Sampling Social Media: Supporting Information Retrieval from Microblog Data Resellers with Text, Network, and Spatial Analysis

Cody Buntain
University of Maryland
College Park, MD 20742
cbuntain@cs.umd.edu

Erin McGrath
University of Maryland
College Park, MD 20742
ecmcgrat@umd.edu

Brandon Behlendorf
University at Albany, SUNY
Albany, NY 12222
bbehlendorf@albany.edu

Abstract

This paper presents a computationally assisted method for scaling researcher expertise to large, online social media datasets in which access is constrained and costly. Developed collaboratively between social and computer science researchers, this method is designed to be flexible, scalable, cost-effective, and to reduce bias in data collection. Online response to six case studies covering elections and election-related violence in Sub-Saharan African countries are explored using Twitter, a popular online microblogging platform. Results show: 1) automated query expansion can mitigate researcher bias, 2) machine learning models combining textual, social, temporal, and geographic features in social media data perform well in filtering data unrelated to the target event, and 3) these results are achievable while minimizing fee-based queries by bootstrapping with readily-available Twitter samples.

1. Introduction

Online social media platforms are changing how researchers study citizen engagement. Through social media, political discussions have moved from private affairs to public debates, shedding light on contentious issues. Broad availability of these platforms have also provided a forum for the general population to speak their minds and share their opinions of the day's events.

Social media data could be a significant boon to studies of complex political and social systems, as social media data is generally more abundant and rapidly produced than surveys or more typical research instruments. Recent attempts to leverage this data, however, have raised important methodological issues regarding data collection, including controlling sampling bias, applying human expertise to large-scale data, and minimizing costs. Addressing these concerns requires strategies that balance comprehensiveness and limited resources while bounding bias in the system as a whole [1].

This paper describes a process that integrates such strategies, optimizing cost, improving efficiency, and reducing biases. Using the context of a larger effort to predict violence during elections in Sub-Saharan Africa, this paper presents an implementation of this process and its evaluation. This study is grounded in Twitter, a popular microblogging platform, and focuses on collecting politically relevant conversation before, during, and after national elections in six countries (analysis of datasets built using this process are available in [2, 3, 4]).

Following a brief discussion of related work, we formulate this data collection process as a computationally assisted information retrieval (IR) task and describe the high-level structure of our technical framework and its implementation. Comparing across six elections in Sub-Saharan Africa, we demonstrate the method's performance, lessons learned from its deployment, limitations, and the weaknesses to be addressed in future work.

2. Related Work: Limitations and Opportunities

Data made available from the proliferation of online social media platforms have enabled new applications of network science at previously intractable scales and resolutions [5, 6, 7, 8, 9, 10]. Researchers, however, face a formidable validation challenge in controlling sampling bias stemming from missing data (textual content or users/groups). Sampling techniques for these online social media platforms remain unsatisfactory, often focusing on unrepresentative convenience samples of unknown populations using opaque interfaces with constrained access [11]. These access restrictions compound omitted data problems by introducing additional biases from closed-source samples, like Twitter's 1% public stream. Although popular due to

URI: http://hdl.handle.net/10125/50138
ISBN: 978-0-9981331-1-9

their low cost, it is unclear how these samples are created, introducing methodological inconsistencies and hindering inference [12, 13].

Even if a researcher had unconstrained access to these platforms, however, the expense in analyzing and filtering the data would likely be both prohibitive and inefficient given the volume of noise in such platforms. A conservative estimate suggests at least 90% of messages in Twitter are unrelated (i.e., noise) to even high-impact events like terrorist attacks [6].

To extract data relevant to the researcher's subject of interest and avoid the need to analyze such massive collections, researchers generally rely on keyword-based searches. Using IR techniques, documents that include these keywords are retrieved for the researcher's analysis. These keyword searches help constrain the dataset but can introduce a new source of bias in the query construction. Query expansion mitigates this intrinsic researcher bias by suggesting new keywords the researcher may have missed. Pseudo-relevance feedback (PRF) is the most common method for this expansion, which identifies tokens (meaningful sequences of characters or parts of text identified from breaking up documents) absent in the researcher's original query and which occur frequently in a query's result set [14]. Lexicons are another common approach. For example, research identifies tokens that frequently occur in tweets mentioning disasters across a broad range of crisis events [8]. Others have used content from webpages referenced in links in tweets for expansion [15]. Challenges in these methods virtually ensure resulting samples will have missing data, however, because researchers' vocabularies significantly vary [16] and brevity in social media platforms reduces the available context.

Specific to querying for election-related content, recent work detects relevant messages with a neural network, increasing query size by 200%, and uses automated techniques to reduce noise [17]. The approach outlined herein differs in cost-efficiency by leveraging free social media samples. In an alternative approach, Vosecky et al. improve their queries using models that detect both latent structure through topics and word order through language (topic language models) but do not seek to reduce biases due to omitted data [18]. All of these approaches generally assume access to the full population data, free from constraints imposed by data resellers. Moreover, many such approaches are fundamentally dependent on seed vocabularies and textual characteristics, which can be difficult to define as concepts become more complex.

The proposed computationally assisted process addresses these weaknesses and uses textual, social, and geographic attributes to perform query expansion. Leveraging free data for cost-effectiveness, human expertise to identify relevant content across countries, and machine learning models to automate this expertise allows for a more efficient and scalable approach than those above.

To ground this method's development, its use case is drawn from a larger research effort for a typical quantitative political science research question: do frustrations with economic, social or political inequalities attributed to governing authorities change during elections? Previous assessments [19, 20, 21] have used static instruments (like structural data or surveys) to capture perceptions, yet political discourse is dynamic and sensitive to changing events, particularly online [22, 23]. Leveraging social media to capture these perceptions provides a new avenue for advancing research. More importantly, the comparative nature of this research question requires diverse contexts, ranging from Botswana to Nigeria to Zambia, and from extremely violent to relatively peaceful elections. To our knowledge, a flexible, easy-to-implement technical framework that can bound systems of citizens discussing desired concepts during target events (like elections) across multiple countries has not been sufficiently presented in the literature. Rather than proposing a new lexicon for studying elections or other social or political phenomena; this effort presents a new *process* to capture representative and relevant social media data of the system in a rigorous, replicable, and scalable approach.

3. Computationally Assisted, Scalable Data Collection for Social Media

As previously mentioned, studying social and political systems via social media data imposes unique constraints on IR tasks. This section presents an overview of the proposed process and how it addresses these constraints. To set up this task, it expects the researcher to have a desired social or political research concept, to include: a textual description, location, and timeframe. With this initialization data, this framework retrieves a more comprehensive and less biased set of messages from a target online social media platform, subject to the following constraints and implementation details.

3.1. Search Constraints

A high-quality dataset should be unbiased, or representative of the true population, and relevant to the target concept. Collecting such a dataset is complicated by three chief constraints: 1) descriptions are always incomplete; 2) data access is restricted and costly; and 3) human expertise is limited. First, even given multiple experts on a topic, an exhaustive description of a concept is difficult to construct. For instance, when searching for keywords describing an event, an expert must make judgments with limited information and use cognitive heuristics to aid decision-making in the face of high volumes of information [24]. Identifying a perfect set of keywords is a "near-impossible task" for human experts, and even the most sophisticated text analysis systems are limited in practice by bias in these selections. Multiple experts are likely to uncover different information related to the same concept when describing even simple phenomena nearly 90% of the time [25]. Human experts also have implicit associations that may bias keyword selection [26], depending on their perceptions of the concept's related actors and the context in which they are situated. This proposed method supports these likely incomplete queries by automatically *supplementing* information.

This automated expansion, however, is still subject to constraints in the data or data source. Often, such data sources are opaque and incomplete. Small, public data samples (like Twitter's public sample stream) may be adequate for some questions, but research has shown their closed sampling processes introduce new, unknowable biases [12, 13]. Moreover, the small samples limit access to rare features, such as geolocation (estimated to be in just 3% of Twitter), or important aspects of communication, such as those between a small but influential group of users. Clever query strategies can alleviate some of these issues, but such methods generally rely on real-time access to the target platform, and are therefore limited in utility for retrospective search tasks [27, 28]. Furthermore, many search interfaces further restricts public access by imposing strict limits on query sizes. Twitter is no exception, generally restricting access to 1% of the full dataset.

Contracting for access to the full historical dataset is the best way to overcome these obstacles but also introduces a different set of restrictions. As mentioned in the description of Gnip's query interface and cost structure for instance, queries must be optimized *before execution* to create as complete a sample as possible. As such, the technical framework must control for *costly, query-based access to the full dataset* rather than relying solely on arbitrary access.

While access to the full dataset is valuable, it also creates a problem of scale: While much of the sample created will likely be irrelevant, an unknown but large portion is likely noise. Human input is required to draw appropriate meanings for complex concepts across diverse contexts in this data. The sheer scale of social media data in general, and Twitter in particular, makes human assessment intractable for all but a fraction of the content, as well as prohibitively expensive. For example, Amazon's Mechanical Turk platform has a minimum per-task charge of $0.03, and requires at least two coders for reliability, leading to more than $30K in costs for a single long-term event. Finally, an appropriate technical framework must *scale to the volume of data*.

3.2. Algorithm Overview

A computationally assisted framework to support social science-oriented search must therefore balance reducing bias, costs of acquiring data, and efforts to incorporate human expertise. This framework balances these issues by extending query axes to include textual, social, and spatial attributes; by integrating public Twitter archives to bootstrap queries; and by developing machine learning models to capture human expertise for judging relevant content.

In the process's first stage, three assumptions are made. First, the researcher has access to a set of social media messages T and identifies a particular event E to study (or concept within an event, such as grievances during elections). Second, the researcher provides an expert description of the event E or concept, including the timeframe of interest [d_{start}, d_{end}]. This description should include keywords, phrases, hashtags, user accounts Q_o, relevant to the event or concept but non-exhaustive, and a geographic area G covering the target population, if one exists. Third, access to the dataset T is restricted to a query interface (i.e., the researcher can not run arbitrary queries on the dataset).

The process, as shown in Algorithm 1, takes the expert description, the data archive T, and query expansion rounds as primary inputs. It then extracts a random sample T_{sample} from T in the time period of the event. To address the issue of incomplete data archives, this process bootstraps samples by expanding the relevant keyword set Q_E and identifying the most central or authoritative users A taking part in the event's

discussion, as described below. This social expansion is especially key: research shows small samples of Twitter data preserve significant network structure [12].

The algorithm then turns to the full dataset T and extracts messages matching the expanded query (T_q), messages authored by or mentioning the central accounts (T_a), and messages posted from the geographic region during the target timeframe (T_g), as the scarcity of geocoded messages on Twitter ensures this set will be limited. The framework then samples an equal number of messages from each query type — keywords, central accounts, and geocoded — and passes each set to human coders. The human coders assess each set of tweets for relevance to the event or concept, producing labels. These labels train a classifier to differentiate between relevant and non-relevant content at scale, addressing the cost of human assessment of millions of tweets. Irrelevant data is removed from the message set T_{all} with this classifier to create a comprehensive, highly-relevant dataset.

4. Methodology for Twitter Sampling

The design of this technical framework is intended to support flexibility and easy implementation to address a range of social and political research problems on social networking platforms. Query expansion and account centrality can be performed in many ways. This section discusses implementation details and data sources for sampling content from Twitter across six diverse countries in Sub-Saharan Africa, including query formats and methods for incorporating human expertise through sampling and machine learning. Events used in this evaluation focus on elections in these countries because such events provide a succinct timeframe in which grievances are popular and salient [20]. Social media is thus an ideal medium to test whether such event-driven and reactive communication [6, 29] has predictive potential.

4.1. Expert Descriptions of Concepts/Events

To capture content about a social or political concept or event on Twitter, our technical framework's first phase (Algorithm 1, line 1) incorporates an initial set of expert-provided parameters. These parameters include an initial set of keywords for which to search, a list of relevant social media accounts, and the geographic and temporal bounds (e.g., start and end dates) specific to the research question. For these election cases, this input includes: a summary of individuals involved (such as leaders, candidates, journalists, and activists), associated organizations, supporting constituencies, issues leading to contentious actions, and specific events in the relevant time period (e.g., protests, rallies, court cases, voter registration problems, candidate or party announcements, scandals, or political violence).

These summaries are distilled into topics and sub-topics, and keywords are generated through an iterative search process. For each topic, researchers choose keywords describing the target concept, while avoiding those that would overlap with other events (e.g., several elections use the hashtag #elections2014) where possible. These search terms then constitute an initial query set, which can be searched on Twitter's public user interface to identify hashtags and accounts for these topics. Within this sample, researchers then identify additional potential topics and iterate until no additional topics, central users, or hashtags seem to be uncovered, with the maximum iterations taking 8-10 hours.

4.2. Data Sources

Nearly half a billion tweets are shared on Twitter daily, but researchers cannot directly access this full volume. Instead, Twitter monetizes this content through "data resellers," or companies that provide more complete, fee-based access. For this effort, we use Gnip, Twitter's chief data reseller, which offers access to all historical tweets via their Historical Powertrack interface. Users pay for a subscription to this resource that provides an allocation of tweets and/or timeframes per month. Even with this interface, hoewver, researchers can only access data through queries, which can contain keywords, user accounts, geolocation bounds, hashtags, and other rules[1]. Cost is based on the number of days a query spans as well as the number of tweets retrieved: running new queries over the same timeframe increases costs, rendering the platform costly for use for query expansion. We alleviate these costs by maintaining an archive of Twitter's 1% public sample stream, as an analog for the second step (line 2 in Figure 1) of our technical framework. The sample stream is the primary source many researchers use for Twitter analysis, and while containing biases, research has shown network structure is generally preserved [12], making it an adequate platform for identifying central accounts. Researchers can also access the Internet Archive's Twitter Stream Grab, and Gnip's Powertrack API has a sample function

[1] http://support.gnip.com/apis/historical_api2.0/

Algorithm 1: Social Science Event Retrieval in Large-Scale, Restricted Microblog Collections

Data: $T, E(d_{start}, d_{end}, Q_0, G, queryExpansionRounds)$
Result: a set of microblog messages relevant to the target event

```
1  begin
2      T_sample ⟵ sample(T, d_start, d_end) ;    /* randomly sample T in target timeframe */
3      Q_E ⟵ Q_0 ;
4      A ⟵ List() ;
5      while i < queryExpansionRounds do
6          Q_E ⟵ expandQuery(T_sample, Q_E) ;   /* expand the initial query using PRF */
7          A ⟵ processNetwork(T_sample, Q_E) ;   /* identify socially relevant users */
8      T_q ⟵ queryKeywords(T, d_start, d_end, Q_E) ;      /* keywords-based textual query */
9      T_a ⟵ queryAccounts(T, d_start, d_end, A) ;          /* user-based social query */
10     T_g ⟵ queryGeo(T, d_start, d_end, G) ;          /* geolocation-based spatial query */
11     T_all ⟵ T_q ∪ T_a ∪ T_g ;
12     T_candidates ⟵ balancedSample(T_q, T_a, T_g) ;      /* sample equally from each set */
13     T_labeled ⟵ labelTweets(T_candidates) ;          /* human assessments for relevance */
14     Clf ⟵ trainClassifier(T_labeled) ; /* train a classifier with the labeled data */
15 return filter(Clf, T_all) ;                    /* use the classifier to filter data */
```

allowing subscribers to take random and repeatable samples of the full Twitter archive if no public archive is available for the timeframe or the 1% sample is too sparse for a given research concept.

4.3. Multi-dimensional Query Expansion

The next phase in the technical framework expands the query along textual and social dimensions and integrates the spatial dimension (lines 5-7 and line 10). Textual expansion identifies keywords whose frequencies in the set of matching messages are much higher than in the general sample (line 6). While typical query expansion finds unigrams and/ or bigrams that co-occur with keywords in the original query, our method leverages the samples to calculate high-signal keywords through a technique called Kullback-Leibler (KL) divergence [30] and finds keywords that occur more frequently in the set of query matches than in the sampled set. We find all messages in the random sample that match the original query, tokenize these messages into bags of words, rank words by how prevalent they are, and add the top ten words not present in the original query. Only unigrams are included as preliminary experimentation showed tokens in highly-divergent bigrams were often captured by top unigrams as well.

To identify highly-relevant users, we convert Twitter's retweet and mention activity into a directed graph of interactions. Vertices in this graph represent Twitter users, and edges denote mentions where vertex A has a directed edge to vertex B if A mentions or retweets B. Research shows highly followed or retweeted users are often not the most influential users [31], so we follow Kwak et al. and use a version of Google's PageRank algorithm to identify important accounts in this network [32] (other centrality measures could be used here as well). We then rank users by their PageRank score and append the top five users to the list of central accounts. We also perform only one round of query expansion, as preliminary results suggested additional query expansion rounds became too noisy.

We also experimented with named entity extraction from journalistic media to provide an additional frame of reference. Many entities extracted from journalistic media, however, were either already identified in the initial description or were place names that were rarely mentioned on Twitter.

4.4. Full Query Methods

Once new keywords and central accounts are extracted from the random sample, Gnip's full historical data archive is queried using the expanded query and a descriptor of the target geographic region (lines 8-10 in Figure 1). Messages in the subset T_q are retrieved by matching keywords, phrase, and substrings from the full set of expanded keywords. Messages from central accounts T_a gather important discussants in the network and contain messages authored by, is a retweet of, or

mentions a user in A (we also account for users who may have changed their Twitter handles over time). T_g matches the spatial dimension and includes all tweets with geocodes indicating the tweet was published from the region of interest. Our experiments focus on specific countries, so we construct a bounding box that covers the entire country. This technique introduces noise, since most countries are not rectangular in shape,. While the resulting tweet set T_{all} is likely to have lower precision than an unexpanded dataset, the following section discusses semi-automated methods for filtering this irrelevant data.

One tradeoff in this design is including a broad set of queries. Methods exist for developing better query sets, such as overlap filters [27], but these methods are motivated by limitations on the number of queries one can include in Twitter's public search APIs. Gnip has a much higher limit, allowing for thousands of such rules to be applied in a single Gnip query, so this motivation is not as strong in our research. In addition, we are more constrained by the number of days our searches cover than the number of tweets we retrieve, so high-precision queries are less critical.

4.5. Scaling Up Human Expertise

Finally, noisy data is filtered to enhance precision (lines 11-14), a necessary step since one should assume many of the tweets acquired in query expansion are noise. Prior to constructing a classifier to filter this noise, the algorithm first incorporates human relevance feedback as training data. These relevance assessments are generated for candidate tweets $T_candidates$ extracted from a balanced sample of approximately 300 tweets from each query type T_q, T_a, and T_g. Balancing these sets addresses the scarcity of geocoded tweets and the likelihood that many more messages will match keywords than either geolocation or central accounts.

This balanced set of candidate tweets and the event description are passed to a pair of human coders, who rate them as relevant, irrelevant, or unknown/not English. We then assess agreement between coders using Cohen's K: for each country-specific task, if K is *less than 0.61 for the first pair of coders*, a new annotator is added rather than retraining the original pair of coders. This process continues until K is greater than 0.61 for the highest-scoring pair of annotators, after which, the most-agreeing pair of coders iterate on the data until agreement exceeds 0.75 (which we achieved in at most one additional coding round). Replacing rather than retraining coders in these cases seemed to be an artifact of using non-native-English speakers for assessment, as non-native English speakers had difficulty understanding the brief content.

After this training, we perform an additional round of manual validation on a subsample of tweets. Tweets agreed to be relevant and irrelevant $T_{labeled}$ are then saved for classifier training, with all other messages being discarded (i.e. messages with inconsistent labels or messages labeled as unknown/non-English).

The last step (line 13) is to scale up relevance assessment, which is accomplished by training a classifier on human-labeled tweets and filtering the full set of tweets using this classifier (line 14). Each tweet is featurized into a bag of words and weighted by term frequency-inverse document frequency (TF-IDF) to avoid bias toward frequent terms. In this implementation,we train a set of Gradient Boosted Trees (GBTs), an ensemble classifier, on these feature vectors and their relevance labels [33]. GBTs consist of a series of iteratively-trained decision trees, and in each iteration, the training set is re-labeled to boost the importance of incorrectly labeled instances. We use a maximum of 100 iterations in our implementation. Other classifiers were considered as well (support vector machines, random forests, etc.), which performed equivalently, but GBTs were easier to distribute across clusters with parallel processing. Researchers without access to distributed processing systems can leverage existing analysis-as-service platforms from Amazon or Microsoft or can run these models locally. Depending on data sizes, ensemble methods like random forests work well in a serialized, non-cluster environment.

Once the classifier is trained, the framework applies it to the full set of tweets T_q, T_a, and T_g and returns all instances classified as relevant, completing the search algorithm.

5. Application to Sub-Saharan African Elections

Two case studies are presented here to explore information retrieved across query types, and to test this method across long-term samples from countries with relatively recent elections in Sub-Saharan Africa. The first case study explores six 90-day periods, shown in Table 1. These cases include national-level elections in Botswana (2014), Ghana (2012), Kenya (2013), Nigeria (2015), South Africa (2014), and Zambia (2015). The election samples draw from two months prior to the vote, and one month after. All elections take place in countries with Twitter populations that tweet

predominantly in English and have some presence on Twitter in the time period.

Table 1: Sub-Saharan African Election Periods

Country	Date of Election	Time Period
Botswana	24 October 2014	1 Sep — 18 Nov
Ghana	7 December 2012	10 Oct — 2 Jan
Kenya	4 March 2013	8 Jan — 3 Apr
Nigeria	28 March 2015	28 Jan — 28 Apr
South Africa	7 May 2014	15 Mar — 14 Jun
Zambia	20 January 2015	26 Nov — 18 Feb

5.1. Query Dimensions: Textual, Social, Spatial

The first study explores the relevance of information across query types: keywords, central accounts, and geographic location. We test whether each query type contributes relevant information, and if so, what volume, comparatively.

To answer these questions, we apply this technical framework to each of the six elections in Table 1, using queries shown in Table 2. For each election, we evaluate Cohen's K and find the task of assessing relevance achievable for humans annotating the balanced set $T_{candidates}$. Tables 3 below shows the precision, recall, F1, and F2 scores per query dimension per country. Precision, recall, and F-scores are widely-used in machine learning metrics.[2] To assess the classifier, we calculate average precision P, recall R, F1 score and F2 score using 10-fold cross validation on the set of labeled messages $T_{candidates}$. The F2 score is the primary metric for evaluation because it lends more weight to recall, consistent with the priorities of retrieving a sample with high recall to ensure full coverage of the population and discourse around the election.

The tables show each dimension contributes to generating relevant content within the corpus of contentious communication during election periods. However, the textual dimension (expanded keywords) generally outperforms the social dimension (central accounts), which generally outperforms the spatial dimension (geolocation), except for Kenya, where social outperforms textual, and Botswana, where spatial outperforms social. We find that these exceptions are likely due to the relative low Internet penetration rates and social media usage rates in Botswana, which are conversely quite high in Kenya, but not nearly as high as Nigeria or South Africa. Yet each brings a different type of information to bear that would otherwise be omitted from the dataset. This is key for valid inference for research questions about social and political systems online.

[2]Recall measures the ratio of true positives out of actual positives, while precision measures the ratio of predicted positives. The F1 score is the harmonic mean of precision and recall, while the F2 score is similar but gives more weight to recall.

5.2. Classifying Relevance

Second, we determine how well human relevance decisions can be recovered using the GBT classifier. Our success in classifying relevance in two stages, across query types, is presented with two metrics: weights and Area Under the Precision Recall Curve (AUCPR) in Table 4. AUCPR shows how well we can classify based on human annotations, and feature weights show how much each dimension contributes to this classification. The AUCPR statistic shows a classifier using all four dimensions achieves a mean score of 0.9725 for classifying relevant information.

6. Lessons Learned and Limitations

The major implication of this work is demonstrated in the per-query-type scores: combining the textual, social, and geographic dimensions of relevance (i.e., keyword queries, central accounts, and geographic bounds) yields superior performance (in terms of F2 score) than standard keyword queries. This computationally assisted framework achieves performance while minimizing fee-based queries by bootstrapping the query process with free (or cheaply acquired) Twitter samples. Such samples can be developed in-house or taken from online sources like the Internet Archive[3].

High variance in relevant coverage across elections is an expected part of our recall orientation, which we address through human annotation and relevance feedback. This feedback provides a method for filtering these large datasets even when relevance is low.

This work further demonstrates the value of non-textual signals of relevance, which are essential given the difficulty in creating unbiased event or concept descriptions [16, 34, 24]. Social and spatial dimensions of relevance reduce dependence on keyword sets and expand the set of content to which annotators are exposed. Using a sample of the social network, we identify accounts that play a major role in the event's discussion regardless of the verbiage they use, and the spatial dimension allows us to target users who are located near the event of interest. While PRF-based expansion is helpful, we demonstrate that social and spatial expansions add value as well, with approximately

[3]`http://archive.org`

Table 2: Sub-Saharan African Elections: Representative Keywords and Central Accounts

Country	Keywords	Hashtags	Central Accounts
Botswana	botswana∗, ward∗, gaborone, wins∗, south∗, khama	#botswana, #elections2014, #bwelections2014	@BDPnews, @BCPBotswana, @BWGovernment, @iecbw
Ghana	ghana∗ elect∗, "Akufo-Addo", "Atta-Mills"	#Ghanadecides, #Ghanaelections, #Mahama, #mahama100	@jdmahama, @ghanaelections, @GhanaDecides, @NAkufoAddo
Kenya	Kenya, election, kenyatta, nairobi, presidential, doctored	#Kenya, #JubileeGovt, #election, #Kenyatta	@IEBCKenya, @marthakarua, @Peter_Kenneth, @JamesOleKiyiapi
Nigeria	nigeria∗, elect∗, postpon∗, Goodluck Jonathan	#NigeriaDecides, #2015Elections, #BringBackJonathan	@GEJonathan, @ThisIsBuhari, @PDPNigeria, @INECNigeria
South Africa	south africa∗, elect∗, corrupt∗	#elections2014, #Ayisafani, #Siyanqoba, #MyVoteSA	@SAPresident, @helenzille, @Julius_S_Malema
Zambia	zambia∗, elect∗, lusaka, snap election, president∗	#zambiadecides, #edgarlungu, #lungu, #guyscott, #hichilema	@FDDZambia, @NAREPzambia, @ZambiaElections, @mmdzambia

Table 3: Per-Feature Relevance Classification Performance

	Prec	Recall
Botswana	0.260	0.696
Ghana	0.296	1
Kenya	0.489	0.701
Nigeria	0.556	0.495
South Africa	0.401	0.298
Zambia	0.396	0.772
Mean	0.416	0.660
	F1	0.511
	F2	**0.591**

(a) Text Analysis

	Prec	Recall
Botswana	0.215	0.152
Ghana	0.412	0.774
Kenya	0.433	0.190
Nigeria	0.535	0.383
South Africa	0.409	0.512
Zambia	0.335	0.215
Mean	0.400	0.371
	F1	0.380
	F2	**0.375**

(b) Network Analysis

	Prec	Recall
Botswana	0.224	0.205
Ghana	0.427	0.130
Kenya	0.384	0.134
Nigeria	0.533	0.131
South Africa	0.356	0.182
Zambia	0.340	0.198
Mean	0.377	0.163
	F1	0.228
	F2	**0.184**

(c) Spatial Analysis

Table 4: Relevance Classifier Performance (Weights and AUCPR)

Election	Textual	Social	Spatial	Time	All
Botswana	2.5/ 1.0	0.5/ 0.8	0.2/ 0.7	0.0/ 0.3	0.96
Ghana	1.9/ 1.0	0.5/ 0.9	0.0/ 0.7	0.0/ 0.5	0.98
Kenya	2.1/ 1.0	1.0/ 0.9	-0.04/ 0.6	-0.0/ 0.3	0.98
Nigeria	2.0/ 1.0	1.2/ 0.9	0.11/ 0.7	-0.5/ 0.6	0.96
South Africa	2.2/ 1.0	0.3/ 0.9	0.0/ 0.4	0.0/ 0.4	0.96
Zambia	1.7/ 1.0	0.9/ 0.9	0.3/ 0.7	-0.2/ 0.4	0.99
Mean	2.1/1.0	0.7/0.9	1.0/0.7	-0.1/0.4	0.97

100,000 and 1.6 million additional relevant tweets captured, respectively, that were originally missed by standard query expansion approaches.

Like central accounts, spatial queries (matching all tweets posted from the target region regardless of keywords) introduce unique information and address bias by exposing human annotators to new content, but the loss in scores and increased financial cost of acquiring them cast doubt on their value. Since geolocation queries are essentially a pseudo-random sample of the full Twitter stream, this low precision is consistent Twitter's limited coverage of major events [35]. One can mitigate the cost of geolocation queries by reusing them in other contexts: researchers interested in multiple events in a given country and timeframe need only pull this data once.

Despite these advantages, this technique's precision is limited by the performance of the machine learning model. While it addresses the untenable task of annotators reading millions of messages, its reliance links performance to that of the classifier. While GBTs perform well in recovering human labels, false positives are still present, especially in the long-term events. For instance, elections occur in multiple countries simultaneously, and since we focus on long-term events, the classifier can incorrectly identify tweets about elections in the United States and India as relevant (the hashtag #elections2014 was particularly popular across several national elections). Likewise, references to voting tended to result in a relevant label, leading to acquiring tweets about several vote-based reality television shows. Early results suggest using the social structure in the retrieved data to filter information about unrelated events; e.g., Twitter users discussing US elections tend to form a community in the Twitter graph that is well-separated from those users discussing an election in South Africa. This community analysis is an open and promising research area.

Besides precision, we also cannot ensure our algorithm is retrieving all possible relevant messages. Results demonstrate including central accounts and geolocation mitigate this recall issue by capturing unique messages, but without human assessment of every single message posted (impractical given volume and access restrictions), one cannot be sure all data has been captured. A possible solution could use Gnip to acquire all tweets posted during our short-term events and use our classifiers to infer relevance, but with more than half a billion tweets posted per day, our monthly allocation would only allow us to acquire about one hour's worth of data for a single event. As such, the

samples collected through querying must be sufficient. These queries are also implicitly connected to our evaluations: Since many query expansion mechanisms exist, other approaches (e.g., co-occurrence networks) may exhibit equivalent recall, which also could be explored in future work. Crowdsourcing and hidden stream methods could address this issue to some degree, but one would need to know the event of interest at time of collection.

Recall and precision aside, we present our technical framework as a general approach for retrieving relevant information on social media for concepts in social and political systems that are location-based (e.g. elections) and show it performs well for our purposes. The presented technical framework addresses three needs: 1) reducing bias in retrieved data, 2) reducing data acquisition costs, and 3) supporting relevance evaluation at scale. Results show we can reduce query costs by leveraging public Twitter archives, expand keyword queries by identifying users who are central in an event's discussion, and support relevance evaluation at scale with machine learning models. By tracking central users and analyzing geolocated tweets, we retrieve thousands of relevant tweets that would be missed if one relied solely on keywords and keyword expansion. Combining standard keyword expansion queries with the social dimension also achieves the highest F2 scores in our evaluations, regardless of event duration. Furthermore, integrating human relevance feedback on a few hundred messages allows us to train a classifier to identify relevant content at scale and with high accuracy.

While there is no substitute for human input when identifying relevant information in microblogs, and open issues still exist, the procedure outlined herein supports the unique constraints of large-scale sampling for social science topics. Different events or research subjects may vary in affordances (e.g., not all subjects may have a geographic component), but this framework is sufficiently flexible as to support different contexts. By codifying this process, this work can accelerate research into and replicability of microblog analysis in socially and politically relevant research across a broad spectrum of questions.

7. Acknowledgements

This research was supported by the National Science Foundation and an appointment to the Intelligence Community Postdoctoral Research Fellowship Program at the University of Maryland, administered by Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the Office of the Director of National Intelligence.

References

[1] W. N. Dunn, "Probing the boundaries of ignorance in policy analysis," *American Behavioral Scientist*, vol. 40, no. 3, pp. 277–298, 1997.

[2] K. Donnay, E. Dunford, E. McGrath, D. Backer, and D. E. Cunningham, "Integrating conflict event data." *Working Paper*, 2017.

[3] E. McGrath, E. Dunford, C. Buntain, K. Donnay, D. Backer, and D. E. Cunningham, "Grievance and electoral political instability in sub-saharan africa." in *American Political Science Association*, August 2017.

[4] G. Mohler, E. McGrath, and C. Buntain, "Hawkes binomial topic model with applications to coupled conflict-twitter data," *Working Paper*, 2017.

[5] P. Barberá, "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data," *Political Analysis*, vol. 23, no. 1, pp. 76–91, Jan. 2015. [Online]. Available: http://pan.oxfordjournals.org/content/23/1/76

[6] C. Buntain, E. Mcgrath, G. Lafree, Golb, C. Park, C. Park, E. Mcgrath, G. Lafree, C. Park, and J. Golbeck, "Comparing Social Media and Traditional Surveys Around the Boston Marathon Bombing," *#Microposts: 6th Workshop on Making Sense of Microposts*, 2016.

[7] V. Lampos, T. De Bie, and N. Cristianini, "Flu detector-tracking epidemics on twitter," *Machine Learning and Knowledge Discovery in Databases*, pp. 599–602, 2010.

[8] A. Olteanu, S. Vieweg, and C. Castillo, "What to Expect When the Unexpected Happens: Social Media Communications Across Crises," in *In Proc. of 18th ACM Computer Supported Cooperative Work and Social Computing (CSCW'15)*, no. EPFL-CONF-203562, 2015.

[9] M. J. Paul and M. Dredze, "You Are What You Tweet: Analyzing Twitter for Public Health," *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pp. 265–272, 2011. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2880

[10] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Tweet the Debates: Understanding Community Annotation of Uncollected Sources," in *Proceedings of the First SIGMM Workshop on Social Media*, ser. WSM '09. New York, NY, USA: ACM, 2009, pp. 3–10. [Online]. Available: http://doi.acm.org/10.1145/1631144.1631148

[11] A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman, "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries," *SSRN Preprint (8 March 2017)*, pp. 1–44, 2016.

[12] S. González-Bailón, N. Wang, A. Rivero, J. Borge-Holthoefer, and Y. Moreno, "Assessing the bias in samples of large online networks," *Social Networks*, vol. 38, no. 1, pp. 16–27, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.socnet.2014.01.004

[13] F. Morstatter, J. Pfeffer, H. Liu, and K. Carley, "Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose," *Proceedings of ICWSM*, pp. 400–408, 2013. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewPDFInterstitial/6071/6379

[14] C. D. Manning, P. Ragahvan, and H. Schutze, *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2009. [Online]. Available: http://www.informationretrieval.org/

[15] R. McCreadie and C. Macdonald, "Relevance in Microblogs: Enhancing Tweet Retrieval Using Hyperlinked Documents," in *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, Paris, France, France, 2013, pp. 189–196. [Online]. Available: http://dl.acm.org/citation.cfm?id=2491748.2491787

[16] S. Gouws, D. Metzler, C. Cai, E. Hovy, and M. Rey, "Contextual Bearing on Linguistic Variation in Social Media," *Proceedings of the workshop of language in social media*, no. June, pp. 20–29, 2011.

[17] P. Vijayaraghavan, S. Vosoughi, and D. Roy, "Automatic detection and categorization of election-related tweets." in *ICWSM*, 2016, pp. 703–706.

[18] J. Vosecky, K. W.-T. Leung, and W. Ng, "Collaborative personalized twitter search with topic-language models," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 53–62.

[19] L.-E. Cederman, N. B. Weidmann, and N.-C. Bormann, "Triangulating horizontal inequality Toward improved conflict analysis," *Journal of Peace Research*, vol. 52, no. 6, pp. 806–821, 2015. [Online]. Available: http://jpr.sagepub.com.proxy-um.researchport.umd.edu/content/52/6/806

[20] H. Fjelde and K. Höglund, "Electoral Institutions and Electoral Violence in Sub-Saharan Africa," *British Journal of Political Science*, vol. 46, no. 2, pp. 297–320, 2016. [Online]. Available: http://journals.cambridge.org/article_S0007123414000179

[21] D. Miodownik and L. Nir, "Receptivity to Violence in Ethnically Divided Societies: A Micro-Level Mechanism of Perceived Horizontal Inequalities," *Studies in Conflict & Terrorism*, vol. 39, no. 1, pp. 22–45, 2015. [Online]. Available: http://dx.doi.org/10.1080/1057610X.2015.1084162

[22] F. Diaz, M. Gamon, J. M. Hofman, E. Kcman, and D. Rothschild, "Online and Social Media Data As an Imperfect Continuous Panel Survey," *PLoS ONE*, vol. 11, no. 1, p. e0145406, Jan. 2016. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0145406

[23] D. Garcia, A. Kappas, D. Küster, and F. Schweitzer, "The Dynamics of Emotions in Online Interaction," *Royal Society Open Science*, vol. 3, no. 8, p. 160059, 2016. [Online]. Available: http://arxiv.org/abs/1605.03757

[24] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," in *Utility, probability, and human decision making*. Springer, 1975, pp. 141–162.

[25] C. Jolls and C. R. Sunstein, "The law of implicit bias," *California Law Review*, pp. 969–996, 2006.

[26] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The Vocabulary Problem in Human-system Communication," *Commun. ACM*, vol. 30, no. 11, pp. 964–971, nov 1987. [Online]. Available: http://doi.acm.org/10.1145/32206.32212

[27] P. Bommannavar, J. Lin, and A. Rajaraman, "Estimating Topical Volume in Social Media Streams," in *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2016, pp. 1096–1101. [Online]. Available: http://doi.acm.org/10.1145/2851613.2851810

[28] J. Sampson, F. Morstatter, R. Maciejewski, and H. Liu, "Surpassing the Limit: Keyword Clustering to Improve Twitter Sample Coverage," *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pp. 237–245, 2015.

[29] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "AIDR: Artificial intelligence for disaster response," *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, no. October, pp. 159–162, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2577034

[30] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951. [Online]. Available: http://www.jstor.org/stable/2236703

[31] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," in *Fourth International AAAI Conference on Weblogs and Social Media*, may 2010. [Online]. Available: http://aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1538/0

[32] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 591–600.

[33] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[34] P. J. Hayes and S. P. Weinstein, "CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories." in *IAAI*, vol. 90, 1990, pp. 49–64.

[35] C. Buntain and E. Mcgrath, "Tracking Emigration from Conflict Areas using Social Media," in *Workshop on Social Media Demographic Research at ICWSM*, 2016.