption; demanding political action; supporting specific candidates in the 2018 midterm election), and *guns* (such as supporting/against gun-control, against gun-violence) significantly differ from all the other movements, each warranting its own cluster. The remaining movements are separated into 2 clusters. Factor analysis shows that the number of events, unique days of events, and the number of claims load into the first principal component. In other words, movements with larger x-

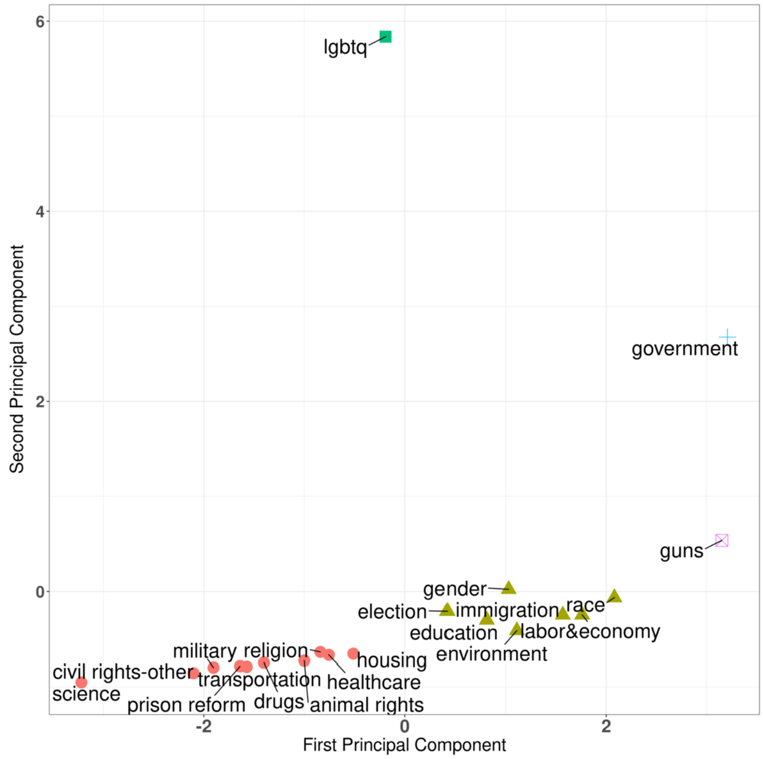


Figure 2 Movements Clustered Using Their Offline Attributes

Table 1 Summary Stats for Sampled Movement

Movement	Total protests	Ave protest size	Activity level	Month	Protest frac.	Seed word count
Government	2111	2049	high	2018-10	0.11	66
			average	2018-08	0.05	
Guns	8840	574	high	2018-03	0.82	34
			average	2018-06	0.01	
Healthcare	159	163	high	2018-05	0.15	12
			average	2019-03	0.09	
Immigration	1712	443	high	2018-06	0.60	20
			average	2018-07	0.07	
lgbtq	228	20,061	high	2018-06	0.26	30
			average	2019-04	0.06	

axis values have more events, and more varied causes of action or demands. Additionally, the average and max participation sizes load into the second principal component. That is, movements with larger y-axis values have broader participation.

For our work, we randomly select a subset of representative movements from each of the five clusters in Fig. 2: *lgbtq, government, guns, immigration*, and *healthcare*. As shown in Table 1, *guns* has a total of 8.8K offline events, significantly exceeding other movements; the *lgbtq* movement has the largest average protest size; and *healthcare* (such as for/against universal healthcare, bringing awareness to specific illness) has the fewest and smallest offline events.

3.1.2 Twitter data

Our second dataset is a 10% sample of all tweets known as the Twitter Decahose. From February, 2018 to April, 2019, this dataset contains an average of 92.8 million original tweets per month (excluding retweets). As such, processing the entire data is resource intensive. For our work, we sample 2 months of Twitter data for each movement according to 2 offline activity levels: (i) high-activity level (month when the largest fraction of offline events had taken place for a movement), and (ii) base-activity level (randomly sample a month with 1% to 10% offline events). Summary stats are located on Table 1. For example, 82% of all *guns*-related protest events between February, 2018 and April, 2019 took place during March, 2018.

This stratification results in 10 unique datasets (5 movements $*$ 2 activity levels), and allows us to test the robustness of existing keyword expansion pipelines as a function of the nature of the movement and the activity levels at the time of data collection.

3.1.3 Identifying the seed keywords

Similar to many prior studies focused on collecting Twitter data for political and social research [1, 2, 46, 47], we identify the most relevant *hashtags* for each movement as seed keywords. For each movement, we use 3 distinct processes to aggregate related hashtags. First, we identify five days with the largest number of offline events. We then manually review the most popular hashtags on these five days (minimum tweet frequency of ≥ 100) and select the relevant hashtags. Second, we identify the locations and dates of the large offline events (minimum participants ≥ 1000) and collect tweets with matching coordinates. We then manually review the most popular hashtags extracted from these tweets (minimum tweet frequency of ≥ 25). For our third approach, we first identify the Twitter accounts of the organizers of offline events by using Twitter's native *search_user* API to search for these organizers' names (e.g., "twin city pride"). We collect 2.6K potential Twitter accounts. We then review the most popular hashtags (minimum tweet frequency of ≥ 5) posted by these accounts, and manually identify the relevant ones. The third approach is particularly useful for movements with few small offline events like *healthcare*. The minimum tweet frequencies chosen for each of the three approaches are different, as noted above. These thresholds are selected primarily based on our expected level of manual labor (each approach should, at max, generate close to 1K hashtags for manual review).

Next, two people on the research team manually assessed whether the retrieved hashtags are related to the movement for a randomly selected movement. The inter-coder reliability showed substantial agreement (Cohen's kappa = 0.755). As such, following content analysis research procedures [48], we only have one of the researchers manually review all hashtags for the remaining movements. As a result of this process, we identify 66, 34, 12, 20, and 30 relevant seed keywords for *government, guns, healthcare, immigration*, and *lgbtq* respectively. Note that initial seed words are the same for a movement irrespective of activity levels given that numerous hashtags are consistently used in a movement. Additionally, while the initial keywords are all hashtags, the keyword expansion pipelines will identify new non-hashtag keywords.

3.2 Pipelines

Here, we discuss the setup for 5 existing pipelines included in our paper. This section corresponds to Component 1, 2, and 3 in Panel 1 of Fig. 1. Additionally, Sect. 2.2 contains the high-level description for each pipeline.

3.2.1 Word2Vec

Given a movement m, activity-level a, and its corresponding month of tweet data $D_{m,a}$, we first use FastText [39] to populate 300-dimension vectors for words with minimum frequency of 5 using $D_{m,a}$. Here, the maximum word length is set to 1. Then, for each keyword $k \in K^0_{m,a}$ where $K^0_{m,a}$ is the initial seed keywords for m at activity level a, we identify the top 10 words that have the highest vector cosine similarity to k (this constraint is set due to budgetary considerations). We then export the union of these words for crowdsourced manual labeling.

3.2.2 TFIDF

We first identify the subset of tweets in $D_{m,a}$ that contain at least 1 keyword k where $k \in K^0_{m,a}$. We denote this subset as $D'_{m,a}$. We then calculate the tf-idf scores for words that occur in at least 100 documents and at most 25% of all documents in $D'_{m,a}$. Finally, we select a maximum of 300 words (1-2 ngrams) with the highest tf-idf scores. Again, this constraint is initially set due to budgetary considerations. However, our results show that on average *TFIDF* provides fewer than 100 potential keywords per m, a combination.

3.2.3 King

Given $D_{m,a}$ and $K^0_{m,a}$, we first identify the set of tweets that contain at least 1 keyword $k \in K^0_{m,a}$. We then downsample remaining tweets to an equal number to generate a balanced dataset. We train 6 separate classifiers[2] using the balanced dataset, and then label

[2] These classifiers are Random Forest, Naive Bayes, Linear Support Vector Machine, MultiLayer Perceptron, Logistic Regression, and Decision Tree. Note that we substituted the K-nearest Neighbor model in their original paper with MultiLayer

all the tweets in $D_{m,a}$. This process generates 6 probabilities for each tweet. We then use HDBSCAN [49] to cluster tweets using the distribution of classifier-assigned probabilities. Our approach differs from the original paper in various ways. First, we choose HDBSCAN over the CLARA method stated in [22] because the former doesn't require the user to set the number of clusters. Further, [22] clusters all tweets. This process is straightforward in the original study because the datasets used in [22] are significantly smaller and only contain tens of thousands of documents. Here, due to scaling issues, we filter out tweets without at least one positive label and also downsample the remaining tweets to a max of 2 million. Next, for each cluster that has at least 1K tweets, we use FP Growth Tree [50] which is faster than Apriori [51] to generate frequent itemsets (i.e., potential new keywords). Itemsets below a minimum support of 0.1 (less than 10% of documents contain this keyword) are not considered. Further, we set max length of 2 words (1-2 ngrams) per keyword (e.g, "gun"+"rights"). Finally, we rank the itemsets using Beta-Binomial likelihood function proposed by [22] and select the top 10 keywords (comparable to original paper). A flowchart of this complex pipeline is shown in Fig. 1 (Panel 2).

3.2.4 Linder

We again first generate a balanced dataset using $D_{m,a}$ and $K^0_{m,a}$ such that the number of movement-related tweets is equal to the number of non movement-related tweets. Following the approach proposed by [23], we first train a stochastic gradient descent (SGD) classifier using the balanced dataset, label all the tweets in $D_{m,a}$, and use Lasso regression to identify words that have the highest positive coefficient (more likely to occur in movement-related tweets). Finally, we select at most the top 100 (same threshold set by the original paper) potential keywords that are 1–2 ngrams.

3.2.5 Linder-BERT

The pipeline replicates *Linder* except we replace the SGD Classifier with DistilBERT [40]. Additionally, due to limitations in available processing resources, especially GPU units, we only label 20% of $D_{m,a}$ when running this pipeline.

3.3 Crowdsourcing

For each movement m at level activity level a, we aggregate all potential keywords generated through the five aforementioned pipelines and assign these words into batches of 20. We then randomly sample 2 gold standard keywords (that we know the label of) and inject them into each batch, resulting in 22 keyword batches. Next, we export each batch as a HIT task and assign it to three Amazon Mturk workers. To ensure the quality of work, we require that workers: (1) reside in the U.S. (2) have successfully completed at least 1,000 HITs; and (3) have an approval rate of at least 98%. We also provide workers with a codebook containing detailed examples of related and unrelated keywords. Additionally, for each potential keyword, workers are given 5 randomly sampled tweets containing the keyword. They are instructed to label a keyword as true positive if at least 3 out of the 5 tweets are movement-related. We choose this threshold given that prior work [1, 16] has highlighted recall issues when there are fewer keywords. This threshold allows us to capture a more comprehensive list of potential keywords. Outside the scope of our project,

Perceptron due to the former's slow performance. We also use GridSearchCV to optimize parameters, which is an extra step not taken in the original work.

researchers can easily find the nosiest keywords and include additional steps to improve tweet-level precision [52, 53]. Finally, a worker's labels are rejected if she fails to identify the gold standard keywords correctly, and the worker is thereafter barred from completing additional tasks.

The average Krippendorff alpha across all movements ranges from 0.3 to 0.4. While these numbers are below the thresholds used in traditional content analysis [48], they are common in crowdsourcing tasks (e.g., [54]). To ensure quality, we manually review all potential keywords labeled by at least 1 worker as relevant and keep the true positives. Our process of combining crowd work with expert labels is similar to [23].

4 Results

We execute our data collection pipelines for 3 iterations (or loops) using the process described in Fig. 1 (Panel 1). As shown in Fig. 3, the vast majority of pipelines experience considerably reduced tweet gain for the second and third loops. In fact, the fraction of additional tweets identified by each pipeline drops from an average of over 500% fraction gain, to an average of 5.3% and 6.2% gain on the third loop for movements in the base-activity level and high-activity level respectively. This suggests significant diminishing returns for a fourth loop. As such, we will use results from the 3rd loop for pipeline performance assessment.

In the following sections, we first summarize the overall performance of each pipeline focusing on the absolute volume of collected datasets. We then evaluate the pipelines using (i) tweet and user coverage; (ii) content similarity and topic diversity; (iii) temporal similarity and burst analysis; and (iv) cost analysis. The first three measurements incorporate the most popular types of analyses in social movement research (see Sect. 2.1). Finally, the remaining fourth analysis addresses real-world resources constraints researchers face.

4.1 Performance overview

First, we focus on each individual pipeline's performance by determining the number of relevant tweets they identify across different movements of varying activity levels. Results are summarized in Fig. 3. As shown, *Word2Vec* (red color) has the highest absolute number of tweets for most movements during average-activity level. Notably, we observe that no single pipeline consistently returns the highest number of tweets for movements at high-activity level. This could be due to movements having a high level of offline mobilization

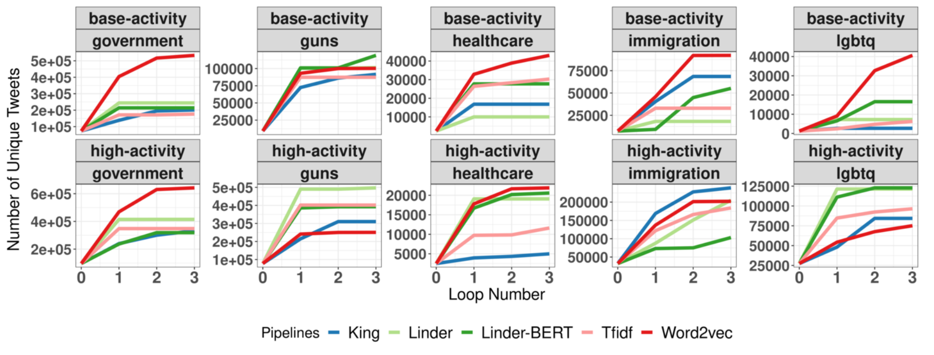


Figure 3 Movement-related Tweet Gain Per Loop

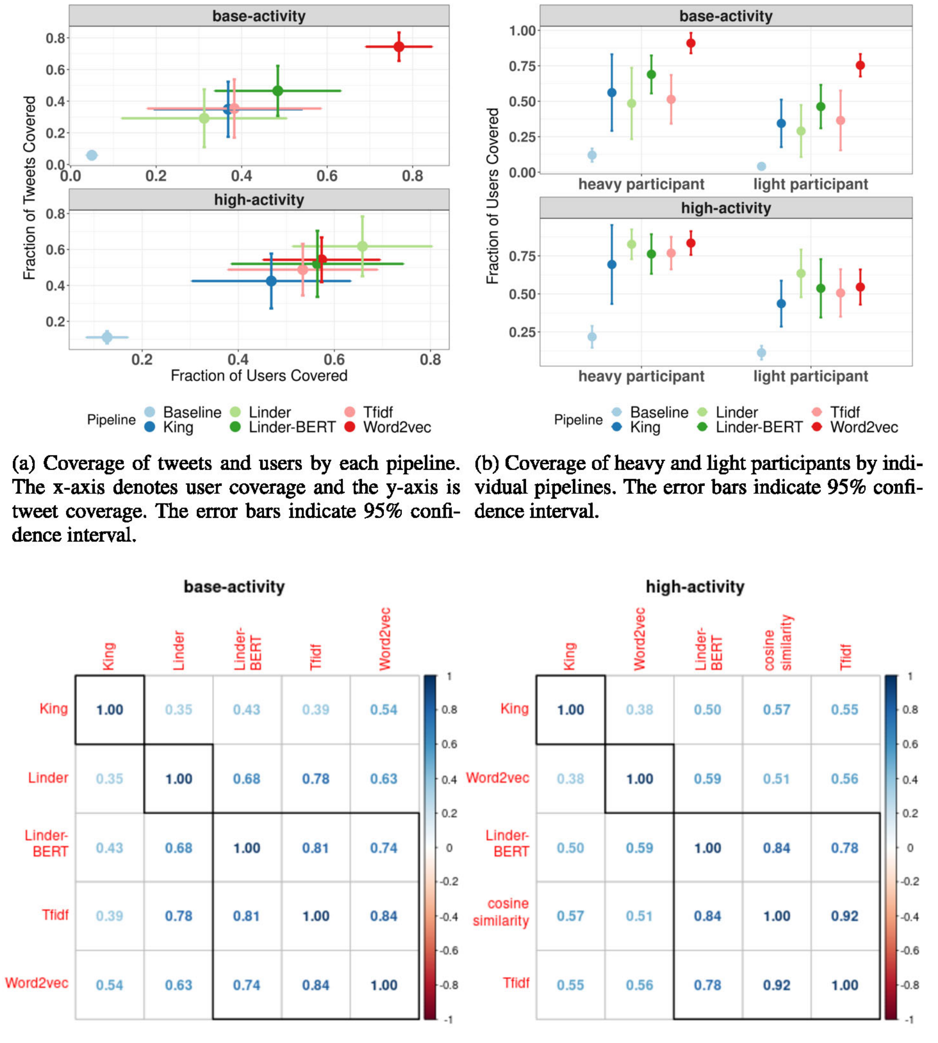
and public attention resulting in increased heterogeneity in online participants. This can lead to language use being too diverse for a single data collection approach to regularly outperform others. Nevertheless, we still see that both *Word2Vec* and *Linder* tend to have one of the highest numbers of tweets. Surprisingly, we do not observe an improvement in performance for *Linder-BERT* compared to *Linder*. While language models have demonstrated significant promise in past work for classification tasks [41], our results show that their promise is not universal. Notably, given the complexity and cost of this model, we had to downsample data to perform training. *Linder*, on the other hand, was run on the full data.

How do different pipelines achieve this aggregate performance? This is a function of two aspects: the number of movement relevant keywords they identify and the average number of tweets identified per keyword. *Word2Vec* has the highest number of possible keywords and one of the highest keyword precision scores across all pipelines. Specifically, it identified 1.1K possible keywords across 5 movements at average-activity level, and its average precision is 21.6% (i.e., 21 out of 100 possible keywords suggested by *Word2Vec* are true positives). In comparison, the same measure is less than 10% for *Linder-BERT, King*, and *Linder*. However, the average number of new tweets gained per keyword is only 570 for *Word2Vec*, significantly smaller than other pipelines (e.g. *Linder-BERT* gained an average of 4.3K tweets per keyword), suggesting that many of its keywords have low occurrence and/or high co-occurrence. Additionally, we also observe that the average keyword precision score is lowest for *healthcare* (12.3%) and highest for *government* (29.8%). Similarly, the average number of new tweets gained per keyword is lowest for healthcare (1.09K) and highest for *guns* (24.6K). Results are comparable when we examine performance for each activity level separately. Observations here suggest that pipelines have better performance for larger movements with more frequent offline activities. Note that the plots summarizing these analyses for all movements are omitted due to space constraints.

Finally, to ensure robustness, we use the total number of unique users instead of tweets identified by each pipeline to assess performance. The results are consistent.

4.2 Tweet and user coverage

Here, we assume that the union of the sets of tweets extracted by the pipelines is the best possible set of relevant tweets for a fixed movement and activity level. This is our ground truth set. We then measure each pipeline's coverage of this ground truth set using the following. Given a movement $m \in M$, and an activity level $a \in A$, let $S^{\star}_{m,a}$ be the union of all relevant tweets identified by all pipelines (i.e., the ground truth set). We compute the coverage of $S^{\star}_{m,a}$ by a given pipeline p as $\frac{|S^{p}_{m,a}|}{|S^{\star}_{m,a}|}$, where $S^{p}_{m,a}$ is the set of relevant tweets identified by p, and $|S^{p}_{m,a}|$ is the size of $S^{p}_{m,a}$. Similarly, we compute p's user coverage using the ground truth set of users identified. We calculate the coverage for all movements across both activity levels and plot the results in Fig. 4(a). Colors denote pipelines; the x-axis indicates user coverage, and the y-axis indicates tweet coverage. Focusing on base-activity level, we see that *Word2Vec* alone, on average, accounts for approximately 75% of tweets in $S^{\star}$, and significantly outperforms the next best pipeline *Linder-BERT*, which has an average coverage of 46%. For movements during high-activity level, *Linder* has the highest average coverage of 61%. Though, the difference between *Linder* and the 2nd highest coverage pipeline *Word2Vec* is insignificant.



(a) Coverage of tweets and users by each pipeline. The x-axis denotes user coverage and the y-axis is tweet coverage. The error bars indicate 95% confidence interval.

(b) Coverage of heavy and light participants by individual pipelines. The error bars indicate 95% confidence interval.

(c) Coverage similarity between pipelines measured using overlap coefficient of tweets.

Figure 4 Coverage Overview

For robustness checks, we also differentiate coverage of heavy and light participants. To elaborate, many social movement studies [8, 55] categorize Twitter users into activists (i.e., heavy participants) who meaningfully contribute to a movement, and light participants who are weakly involved in a movement. The values of both types of participants are then examined separately. Similar to [8], here we label a user as a light participant if they have only posted 1 or 2 movement-related tweets. We then recompute $\frac{|S^{p}_{m,a}|}{|S^{\star}_{m,a}|}$ separately for light and heavy participants. We observe comparable results. That is, during base-activity level, *Word2Vec* has, on average, over 75% coverage of both heavy and light participants, significantly outperforming other pipelines. However, performance is comparable across pipelines when movement activity-level is high. Additionally, we also see that all pipelines are significantly better at identifying heavy participants than light participants. For instance, when movements experience high offline activities, the pipelines on average are able to identify 75% of heavy participants, but only 50% of light participants (Fig. 4(b)).

Next, we assess the similarity of coverage S^p and S^q by a given pair of pipelines (p, q) by calculating their weighted overlap coefficient. The coefficient is defined as $\frac{\sum_{m\in M,a\in A} |S^p_{m,a} \cap S^q_{m,a}|}{\min(\sum_{m\in M,a\in A} |S^p_{m,a}|, \sum_{m\in M,a\in A} |S^q_{m,a}|)}$ across all movements M and activity levels A. Next, we arrange pipelines into hierarchical clusters using *hclust* function in R stats library ("euclidean" distance and "complete" linkage). Results for coverage similarity measures are summarized in Fig. 4(c). As shown, the pipelines *Linder, TFIDF,* and *Linder-BERT* tend to have high similarities with each other. Additionally, *Word2Vec* is more similar to *TFIDF, Linder, Linder-BERT* in base-activity level than high-activity level. This is unsurprising given that *Word2Vec* is able to capture 75% of $S^\star$ for movements in base-activity level. Further, we also see that the pipeline *King* significantly differs from the remaining pipelines for both activity levels. A likely explanation is that the pipelines *TFIDF, Linder, Linder-BERT* and *Word2Vec* only identify potential keywords that are (i) a single word, or (ii) consecutive n-grams, whereas *King*'s use of frequent itemset mining allows it to identify nonconsecutive keywords. (e.g., the keyword "pride+twibbon" from tweets such as "please help support pride, add a #twibbon now!").

Finally, we calculate the fraction of tweets covered by at most x pipelines, where $x = \{1, 2, 3, 4\}$. We select the top pipelines using 2 ranking strategies: (i) optimal and (ii) random. For *optimal*, we exhaustively calculate the coverage using all combinations of pipelines, and select the best combination. We see that even choosing two pipelines at random increases coverage significantly compared to randomly choosing a single pipeline (from an average of 52% to an average of 71% for high-activity level). Though this value is less than that of the optimal selection (90% with two pipelines).

Implications Results here suggest that *Word2Vec* has the highest coverage across all pipelines when movements are in base-activity levels. A likely explanation is that when a movement is experiencing a normal offline activity level, tweets focused on the movement are far fewer. As such, pipelines that use classifiers for keyword expansion (*Linder, Linder-BERT, King*) may suffer performance issues due to a shortage of training data. In comparison, *Word2Vec* can leverage the entire corpus to generate vectors. Additionally, we also demonstrate that using a single data mining method is insufficient when social movements have high offline activities. Indeed, in the best case, a single pipeline results in only three-fifths of all tweets being identified. Furthermore, pipelines are also significantly worse at mining light participant data. As such, studies that rely on a single mining pipeline likely underestimate the involvement of light participants. Instead, researchers should consider using two or more pipelines. Our pipeline coverage similarity analysis provided indications for which subset of pipelines should be chosen. For example, researchers can use *Word2Vec, King,* and *TFIDF* (which is similar to *Linder* and *Linder-BERT*).

4.3 Content similarity and topic diversity

So far we look at coverage defined as the number of unique tweets (unique tweet ids) or users (i.e. unique user ids) identified as relevant by each pipeline. However, tweets can be quite similar to each other in content. In this section, we go beyond merely counting tweet ids and focus on the actual text of each tweet. We conduct two analyses: (i) we measure content similarity across the pipelines; and (ii) we also explore topic coverage by each pipeline. A notable concern here is that data sizes vary significantly across pipelines. As such, content analyses where tweets from all pipelines are simply aggregated to define the complete corpus can lead to large pipelines dominating the detected topics/language. In short, in such an analysis, pipeline text size is a strong confound. To remedy this matter, we conduct our analysis using proportional samples from each pipeline. That is, we randomly sample the same number of tweets from each pipeline (the number of tweets from the smallest pipeline) for each movement m and activity-level a. A pipeline p's sampled tweets are denoted as $\widetilde{S}^{p}_{m,a}$, and the union of the tweets across all pipelines is $\widetilde{S}^{\star}_{m,a}$.

We first evaluate the similarity of content identified by each pipeline using the Jensen Shannon distance [56] To elaborate, let $W^{p}_{m,a}$ be all words that occurred in $\widetilde{S}^{p}_{m,a}$. We denote $F(W^{p}_{m,a})$ as the probability distribution of $W^{p}_{m,a}$ such that $\sum_{w\in W^{p}_{m,a}} f(w) = 1$. We can then calculate the distance between a pair of pipelines (p,q) using $JSD(F(W^{p}_{m,a})||F(W^{q}_{m,a}))$ where JSD is the Jensen Shannon distance equation. Finally, we compute the average distance between pairs of pipelines across all movements and activity levels. And, we again apply hierarchical clustering using *hclust* to identify pipelines with the highest similarities. We observe results comparable to previous coverage-based analysis. That is, *TFIDF, Linder*, and *Linder-BERT* have the highest similarities with each other. Content by *Word2Vec* and *King*, however, are dissimilar with the other 3 pipelines and with each other.

Next, we focus on topic coverage analysis. To do so, we fist identify topics in $\widetilde{S}^{\star}_{m,a}$. We then determine whether tweets in $\widetilde{S}^{p}_{m,a}$ are present in significant portions in each topic. In other words, we evaluate whether tweets collected through p are representative of all identifiable topics for a movement. Here, we sample 2 specific movements, *lgbtq* and *government*, to examine the topic coverage by pipelines. We pick the two movements due to the significant difference in their offline event frequency (*government* has a total 2.1K offline events compared to *lgbtq*'s 228) and size of crowd events (the average size of *government* is 2.05K compare to *lgbtq*'s 20.0K).

Here, we describe our topic generation and evaluation process. First, given movement m, activity-level a, and all tweets $\widetilde{S}^{\star}_{m,a}$, we use SIF embedding [57] to convert each tweet in $S^{\star}_{m,a}$ to its corresponding sentence vector. Next, we use KMEANS [58] to cluster these vectors. To identify the best k value, we set $k = \{6, 8, 10, \dots, 50\}$ and run KMEANS for each k. We then choose the best k using the elbow method [59]. To evaluate this technique, we sample 10 tweets from each cluster and manually review these tweets to identify coherent topics. Our analysis shows that this approach leads to coherent clusters. For instance, for *lgbtq* during *high-activity* (i.e., $S_{lgbtq,high}$), we observe that 16 (or 80%) of the clusters have a coherent topic, with an average sample precision of 87.5%, and accounting for 53% of all tweets. Clustering results for coherent topics are summarized on Table 2. For each topic, we include the fraction of total tweets in $S_{lgbtq,high}$ accounted for by the topic (*tweets frac*), precision score for the manually reviewed sample tweets (*precision*), and a randomly sampled tweet from the true positive samples (*sample tweet*).

Thereafter, we describe the topic coverage of pipeline p for movement m and activity a as $G(C,p)_{m,a} = \{g(c_{0,m,a},p), g(c_{1,m,a},p), g(c_{2,m,a},p), \dots\}$, where $g(c_{0,m,a},p)$ is the fraction of tweets in cluster c_0 that's identified by p for movement m and activity level a. As an hypothetical example, $G(C,p)_{m,a} = \{0.4, 0.25, 0.1\}$, indicates that the pipeline p is able to identify 40% of tweets in the first cluster, and 25% and 10% of the tweets for the 2nd and 3rd clusters (note that $\sum_i g(c_{i,m,a},p) \neq 1$). Finally, we derive the fraction of topics that p is able to identify at least $g(c,p) \geq x$ percentage of tweets. We perform this for movement and activity pairs $m \in \{lgbtq, government\}$ and $a \in \{high\ activity, base\ activity\}$.

Table 2 Coherent Clusters for the LGBTQ Movement in High-activity Level. To protect privacy, we change all user handles for political elites to "@politician", individuals to "@username", NGOs to "@NGO", and news organizations to "@newsmedia"

Cluster description	Tweets frac	Preci-sion	Sample tweet
arts and entertainment	0.099	0.7	running a 20% off sale till the end of this month in honor of pride! all our rainbow themed items are dis
specific location prides	0.081	0.8	columbus will greet pence with a big lgbtq dance party when he arrives during pride via /r/atheism https://t.co/xm64ndvse0
anti-lgbtq related	0.074	0.8	@username i can attest to this assertion–coming from a hometown that had an anti-lgbtq movement in which bigots
political action	0.066	0.9	retweeted @newsmedia (@newsmedia): the world health organization announces it will stop classifying transgender people as menta
genderqueeredness	0.055	1.0	@username yeah i feel like, theres a whole dichotomy in how my queerness presents to white/cishets, vs native/lgbt folk
@notable politician	0.050	1.0	@politician what are you gonna tell the truth and tell the lgbt community u lied to them in order to get their votes???
pride parade	0.036	0.9	happy #pride! leo luckett rides a rainbow-colored dildo in honor of #gaypridemonth!
specific celebrities	0.024	0.8	@username i love you so much. thank you for being so open and honest and creating a safe space. happy #pride angel.
business and sports	0.010	0.7	the fa is supporting lgbt+ pride football flags at russia world cup – https://t.co/zm95begmhw
twibbon campaigns	0.008	1.0	please help support support nb nonbinary pride, add a #twibbon now! https://t.co/z5ymrjbqmy
lgbtq canada	0.007	1.0	we support and appreciate the lgbtq+ community and are choosing to donate to @NGO and @NGO #happypridemonth
positive initiatives	0.006	0.8	#whydontyou hat's off to the bold transgender modeling organisation for this brilliant initiative. it just made me
lgbtq books	0.005	0.9	28 fabulous works of queer historical fiction for pride month https://t.co/tpxtamem6b
transgender	0.004	0.9	@username please help me spread the word about my #lgbtq+ #horror short #conversiontherapist starring transactr
tv shows	0.004	1.0	#saveshadowhunters because it is a show with positive lgbtq relationships and has one of the most diverse casts on tv @username
ask for follow	0.002	1.0	if you love this pic follow me i upload daily. #crossdresser #transvestite #sissy #transgender #cd #sissycaptions

Results from the above process are shown in Fig. 5. The y-axis denotes the fraction of topics that a pipeline (denoted by color) can identify at least x (x-axis) fraction of tweets for. As an example, *Word2Vec* is able to identify at least 25% of the tweets for 94% of topics for the *lgbtq* movement in high-activity level – the highest among all pipelines. We observe comparable results for *government*, *high-activity*), and (*government*, *base-activity*). Results differ for (*lgbtq*, *base-activity*), likely due to small datasets and a very small number of clusters (4 of the 6 clusters have a coherent topic).

Next, we again assess whether having multiple pipelines increases the overall topic coverage. We observe that using multiple pipelines rather than a single pipeline significantly increases the topic coverage. This is true even when pipelines are randomly selected. This result is consistent with our overall coverage findings presented in Sect. 4.2.

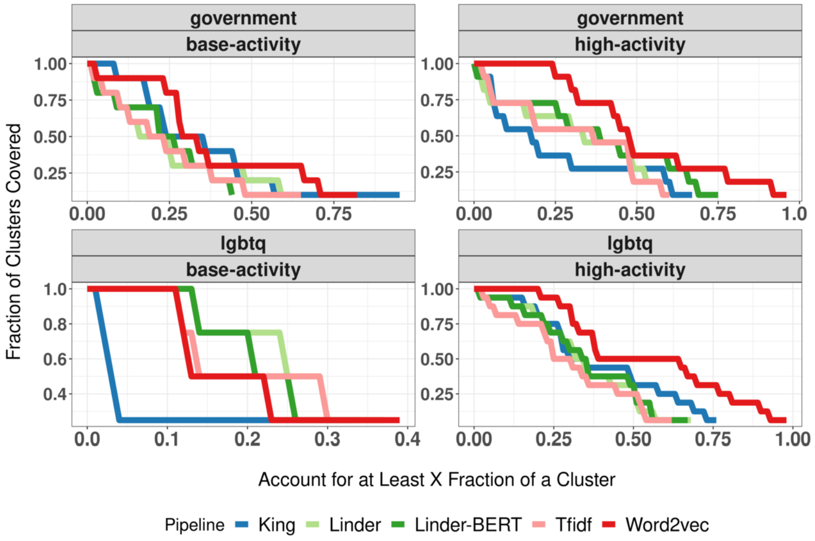


Figure 5 Topic Coverage. The *y*-axis denotes the fraction of topics that has at least X fraction of coverage by a pipeline (separated by color). As a example, we see that for the *lgbtq* movement in high-activity level, *Word2vec* is able to identify at least 25% of the tweets for 94% of all topics

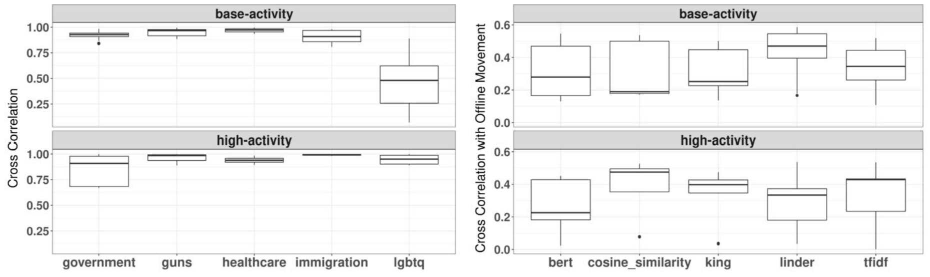
Finally, we resample tweets to generate new $\widetilde{S}^{p}_{m,a}$ and re-run the prior analyses. Our results are consistent.

Implications *Word2Vec* has the highest topic coverage compared to all the other pipelines. Similar to results from Sect. 4.2, we observe that (i) using multiple pipelines significantly increases topic coverage and (ii) content generated by *Linder, Linder-BERT* and *TFIDF* are similar to each other, but are dissimilar to content by *King* and *Word2Vec*.

4.4 Temporal similarity and burst analysis

In this section, we first examine whether the time-series generated using data collected by different pipelines are similar to each other. Given a movement m, activity-level a, and pipeline p, let $s^{p,t}_{m,a}$ be the number of tweets with a creation date of t identified by p and $s^{p,t}_{m,a} = \{s^{p,1}_{m,a}, s^{p,2}_{m,a}, \ldots, s^{p,n}_{m,a}\}$ be the time series data where n is the number of days in the given month (as noted in Sect. 3.1, we have 1 month of data from each movement and activity-level combination.). We standardize these time-series and then calculate the cross-correlation between pairs of pipelines (p, q). Results are summarized in Fig. 6(a). As shown, the average correlation between pipelines is very high: 0.84 and 0.94 for movements in base and high-activity levels, respectively. Further, we use the sample cross-correlation function [60] to determine the value of the lag with the highest correlation coefficient between the two series. Our results show that the median lag, which is defined as the number of days a time-series lags behind another that maximizes the cross-correlation between the two time series, is 0 days.

Next, we compare time-series of online movements to their offline counterparts. As shown in Fig. 6(b), the average coefficient is only 0.35 and 0.29 for movements in base and high-activity levels respectively. We repeat this assessment using the number of people who participated in offline protests and find similar results. Observations here suggest that temporal patterns of online and offline movements are mismatched. This result is



(a) Online Time-series

(b) Offline Time-series

Figure 6 Time-series cross-correlation Results

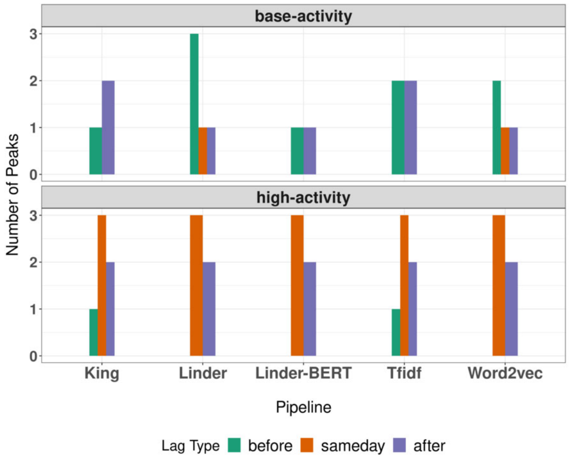


Figure 7 The *x*-axis indicates the pipelines, and the *y*-axis is the number of matched bursts with lag type (e.g., green indicates that an online burst took place before an offline burst)

surprising given that numerous prior work has used online movement activities to predict offline activities, or vice versa [61]. One possible explanation is that many prior studies are only focused on predicting large, notable offline events (i.e., when a movement has mobilized a significant part of the population and gained much public attention) but not multitudes of small offline events with a handful of participants.

Given this consideration, we next explore the relationship between a movement's online and offline components by focusing on bursts – days when a large number of sizable events take place online or offline. Here, we first use IQR [62] (threshold of 1.5*IQR) to identify bursts in each of the time-series generated. We then match offline bursts to online bursts when possible. Note, an online burst and an offline burst are considered matched when the time difference between them is ±7 days (results are consistent when the threshold is set to ±3 days). If there are multiple matches, the one with minimum time difference is selected. We then determine whether movements' online bursts lag behind (or happen before) the matching offline bursts. Further, we also evaluate whether results are similar for different pipelines. As shown in Fig. 7, in base-activity level, movements' online bursts take place prior to offline bursts, whereas online and offline bursts happen on the same day during high-activity level. In other words, when a social movement is in maintenance mode, a significant increase in its online activities often takes place prior to increased offline activities. However, when a movement is already at its peak activity period, both online and offline bursts take place simultaneously.

Moreover, we observe that online time-series generated through a single pipeline is only able to recall at most 55.5%, and 46% of all offline bursts in base and high-activity levels. Specifically, *Linder* recalls 5 out of 9 total offline bursts for base-activity; *King* and *TFIDF* both recall 6 out of 13 offline bursts in high-activity level. Further, we again assess whether combination of pipelines as opposed to a single pipeline would increase the overall recall. We observe little improvement. In fact, combining all pipelines still results in 66.7% recall for offline bursts taking place during movements' base-activity levels and 54% for those taking place during high-activity levels. This perhaps suggests that there are aspects of offline protests that are not reflected in overall online behavior – a shortcoming not even the best data mining/collection mechanism can solve. Finally, we repeat our analysis using threshold of 3*IQR and observe similar results.

Implication Overall, results in this section show that temporal patterns between a movement's online and offline activities differ. In fact, approximately half of the time, bursts in a movement's offline activities don't have corresponding bursts in the movement's online activities. One possible hypothesis is that the volume of movement-related tweets is, by itself, a poor indicator/predictor for the number and size of offline activities. Perhaps, focusing on mobilization-related tweets (e.g., "rally", "sit-in", "join me") could potentially lead to better prediction. Focusing on the performance of pipelines with respect to temporal analysis, results suggest that the pipelines are comparable with highly similar time-series.

4.5 Cost analysis

We also review each pipeline's resource requirements and consumption (Table 3). We see that *TFIDF* and *Word2vec* take very little time to complete. In comparison, *Linder-BERT* and *King* have the longest run time. Further, *TFIDF* and *Word2vec* can run on much smaller RAM, whereas *Linder* requires compute nodes with 180GB RAM. Additionally, due to *Word2vec* and *King* generating the largest number of possible keywords (see Sect. 4.2), they also have the highest manual labor cost. Note that MTurk workers are paid $0.15 for each keyword identification task. We choose this pay rate aiming for an hourly rate of $15, based on our internal tests that indicated that individuals take 30 to 40 seconds on average to complete each task.

Implications Cost is a significant consideration for numerous reasons. First, not all researchers have easy access to large memory machines or GPUs. Second, the time it takes

Table 3 Pipeline Resource Consumption. Note that the smallest node in our cluster has 120GB RAM. However, testing shows that *TFIDF* and *Word2vec* can both run on much smaller RAM (16GB to 32GB)

Pipeline	Ave. Hours Per Loop (Single Thread)	Total Mturk Cost	PC Requirement	Notes
Linder-BERT	118.0	$28.0	Tesla-K80 GPU	Partial data processing (20% tweets)
Word2vec	1.0	$300.0	120GB RAM	–
King	122.0	$102.0	120GB RAM	Partial data clustering (max 2 million tweets)
Linder	24.0	$31.0	180GB RAM	–
TFIDF	0.0	$29.0	120GB RAM	–

to process data can make testing, piloting, and timely data collection difficult. Here, approaches such as *Word2vec* are more cost-effective, in addition to providing better coverage.

5 Conclusion

In our work, we have provided a systematic evaluation of five existing keyword expansion pipelines. We have assessed their performance for five contemporary movements in two distinct activity levels (baseline and peak mobilization period). Our results showed that pipeline performance significantly differs across activity levels. Word embedding methods readily outperformed other approaches when movements were experiencing a normal activity level. However, when movements' mobilization was high, no single pipeline significantly outperformed others. In fact, performance was low for all pipelines: the best pipeline on average only identified little more than half of all movement-related tweets collected through all pipelines. More importantly, we also showed that using multiple pipelines, even selected at random, can significantly increase tweets' coverage.

We note the following caveats in our work. First, our work didn't separate a social movement into its progressive and reactionary (i.e., countermovement) components. Prior work has shown that popular keywords such as hashtags are often hijacked or appropriated [7, 63, 64]. Similarly, we assigned offline events to social movements using issue areas rather than specific issues. For example, the *immigration* movement consists of multiple interconnected issues or grievances including family separation at the border, termination of DACA, the building of border wall. Future work may study these unique instantiations of the *immigration* movement separately. Furthermore, we also highlight that our paper labels a candidate keyword as a true positive keyword if a majority of sampled tweets are related to a movement. Hence, the resulting tweets contain both true positive tweets and false positives. Keyword based data mining approaches are inherently noisy [22]. Outside of the scope of our paper, improving tweet-level precision is an important endeavor. Strategies explored in past work provide potential solutions here. Similar to [52], researchers can obtain tweet-level labels from crowdsourced platforms, and then build a classifier to filter out irrelevant tweets. For each noisy keyword, researchers can also identify additional 2nd-level keywords that, when combined with the noisy keyword, contribute to higher tweet-level precision [53]. Next, due to resource constraints, we were unable to process all available data when running the BERT variation of the method proposed by Linder [23]. Similarly, when running the method proposed by [22], we downsampled when clustering tweets due to scaling issues. Both methods' performance may increase with a full set of available data and without downsampling. However, computational resources used in this paper are already substantial (e.g., we ran 16 GPUs in parallel for the BERT variation). The resources required by these methods would likely render them cost-prohibitive to many researchers. In comparison, word embedding models are fast while requiring significantly less computing power and memory. As another caveat, we note that researchers will need to already have access to a large static dataset where relevant subsets of movement-related data can be identified. That is, these methods are not built for researchers who, for instance, need to continuously search for relevant tweets through Twitter's search API. However, static archival datasets should be readily available through Twitter's open access policy for academic research. Furthermore, the techniques we examined can still be used to identify a set of keywords using a smaller sample (e.g., 1% data access more broadly available even beyond academia) and later used to extract future content through the

Search API. Next, we implemented existing methods as closely as we could to their descriptions in the original papers. Additional modifications to the pipelines could improve their performance (e.g., incorporating parts of a pipeline into another). Finally, due to the sheer volume of Twitter Decahose data, we only focused on monthly tweet data instead of yearly. And, we only included original tweets (i.e., no retweets).

Despite these limitations, we provided a robust evaluation of the performance of the most representative and commonly used keyword expansion methods. Our work has several noteworthy implications. First, we identified a simple and affordable keyword expansion technique, the word embedding pipeline, that significantly outperforms more complex and expansive pipelines. Second, we also demonstrated that pipeline performance varies significantly across different movement stages (baseline or growth). As such, researchers who are interested in longitudinal analysis of social movements should utilize a more robust data collection approach. Third, pipelines are significantly better at detecting heavy participants who frequently post movement-related tweets than light participants regardless of movement stage. This result is consistent with [16], which showed that keyword based data mining approaches are better at identifying users with stronger political identities. As such, we urge studies focused on politically moderate proponents or opponents of a social movement to assess and mediate potentially data representativeness issues. Finally, to facilitate future work within this space, we make available the initial sets of seed keywords, the final sets of keywords we obtained for each movement, and code for the pipelines.[3]

Social movements are vital to monumental, progressive change, leading to ever increasing academic interest in the subject [65]. In recent years, a wide range of high-profile social movements arose amid massive social, political, and economic crises in the United States. Social media data makes it possible for us to examine these movements at a scale and granularity that was impossible in the pre-Internet era. Yet, this scale also makes it more challenging to collect movement-related content. Our paper undertook this challenge. Insights gained through our work can be used by social movement scholars to identify better quality Twitter data for their research. Finally, social movement scholars also have been using various approaches to mine data from other online platforms [66, 67], and are likely facing similar issues due to the lack of benchmarking. The evaluation process proposed in our study can be adopted and revised for benchmark analysis on other platforms. This process is particularly applicable for platforms that provide robust text-based streaming APIs (e.g., Weibo) or public data (e.g., Reddit's pushshift dataset [68]). Future work can expand benchmark studies to these matching platforms.

Funding

This research was supported by the National Science Foundation (Grant IIS-1815875).

Abbreviations

TFIDF, pipeline that uses the tf-idf weighting scheme; *Word2vec*, word embedding pipeline; *King*, pipeline proposed by King et al. [22]; *Linder*, pipeline suggested by Linder [23]; *Linder-BERT*, modified version of *Linder* that relies on deep learning.

Availability of data and materials

Twitter data is available via MIDAS Twitter Decahose https://midas.umich.edu/twitter-decahose-data/. Additionally, movement offline data is obtainable from https://github.com/nonviolent-action-lab. Finally, we will provide the initial and final sets of keywords for each movement, and the code repository for all pipelines upon the publication of this manuscript.

[3] https://github.com/lbozarth/keywordexpansion

Declarations

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

Authors' contributions
Both authors conceived and designed the study together. The first author collected and analyzed the data. Both authors then wrote the manuscript together. All authors read and approved the final manuscript.

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

Received: 11 November 2021 Accepted: 1 May 2022 Published online: 21 May 2022

References

1. González-Bailón S, Wang N, Rivero A, Borge-Holthoefer J, Moreno Y (2014) Assessing the bias in samples of large online networks. Soc Netw 38:16–27. https://doi.org/10.1016/j.socnet.2014.01.004
2. Theocharis Y, Lowe W, Van Deth JW, García-Albacete G (2015) Using Twitter to mobilize protest action: online mobilization patterns and action repertoires in the occupy wall street, indignados, and aganaktismenoi movements. Inf Commun Soc 18(2):202–220
3. Maginot KB, Chaudhuri S (2015) No shaming this slut. In: Research in social movements, conflicts and change. Emerald Group Pub., Bingley
4. Tillery AB (2019) What kind of movement is black lives matter? The view from Twitter. J Race Ethn Polit 4(2):297–323
5. Conover MD, Ferrara E, Menczer F, Flammini A (2013) The digital evolution of occupy wall street. PLoS ONE 8(5):64679
6. Segerberg A, Bennett WL (2011) Social media and the organization of collective action: using Twitter to explore the ecologies of two climate change protests. Commun Rev 14(3):197–215
7. Gallagher RJ, Reagan AJ, Danforth CM, Dodds PS (2018) Divergent discourse between protests and counter-protests: #BlackLivesMatter and #AllLivesMatter. PLoS ONE 13(4):0195644. https://doi.org/10.1371/journal.pone.0195644
8. Bozarth L, Budak C (2020) Beyond the eye-catchers: a large-scale study of social movement organizations' involvement in online protests. New Media Soc 1461444820943854
9. Olteanu A, Weber I, Gatica-Perez D (2016) Characterizing the demographics behind the# blacklivesmatter movement. In: 2016 AAAI spring symposium series
10. Twitter API for Academic Research|Products|Twitter Developer Platform Twitter. https://developer.twitter.com/en/products/twitter-api/academic-research
11. Jürgens P, Jungherr A (2016) A tutorial for using twitter data in the social sciences: data collection, preparation, and analysis. Preparation, and Analysis
12. McCormick TH, Lee H, Cesare N, Shojaie A, Spiro ES (2017) Using Twitter for demographic and social science research: tools for data collection and processing. Sociol Methods Res 46(3):390–421
13. Earl J, McKee Hurwitz H, Mejia Mesinas A, Tolan M, Arlotti A (2013) This protest will be tweeted: Twitter and protest policing during the pittsburgh g20. Inf Commun Soc 16(4):459–478
14. Tufekci Z (2014) Big questions for social media big data: representativeness, validity and other methodological pitfalls. In: Proceedings of the international AAAI conference on web and social media
15. Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? Comparing data from Twitter's streaming api with Twitter's firehose. In: ICWSM
16. Barberá P, Rivero G (2015) Understanding the political representativeness of Twitter users. Soc Sci Comput Rev 33(6):712–729
17. Barberá P, Steinert-Threlkeld ZC (2020) How to use social media data for political science research. In: The SAGE handbook of research methods in political science and international relations. Sage, London, pp 404–423
18. Ramos J et al (2003) Using TF-IDF to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning, vol 242, pp 29–48. Citeseer
19. Nabeel Asim M, Wasim M, Usman Ghani Khan M, Mahmood W (2018) Improved biomedical term selection in pseudo relevance feedback. Database 2018:bay056. https://doi.org/10.1093/database/bay056
20. Wartena C, Brussee R, Slakhorst W (2010) Keyword extraction using word co-occurrence. In: 2010 workshops on database and expert systems applications. IEEE Comput. Soc., Los Alamitos, pp 54–58
21. Diaz F, Mitra B, Craswell N (2016) Query expansion with locally-trained word embeddings. arXiv preprint. 1605.07891
22. King G, Lam P, Roberts ME (2017) Computer-assisted keyword and document set discovery from unstructured text. Am J Polit Sci 61(4):971–988
23. Linder FJ (2018) Methods for the study of public opinion using social media data with applications to the german refugee crisis
24. Nonviolent-action-lab: crowd counting consortium crowd data. https://github.com/nonviolent-action-lab/crowd-counting-consortium
25. Fisher DR, Andrews KT, Caren N, Chenoweth E, Heaney MT, Leung T, Perkins LN, Pressman J (2019) The science of contemporary street protest: new efforts in the United States. Sci Adv 5(10):5461
26. Dreier P (2020) The decade in 11 movements. https://prospect.org/civil-rights/the-decade-in-11-movements/
27. Xiong Y, Cho M, Boatwright B (2019) Hashtag activism and message frames among social movement organizations: semantic network analysis and thematic analysis of Twitter during the# metoo movement. Public Relat Rev 45(1):10–23
28. Mueller A, Wood-Doughty Z, Amir S, Dredze M, Nobles AL (2021) Demographic representation and collective storytelling in the me too Twitter hashtag activism movement. In: Proceedings of the ACM on Human-Computer Interaction 5(CSCW1), pp 1–28

29. De Choudhury M, Jhaver S, Sugar B, Weber I (2016) Social media participation in an activist movement for racial equality. In: ICWSM, pp 92–101
30. Costa JM, Rotabi R, Murnane EL, Choudhury T (2015) It is not only about grievances-emotional dynamics in social media during the Brazilian protests. In: Proceedings of the ninth international AAAI conference on web and social media, pp 594–597
31. Varol O, Ferrara E, Ogan CL, Menczer F, Flammini A (2014) Evolution of online user behavior during a social upheaval. In: Proceedings of the 2014 ACM conference on web science. ACM, New York, pp 81–90
32. Neto JL, Santos AD, Kaestner CA, Alexandre N, Santos D et al (2000) Document clustering and text summarization
33. Abilhoa WD, De Castro LN (2014) A keyword extraction method from Twitter messages represented as graphs. Appl Math Comput 240:308–325
34. Ahmed S, Chy AN, Ullah MZ (2020) Exploiting various word embedding models for query expansion in microblog. In: 2020 IEEE 8th R10 humanitarian technology conference R10-HTC, pp 1–6. https://doi.org/10.1109/R10-HTC49770.2020.9357016. ISSN 2572-7621
35. Zhang K, Xu H, Tang J, Li J (2006) Keyword extraction using support vector machine. In: International conference on web-age information management. Springer, Berlin, pp 85–96
36. Li C, Wang Y, Resnick P, Mei Q (2014) Req-rec: high recall retrieval with query pooling and interactive classification. In: Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval, pp 163–172
37. Kenter T, De Rijke M (2015) Short text similarity with word embeddings. In: Proceedings of the 24th ACM international on conference on information and knowledge management, pp 1411–1420
38. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
39. Joulin A, Grave E, Bojanowski P, Mikolov T (2016) Bag of tricks for efficient text classification. arXiv preprint. 1607.01759
40. Sanh V, Debut L, Chaumond J, Wolf T (2019) Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint. 1910.01108
41. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding
42. Pressman J, Choi-Fitzpatrick A (2020) COVID19 and protest repertoires in the United States: an initial description of limited change. Soc Mov Stud 20(6):766–773
43. Andrews KT, Caren N, Browne A (2018) Protesting trump. Mobilization Int Q 23(4):393–400
44. Beyerlein K, Ryan P, Abu-Hazeem A, Pauley A (2018) The 2017 women's March: a national study of solidarity events. Mobilization Int Q 23(4):425–449
45. Larson J, Soule S (2009) Sector-level dynamics and collective action in the United States, 1965–1975. Mobilization Int Q 14(3):293–314
46. Conover M, Ratkiewicz J, Francisco M, Gonçalves B, Menczer F, Flammini A (2011) Political polarization on Twitter. In: Proceedings of the international AAAI conference on web and social media
47. Shapiro MA, Hemphill L (2017) Politicians and the policy agenda: does use of Twitter by the us congress direct New York times content? Policy Internet 9(1):109–132
48. Krippendorff K (2018) Content analysis: an introduction to its methodology. Sage, Thousand Oaks
49. McInnes L, Healy J, Astels S (2017) hdbscan: hierarchical density based clustering. J Open Sour Softw 2(11):205
50. Borgelt C (2005) An implementation of the fp-growth algorithm. In: Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations, pp 1–5
51. Bodon F (2003) A fast apriori implementation. In: FIMI, vol 3, pp 63
52. Hassan N, Poudel A, Hale J, Hubacek C, Huq KT, Santu SKK, Ahmed SI (2020) Towards automated sexual violence report tracking. In: Proceedings of the international AAAI conference on web and social media, vol 14, pp 250–259
53. Charquero-Ballester M, Walter JG, Nissen IA, Bechmann A (2021) Different types of Covid-19 misinformation have different emotional valence on Twitter. Big Data Soc 8(2):20539517211041279
54. Lind F, Gruber M, Boomgaarden HG (2017) Content analysis by the crowd: assessing the usability of crowdsourcing for coding latent constructs. Commun Methods Meas 11(3):191–209
55. Simpson E (2018) Integrated & alone: the use of hashtags in Twitter social activism. In: Companion of the 2018 ACM conference on computer supported cooperative work and social computing, pp 237–240
56. Lin J (1991) Divergence measures based on the Shannon entropy. IEEE Trans Inf Theory 37(1):145–151
57. Arora S, Liang Y, Ma T (2016) A simple but tough-to-beat baseline for sentence embeddings
58. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
59. Yuan C, Yang H (2019) Research on k-value selection method of k-means clustering algorithm. Multidiscipl Sci J 2(2):226–235
60. Helleseth T (1976) Some results about the cross-correlation function between two maximal linear sequences. Discrete Math 16(3):209–232
61. González-Bailón S, Borge-Holthoefer J, Rivero A, Moreno Y (2011) The dynamics of protest recruitment through an online network. Sci Rep 1(1):1–7
62. Sun Y, Genton MG (2011) Functional boxplots. J Comput Graph Stat 20(2):316–334
63. Jackson SJ, Foucault Welles B (2015) Hijacking# mynypd: social media dissent and networked counterpublics. J Commun 65(6):932–952
64. Knüpfer C, Hoffmann M, Voskresenskii V (2020) Hijacking MeToo: transnational dynamics and networked frame contestation on the far right in the case of the '120 decibels' campaign. Inf Commun Soc, 1–19
65. Buechler SM (1995) New social movement theories. Sociol Q 36(3):441–464
66. Rho EHR, Mark G, Mazmanian M (2018) Fostering civil discourse online: linguistic behavior in comments of# metoo articles across political perspectives. In: Proceedings of the ACM on human-computer interaction 2(CSCW), pp 1–28
67. Vizcaíno-Verdú A, Aguaded I (2022) # thisismechallenge and music for empowerment of marginalized groups on tiktok. Media Commun 10(1)
68. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J (2020) The pushshift reddit dataset. In: Proceedings of the international AAAI conference on web and social media, vol 14, pp 830–839