Random Sampling in Corpus Design: 
Cross-Context Generalizability in Automated Multi-Country Protest Event Collection

Abstract
What is the most optimal way of creating a gold standard corpus for training a machine learning system that is designed for automatically collecting protest information in a cross-country context? We show that creating a gold standard corpus for training and testing machine learning models on the basis of randomly chosen news articles from news archives yields better performance than selecting news articles on the basis of keyword filtering, which is the most prevalent method currently used in automated event coding. We advance this new bottom-up approach in order to ensure generalizability and reliability in cross-country comparative protest event collection from international and local news in different countries, languages, sources and time periods, which entails a large variety of event types, actors, and targets. We present the results of comparing our random-sample approach to keyword filtering. We show that the machine learning algorithms, and particularly state-of-the-art deep learning tools, perform much better when they are trained with the gold standard corpus from a randomly selected set of news articles from China, India and South Africa. Finally, we also present our approach to overcome the major ethical issues that are intrinsic to protest event coding.
Introduction

Automated protest event extraction promises to overcome the prohibitive costs of human coding of protest events from a variety of languages, sources, countries, and time periods in order to create large and comprehensive datasets of protest events. Nevertheless, there are lingering questions as to the reliability and generalizability of automated text processing tools, that is, whether they are a match to human coding in terms of validity, accuracy, and completeness across different contexts. What automated methods boast as their advantage, that is, their ability to work with thousands of texts from a multitude of sources comes with a new set of challenges which human coding has not had to tackle. Maintaining completeness, validity, and reliability over an infinitely larger pool of data becomes a key challenge that taxes the development process of automated text processing tools. The difficulty of achieving reliability in such a setting has much to do with the variability that the object of inquiry enjoys throughout the immense data universe in a cross-context setting. Put more concretely, the variety in protest characteristics over time and across different social, political, linguistic and cultural contexts and across news sources compounds the difficulty in identifying protest events accurately and extracting event information in a complete manner. Hence, the need for generalizability makes the issues of reliability an ever-challenging concern (Wang et al. 2016a, Ettinger et al. 2017, Ward et al. 2013). Although the context variability that we describe here is more pronounced in multiple country settings, many of the same challenges appear also in studies that focus on single countries. This is due to the fact that different sources of event data, or different time periods will be a source of variety even in a single country.

The approach that we take is part of building the Global Contentious Politics (GLOCON) dataset (glocon.ku.edu.tr), which forms part of the Emerging Welfare Project (emw.ku.edu.tr) funded by the European Research Council. It aims to tackle the issues of reliability by devising generalizable tools capable of adapting to different contexts from scratch. The Emerging Welfare Project aims at explaining the contemporary welfare state development in emerging market economies (focusing on Argentina, Brazil, China, India, South Africa and Turkey) by considering the effects of social movements. This highly quantitative endeavor entails building quantitative measures on social movements from this diverse set of countries, by constructing an event database of all contentious politics events for the mentioned focus countries from the 1990s onwards. The database will be created through using automated text processing tools, the development of which bears the difficulties of context variability in all its forms.

In this study, we describe our method of automated tool development with a particular focus on the creation of our gold standard corpus (GSC) of contentious politics events. A high quality GSC is a widely accepted method of creating training, test and evaluation data for automated text processing. We highlight the sampling process of the documents that comprises our GSC, which we are convinced is a key step in accommodating the variability in the data universe in a way that assures data validity while achieving generalizability. We argue that a GSC built by full-text human reading of unfiltered samples of source data achieves a better approximation of the data universe (i.e. of all news articles that report contentious politics events in our case) compared to the sampling of raw data by keyword filtering. We refer to our approach as “random sampling” while contrasting it to a “keyword filtering” approach, which we find introduces selection bias into document compilation. The greater representativeness of random sampling improves the validity of the resulting database in measurable ways in both single and multi-country settings. Furthermore, it maintains greater consistency in the face of the larger variety of language and texts that our diverse set of focus countries present.
Our article is structured as follows. We begin first with a short survey of the history of protest event collection methods focused on the shift from manual to automated methods. We then present a critical reading of existing automated methods, focusing on their failure to generate reliable and generalizable results. The next section portrays our approach, which consists of creating a GSC based on a random sample of news reports from a variety of contexts (countries, sources, languages and time periods) and applying state-of-the-art machine learning Natural Language Processing tools. Then, we present our result, which illustrates that the random sample method outperforms keyword-based methods across different methodological and contextual settings. The last section presents our ethical concerns about the protection of subjects and data sharing, followed by the discussion and conclusion.

Data Validity and Reliability in Protest Event Research

Protest event analysis is the most common approach for social movement scholars as it is an unobtrusive and context-sensitive technique that can convert unstructured matter into large volumes of data, in a cross-country, cross-time and cross-issue comparative character (Hutter 2014).

As Hutter’s (2014) literature review aptly portrays, the history of protest event analysis from news sources goes back to the 1970s - or even the 1930s, when P.A. Sorokin (1937) developed event coding as part of his book Social and Cultural Dynamics. Manual coding of contentious political events was initiated in the 1960s and 1970s by scholars who were interested in a large-N comparison with a relatively smaller number of event characteristics (Russett et al. 1977; Shorter and Tilly 1974; Tilly et al. 1975). Later, scholars developed more detailed event coding schemes that would allow them to identify domestic and cross-country variations in contentious politics in terms of event types and other protest characteristics (Jenkins and Perrow 1977; McAdam 1982; Tarrow 1989).

A lingering methodological problem in protest event research has been the degree to which newspapers are reliable sources of information on real world protest events. Questions have been raised about the validity and robustness of research based on newspaper data, especially the tendency towards selective coverage of events based on their newsworthiness (i.e. violence, size, drama, etc.), proximity to reporting systems, and the ideology and issue agenda of the news source (Della Porta and Diani 2020; Koopmans 1999). A potential solution is the systematic quality of such selection bias. If bias is consistent over time and space, then researchers may be able to control for this error (Koopmans 1999: 93). Another idea is to use multiple sources, which may reduce selective coverage. In any case, though, the biases of various newspaper sources among themselves, and those that involve print media needs to be assessed if their effect on research validity is to be understood and controlled.

Conducting such an assessment of the validity of news sources or event datasets is not straightforward, as it is really difficult to create or determine a set of protest events that can serve as an absolute benchmark. In the absence of such an accepted ground truth, the only viable method has been a “relativistic strategy” whereby datasets and underlying sources are compared to each other to determine their “relative bias” (Strawn 2008: 149-50; Jenkins & Maher 2016: 53). Jenkins & Maher (2016) argue that since all event datasets are partial and limited, all evaluations is ultimately relative. While drawing on multiple sources, including non-news sources, should reduce selection error, multi-source datasets are also biased and partial, providing an incomplete picture of the real world and introducing the additional problem of duplicate reports of the same events. Furthermore, datasets that draw on multiple news sources which target similar audiences might end up magnifying certain shared biases, as demonstrated by Myers and Schaefer Caniglia (2004). Thus,
the variety of sources may be more important than their sheer number in achieving more completeness. The culmination of this rich debate on source selection and bias is that source heterogeneity, and, to a lesser degree, multiplicity are key features of event databases that enhance the validity of protest data, and with that, the robustness of social movements research.

Multiplying and diversifying sources in order to boost the validity of event data is costly. While the digitization of news and the development of automated approaches to coding events and managing large inventories of event data have blossomed (Imig and Tarrow 2001, Nam 2007), there are still major challenges to be solved. The first automated system was the Kansas Event Data System (Gerner & Schrodt 1995), which relied on sparse-parsing and the coding of single sources, such as Reuters newswire or Agence France Presse (AFP). Using human coding methods, others developed ways of using non-news sources, like local police records (Maney and Oliver 2001) and reports of the National Parks Service (McCarthy et al. 1995). In the last decade, the unprecedented development of computational power based on new generations of processors has enabled the application of artificial intelligence techniques and new fully or semi-automated methods to collect and process contentious political event data from multiple news sources, such as GDELT, ICEWS, EMBERS and SPEED, POLDEM and MMAD (Wang et al. 2016a, 2016b; Wang 2018; Lorenzini et al. this issue; Althaus et al. this issue).

In terms of data validity and reliability, automated event databases have their own issues. Accurately classifying texts and extracting relevant information from them as accurately as humans is still an aspiration in automated coding. In large part, this is due to the difficulties that automated tools face when interpreting the complexity of meaning in the narrative content of texts, as opposed to their structural qualities, such as word counts, presence of specific trigger expressions or syntaxes that are more uniform. Nardulli et al. (2015: 151-152) point out that fully automated methods of event extraction perform better on the manifest content of texts, that is, its morphological qualities such as the presence, frequency, and proximity of a set of linguistic units (i.e. "a bag of words") that relate to a simple topic that is not context-dependent. Humans have contextual knowledge and thus are much better at discerning the nuances that expressions acquire within a particular historical context. The rigidity the automated models display when confronted with context variety is a key factor, which limits their performance in delivering consistent and reliable event data over large spans of space and time. In their study of the most comprehensive and widely used global events datasets, Wang, et. al. (2016b) demonstrate that automated systems like GDELT and ICEWS are less reliable relative to human coding as specificity increases. Wang et al. (2016b: 10) conclude that the flaws of these datasets lies not only in their automation procedures, which can in principle be optimized to a certain extent through de-duplication and various automated mechanisms that improve precision, but also the sources that they use to gather event data, namely English language international news networks.

Utilizing such international news agencies as data sources is a common way of achieving greater uniformity in terms of source material so that automated tools can operate in a more consistent way. The uniformity granted by the single language and similar editorial style, however, comes at the expense of the source heterogeneity that is conducive to representing the diversity of protest events in different parts of the world. Another team of researchers who devised a semi-automated model to create a protest event data underlines the importance of incorporating local news in augmenting the completeness of context-dependent event data (Weidmann & Rod, 2019:41-44). The fact that international news outlets target an elite, Western audience, and thus have a bias to cover events that elicit interest among this audience leads to underrepresenting or even ignoring a good deal of events that are of local significance. After comparing prominent
international news agencies in terms of protest event coverage for Kyrgyzstan, they find that coverage increases significantly with the addition of BBC Monitoring into the source mix, as it contains translated material from local news sources. Herkenrath & Knoll’s (2011) analysis of the coverage of protest events in Paraguay, Argentina, and Mexico in the English language aggregate news source LexisNexis found that the number of protest events gathered from local newspapers of these countries that are included in the database is extremely low. While they affirm that the selection bias stems partly from the size and intensity of the protest, and location of the newspaper, as demonstrated by previous studies, they identify biases related to the location of the protest to be far more critical. The levels of coverage also differ significantly among three countries where events in Mexico have the highest representation, and those in Paraguay have the least. This is in line with the results of Wang et al. (2016b), who found that correlation between automated databases from English language international sources and manual databases with local sources varies extensively across countries. The lesson to be learned from all these studies is that relying on a small number of international sources in the English language harms the completeness of databases in that they carry the potential of augmenting, or at the very least, not resolving selection bias problems.

Selection bias in data sources and source types is not the only issue of data validity in automated databases. An equally important but somewhat less addressed area of concern is methods of data compilation, i.e. methods involved in both sampling (Koopmans, 1999) and acquisition of news reports (Maney & Oliver, 2001; Strawn 2010). With the explosion of the availability of online content, and the emergence of data providers like LexisNexis and Proquest that streamline and facilitate easy access to it, keyword searching has become an integral part of document selection in almost all projects that strive to create event databases. Along with the ease of access, this method introduces an additional layer of selection, which needs to be assessed as a potential challenge to validity. Strawn (2010: 72-73) assessed the validity of a keyword search protocol in Mexican news sources in order to determine the trade-off between “search validity,” i.e. the ability of keywords to identify valid protest events, and “efficiency,” i.e. the ratio of false hits relative to total hits. Full text reads by human coders of all documents sampled from particular times served as the basis for assessment. Search validity varied among event forms and performed best on more standard forms of protest (e.g. sit-ins, marches) and poorly for less conventional protests (e.g. open letters, hostage-taking). Similarly, Maney and Oliver (2001: 159) found earlier that generic keywords “consistently miss event coverage that describes events without reference to conventional terminology.”

In training automated models, a common technique is to train an initial automated classifier to compile relevant articles (Kriesi et. al. 2019; Croicu & Wiedman 2015). The training set for said classifier comprises a human coded set of documents that are initially gathered via use of keyword searches. There are two main reasons for using keyword searches in compiling the initial raw data. Firstly, online collections of news sources such as LexisNexis impose this as they offer the data as results of search engines. Secondly, using protest event related keywords acts as an initial mechanism of oversampling relevant documents, i.e. reduces the noise to signal ratio in the raw data that will be manually annotated by humans, a procedure which is otherwise highly costly. Keyword lists are designed to include many words and synonyms which denote various types of protest action and are considered to be as broad as it takes to include the most relevant news reports.

---

1 It is worth mentioning that which events are conventional and unconventional will also be context dependent. Citizen strikes, during which protesters close their shops and businesses, or force local businesses to close shop would be an unconventional event in most contexts while it is a common occurrence in India. This also highlights the difficulty of coming up with a comprehensive keyword set that would encompass different types of events in different contexts reliably and efficiently.
and quite a few that are irrelevant, which are later eliminated in a supervised setting. Given the above-mentioned shortcomings of keyword searches, it is surprising that this phase of research is not paid the attention it deserves by automated event database projects. We aim to fill this void with this study where we assess the validity and reliability of automatic document classifiers that are trained on the basis of keyword-filtered raw data vis-a-vis full-text human reads of randomly sampled documents from the entirety of available documents in our sources.

The Approach of the GLOCON Project

In the GLOCON Project, we develop fully automated tools for document classification, sentence classification, and detailed protest event information extraction that will perform in a multi-source, multi-context protest event setting with consistent recall and precision for each country context (Hürriyetoğlu et al. 2020). In order to cope with the challenges of developing generalizable tools that can handle source text heterogeneity, we designed the tool development process to incorporate news sources from multiple contexts, which may contain different grammar and diction. However, rather than developing tools from scratch for every country context, which would involve compiling training, test, and development data from each individual country, we distribute these phases into multiple countries. That is, we test the models trained on one particular country context (source country) on a different country data (target country). Then we optimize the automated tool performance by extending the training data with more randomly sampled training data from the target country until recall and precision scores reach satisfactory levels. Each set of new target country data is thus incorporated into the GSC in an incremental fashion in a way that allows the social scientists (domain experts yerine kullandım) to guide the annotation process so as to adapt to the new lexicon of contentious politics in each country.

We define a protest as “a collective public action by a non-governmental actor who expresses criticism or dissent and articulates a societal or political demand” (Rucht et al. 1999: 68), and instances or episodes of social conflict, which are based on grievances or aspirations to change the social and political order. The GLOCON Project creates a database of events from a wide spectrum of collective action forms, such as strikes, rallies, boycotts, protests, riots, clashes between social groups as well as demonstrations – i.e. “repertoire of contention” in a larger sense (Tarrow 1994, Tilly 1984). It aims to identify, by use of automated text processing tools, the participants, organizers, targets, places and facilities of events; whether they are violent, urban or rural; and ethnicity, religion, ideology and caste characteristics of event organizers and participants.\(^2\)

The development process of automated text processing tools depends on a high quality, human-annotated GSC of documents to serve as the basis for each text processing task of text classification, sentence detection, and information extraction. It currently contains 5,978 randomly selected news articles from Indian, Chinese, and South African news sources in English. In addition to local news sources, another random sample from an international English-language news source containing news about China was also included so that the corpus contains both country and source variability. All of these documents are labeled as protest and non-protest at the document level. In the sentence classification task, all sentences of the protest articles are labeled as containing event information or not. Finally, these protest articles are annotated at the token level for the extraction of detailed event information on characteristics such as the event type, protest category of the event, event trigger (a word or a phrase that indicates the occurrence of a protest event, e.g., “protested,” “demonstrated”), any reference to the event trigger, and the event’s place, time, and actors.

\(^2\) Please see our annotation manual for our more detailed definition of contentious political event and our event ontology https://github.com/emerging-welfare/general_info/tree/master/annotation-manuals.
participants, organizers and targets). The corpus has been processed in order to create a pipeline of deep learning-based machine learning (ML) models. The variety of sources has allowed us to study the cross-context robustness and generalizability of the ML models. A GSC of protest events that can enable large scale, multi-source social scientific studies should be representative of the content it aims to capture. Moreover, it should enable quantifying the performance of the automation across contexts. Therefore, using available corpora that are already being used in research such as English Gigaword (Parker et al., 2011) is not an option for this setting. In order to satisfy these requirements, our methodology is designed to contain and incorporate multiple sources and countries and apply a detailed annotation methodology without sacrificing quality (Hürriyetoğlu et al., 2020).

Our corpus design aims at securing completeness and validity through a variety of principles and mechanisms. In order to secure utmost completeness, we start with unfiltered random samples of documents compiled from every news source in order to be able to incorporate country and source-specific characteristics of protest events from scratch. The random sampling approach makes the task challenging but more representative of reality. Keyword lists tend to miss certain protest events due to indirect or less explicit language (e.g. "workers stopped working" to refer to a strike) (Hürriyetoğlu et al., 2019a, Hürriyetoğlu et al. 2019b). They might even exclude a whole class of events due to the lexical variance across contexts when referring to particular event types (e.g. "dharna" and "bandh", referring to “sit-in” and “citizen strike” respectively in India). In order to take full advantage of random sampling in terms of completeness, a domain expert in politics from the target country provides country-specific training on contentious politics to our annotators for better understanding of dynamics of social movements and peculiar and/or prevalent types of protest that are specific to each country. This enables them to better recognize protest events even when reporting is less explicit and indirect, and when they encounter local event characteristics that might not be obvious to someone who is unfamiliar with that context. Country experts also ensure sensitivity to local language and news practices in how protest is reported.

To ensure consistency, the whole annotation effort is based on an annotation manual created and maintained by an expert, who continuously monitors the annotations to achieve a high level of inter-annotator agreement in terms of Krippendorf’s alpha (Krippendorf et al. 2014). This expert also updates the manual through elaborating existing annotation rules to accommodate novel context-specific characteristics based on feedback from annotators. Thus, the same annotation rules can be better applied to data collected from different sources and countries, which enables obtaining comparable measures of automatic tool performance across contexts.

Checking the quality of annotations to maintain high levels of validity and consistency is performed through randomized spot-checks, i.e., 10% of annotator agreements are checked by the annotation supervisor for mistakes, and a cross-task feedback mechanism that is enabled by our process design. The three levels of annotation are separate but integrated in the sense that they form a pipeline in which a single document goes through each individual step, and each step is built upon the result of the previous step. The aim here is to maximize time, resource efficiency and performance by utilizing the feedback of each level of annotation for the whole process. This, in turn, enables error analysis and optimization during annotation and tool development efforts, which improves inter-annotator agreement (IAA) and lessens the time spent for quality checks. We also exploit the segmented workflow of text processing in a way that allows us to diversify our performance markers for different levels for optimizing completeness and validity of the whole

---

3 See our evaluation of four keyword lists, which are reported by Huang et al. [2016], Wang et al. [2016b], Weidmann and Rød [2019], and Kriesi et al. [2019] below in the results section.
process. That is, we can favor higher recall performance on the document and sentence levels to maximize completeness, and focus on improving precision at the token level to boost validity. What is more, the human-annotated GSC compiled from unfiltered random samples in this bottom-up and interlinked work process allows us to measure the markers of completeness and validity transparently and consistently. Human annotation and automated tool performances are constantly compared and checked against one another and all data in the GSC are optimized in multiple steps of checks, error corrections and fine-tuning.

**The Methodology of the Article: A Systematic Comparison of Random Sampling and Keyword Filtering**

In the following, we will lay out our experimental procedure in which we calculated respective performance markers of automated text classification models that were trained using randomly sampled data and data that were filtered by use of keyword lists of four projects that use a similar protest event ontology to ours. We compare the performances of our random sample model with the four above mentioned keyword filtered models, and with an additional keyword filtered model trained on the basis of the aggregation of all four keyword lists. All models will be trained and tested on the same raw data pool, the content of which is laid out in Table 1.

**Table 1. Descriptive statistics for of each dataset of news sources.**

<table>
<thead>
<tr>
<th>Source</th>
<th>Total</th>
<th>Protest</th>
<th>Not-protest</th>
<th>Country</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIE</td>
<td>701</td>
<td>71</td>
<td>630</td>
<td>India</td>
<td>.11</td>
<td>1.0</td>
<td>.19</td>
</tr>
<tr>
<td>PD</td>
<td>801</td>
<td>69</td>
<td>732</td>
<td>China</td>
<td>.09</td>
<td>1.0</td>
<td>.16</td>
</tr>
<tr>
<td>SCMP1</td>
<td>1002</td>
<td>17</td>
<td>985</td>
<td>China</td>
<td>.02</td>
<td>1.0</td>
<td>.04</td>
</tr>
<tr>
<td>SCMP2</td>
<td>2404</td>
<td>121</td>
<td>2283</td>
<td>China</td>
<td>.05</td>
<td>1.0</td>
<td>.09</td>
</tr>
<tr>
<td>ANA</td>
<td>1010</td>
<td>107</td>
<td>943</td>
<td>South Africa</td>
<td>.10</td>
<td>1.0</td>
<td>.18</td>
</tr>
<tr>
<td>Total</td>
<td>5922</td>
<td>385</td>
<td>5,573</td>
<td>-</td>
<td>.06</td>
<td>1.0</td>
<td>.11</td>
</tr>
</tbody>
</table>

These samples were drawn randomly from the complete archives of the aforementioned four news sources in India, China, and South Africa: New Indian Express (NIE) from India, People’s Daily (PD) and South China Morning Post (SCMP) from China; and African News Agency (ANA) in South Africa. Each archive consists of around a million articles and the periods covered by these archives start in 1998 and end in 2019. SCMP1 was randomly drawn only from all articles published in 1998, while SCMP2 was drawn as a separate random sample from stories between January 1, 1999 and December 31, 2019.

All documents have been coded by our annotators as to whether or not they contain at least one protest event mention that has definitely taken place in the past, or is ongoing. We consider our annotations as the ground truth upon which the performance results in experimental steps are based, as our double annotation and adjudication process leaves no false negatives. In Table 1 this is indicated by the recall value that is equal “1.0”. Precision is the ratio of articles mentioning protest relative to the full archive. F1 is an overall test statistic that integrates precision and recall.4

---

4 The F1 score is calculated using the formula “(2 * Precision * Recall) / (Precision + Recall)” and represents the performance of an information retrieval or text classification system. F1 is a single metric that strikes a balance between precision and recall. We always report the macro average of the precision, recall, and F1 for the classes,
From this data pool, every sampling model will compile the training documents according to its method. Random sample method does not filter data so the entire data pool will serve as training data. For keyword filter models, on the other hand, the data pool will be filtered to compile the documents that will serve as training data. In order to replicate an actual semi-supervised machine learning model training, we assume that every set of documents compiled as described above, are human coded to determine protest event containing articles. Our existing annotations will thus serve as the GSC creation for each training model that we will test.

In the next section, we will present our results which concludes that automated text classification models that are based on a GSC compiled from a completely random (i.e. unfiltered) sample of news articles performs much better than keyword filtering in a cross-country context. In the first step, we evaluate the performance of mere keyword lists in detecting true positive documents (i.e. ones that contain protest events) (Table 2). The following three steps compare the performance of the ML models that are trained with keyword filtered data and random sampled data. In step two, we use different keyword lists to filter both training and test data (Table 3). This is the evaluation methodology, which we designate as an optimistic scenario, applied in recent work (Zhang and Pan, 2019; Hanna, 2017). This shows that while traditional ML models trained on keyword filtered data yield higher scores than models trained on random sample, the random sample performs manages to equally well in case deep learning is used in this evaluation setting. In step three, we repeat the same exercise by filtering the training data like in step two, but without filtering the test data using any keyword list (Table 4). We conducted Step 2 and Step 3 in order to show that automated text processing tools will perform well particularly if they are tested on data that have the same biases as their training data (i.e. when both sets are filtered with keywords). The comparison between Step 2 and Step 3 demonstrates the performance of the tools developed using various filtering regimes in a realistic setting, i.e., non-restricted test data, which shows that the filtered training data yields lower scores with non-filtered test data and the random sampling method still yields comparable results to the best performing keyword filtered training set when state-of-the-art deep learning methods are employed. In step four, we show that the higher performance of the random sample method is not due to its larger size (Tables 5a and 5b). In step five, we find that the random sample significantly outperforms the best filtering method in a cross-country training-test setting (Table 6 and Table 7). Overall, our findings show that a random sample-based method is either comparable or better when researchers use the state-of-the-art ML methods to create either single context or cross-context measures.

Results

Our analysis in the first step compares the performance of our random-sample GSC with existing studies that use a keyword filtering approach. We use four major studies whose keyword lists and applications are publicly available: PolDem (Protest Event Dataset 30) by Kriesi et al. (2019); the Mass Mobilization in Autocracies Database (MMAD) by Weidman (2019); the EventStatus Dataset by Huang et al. (2016); and the meta-analysis by Wang et al. (2016b). The detailed list of the keywords is provided in APPENDIX A. The precision, recall, and F1 values are calculated on the presumption that contentious political event definitions of these studies basically overlap with our definitions. Appendix A shows the extent of this overlap. We refer to those which are protest and non-protest in our case, in our dataset. The macro average is not affected from different ratios of the classes in the evaluation dataset. Note that macro averaging of precision and recall may cause the F1-macro scores not to remain between precision and recall.
randomly sampled batches as “non-filtered data” in the rest of the article and we designate a dataset as “random” if it is not filtered using keywords.

Table 2 shows the extent to which each keyword list manages to identify the relevant articles annotated in each randomly sampled batch in our dataset. The table presents precision (P), recall (R), and F1-macro scores for each keyword list. This shows that the keyword list of Wang et al. identifies 72% of the relevant documents identified by the annotations from the random sample from the New Indian Express (NIE) (R=0.72). Similarly, this keyword list identifies 62% of the positive cases identified by our annotation (P=0.62). Precision and recall for the various news sources vary between .53 and .75 and between .57 and .84 respectively. Merging all keywords together yields .57 precision and .78 recall on all batches. In other words, the attempt to maximize recall by combining all the keyword lists only yielded a .78 recall score, meaning that 22% of documents that were found by our annotators as containing a protest event were missed by this keyword filtering. Optimizing data selection for a limited number of sources may cause the coverage to be affected significantly on sources not included in the optimization. Moreover, this effort gave poor precision scores, which would require a substantial amount of human supervision to correct, therefore defeating much of the purpose of keyword filtering. Also, each keyword list, including the aggregation of keywords, performs quite differently across the sources, indicating the difficulty of coming up with keywords that will perform consistently across contexts.

Table 2. Evaluation scores for keyword lists on random batches for India, China and South Africa

<table>
<thead>
<tr>
<th>Source</th>
<th>Wang et al.</th>
<th>PolDem</th>
<th>MMAD</th>
<th>EventStatus</th>
<th>All keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>NIE</td>
<td>.62</td>
<td>.72</td>
<td>.64</td>
<td>.59</td>
<td>.74</td>
</tr>
<tr>
<td>PD</td>
<td>.59</td>
<td>.69</td>
<td>.61</td>
<td>.63</td>
<td>.83</td>
</tr>
<tr>
<td>SCMP1</td>
<td>.53</td>
<td>.77</td>
<td>.51</td>
<td>.53</td>
<td>.82</td>
</tr>
<tr>
<td>SCMP2</td>
<td>.59</td>
<td>.84</td>
<td>.60</td>
<td>.57</td>
<td>.84</td>
</tr>
<tr>
<td>ANA</td>
<td>.66</td>
<td>.84</td>
<td>.68</td>
<td>.62</td>
<td>.82</td>
</tr>
<tr>
<td>All sources combined</td>
<td>.60</td>
<td>.79</td>
<td>.62</td>
<td>.59</td>
<td>.81</td>
</tr>
</tbody>
</table>

In the next steps, we compare different ML performances that are obtained in models that using keyword filtered training and test data from the raw data pool to our random sample based model. This requires training and testing different ML models with subsets of the data derived from our raw data pool with the respective keyword lists. Our procedure consists of randomizing the data, optimizing parameters, and predicting the held-out subset for the test. We follow a k-fold cross-validation (k=5) procedure, in which scores are the averages of executing the same processes for k times, as the dataset is split into k subsets in a stratified manner. At each iteration, k-1 subsets are used for training to create a model and 1 subset is used for evaluating it. The 10% of the training data, which is extracted in a stratified manner, is used as development data during the training. All of the data points receive a prediction in this manner. Then, we calculate the performance scores
on these predictions using the correct label.\(^5\) We use both traditional and state-of-the-art ML approaches. The former consists of Support Vector Machines (SVM), Random Forest (RF), Naive Bayes (NB), and Recall Biased Voting (RBV).\(^6\) The state-of-the-art model is the BERT (Devlin et al. 2019).\(^7\) All these models are binary classifiers that learn feature weights in each class (i.e. protest vs. non-protest) of the documents in the training set using dedicated feature extraction and learning algorithms. The weights enable the prediction of the classification of a document in the test data.

Table 3 presents the performance scores of the keyword lists when both the training and test data are filtered using the respective keyword lists. Across different filtering methods, we observe significant performance differences. In general, keyword lists that restrict the data more strictly create a higher F1-macro score. In particular, EventStatus’s list yields the highest scores among the traditional ML methods. However, this list shows the lowest performance when we use the state-of-the-art ML method (BERT). The keyword list proposed by Kriesi et al. (2019) performs the best when we use BERT. Moreover, the table suggests that the progress in ML methodology affects the performance significantly. BERT clearly outperforms all traditional ML approaches. Table 3 shows that performance scores can be misleading under restricted conditions, i.e. where test data are keyword filtered.

**Table 3. Performance of keyword lists on filtered training data and filtered test data for various ML models**

<table>
<thead>
<tr>
<th>Sampling method</th>
<th>Support Machine</th>
<th>Vector Naive-Bayes</th>
<th>Random Forest</th>
<th>Recall Voting</th>
<th>Biased</th>
<th>BERT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>Wang et al.</td>
<td>.79</td>
<td>.69</td>
<td>.72</td>
<td>.72</td>
<td>.73</td>
<td>.73</td>
</tr>
<tr>
<td>PolDem</td>
<td>.82</td>
<td>.68</td>
<td>.72</td>
<td>.71</td>
<td>.73</td>
<td>.72</td>
</tr>
<tr>
<td>MMAD</td>
<td>.76</td>
<td>.67</td>
<td>.69</td>
<td>.77</td>
<td>.68</td>
<td>.70</td>
</tr>
<tr>
<td>EventStatus</td>
<td>.73</td>
<td>.72</td>
<td>.73</td>
<td>.73</td>
<td>.71</td>
<td>.71</td>
</tr>
<tr>
<td>All keywords</td>
<td>.83</td>
<td>.68</td>
<td>.73</td>
<td>.68</td>
<td>.74</td>
<td>.70</td>
</tr>
<tr>
<td>Random</td>
<td>.88</td>
<td>.65</td>
<td>.71</td>
<td>.61</td>
<td>.77</td>
<td>.64</td>
</tr>
</tbody>
</table>

In the third step, we compare the performances of different keyword lists and different ML models when the ML models use keyword-filtered training data and non-filtered test data (Table 4). This setting yields lower performance scores than in Table 3. Traditional ML approaches are

---

5 The code is available on https://github.com/emerging-welfare/random_sampling-vs-keyword_filtering.

6 We used Scikit-learn version 0.23.2 (Pedregosa et al., 2011). The software is available on https://scikit-learn.org/0.23/ by October 29, 2020. The SVM, RF, and NB are the SVC(C=10, gamma=1, kernel="rbf"), RandomForestClassifier(n_estimators=25, max_depth=35) and ComplementNB(norm=False). The Recall Biased Voting is our implementation that predicts positive in case at least one of the SVM, RF, and NB predicts a document as positive. The hyperparameters provided with the ML objects are determined using grid search on a wide spectrum of possible values. Bag-of-words and TF-IDF are applied using CountVectorizer(max_features=2000) and TfidfTransformer(<default parameters>) from Scikit-learn as feature extraction and selection methods.

7 BERT-base-uncased was fine-tuned using AdamW optimizer using learning rate 2e-5. The best model in 15 epochs is used. The document length is 512.
affected more than BERT. This difference between Table 3 and Table 4 indicates that keyword-based settings are bound to perform worse in test data derived from a random sample, which approximates the actual news archive to a greater extent. Therefore, the performance scores of keyword-based systems should be conceived as limited to restricted settings, which could be regarded as optimistic depending on the ML paradigm.

Besides, as we observe from Table 2, higher F1-macro scores tend to be driven by higher recall scores and this correlation improves when we move from traditional ML to the state-of-the-art BERT. The relatively better performance of the non-filtered (random) and keyword list of Kriesi et al. (2019), 0.84 (Table 4), which are based on recall values of 1.0 (Table 1) and .81 (Table 2) respectively, proves this trend. Finally, we see that the random training data (i.e. non-filtered) yields comparable performance when BERT is used. The random data performs worse than the filtered data in cases in which we use traditional ML models due to the greater imbalance between positive and negative cases in random data than in the filtered data. We find that the random data is in a position to benefit the most from approaches that oversample the positive class for handling imbalanced data (Branco et al. 2016).

In the fourth step, we investigate the effect of additional information on performance. Does having more information improve the scores in comparison to filtered batches? Table 5a illustrates the effect of the data size on the random sample and the Kriesi et al. (2019) keyword filtered versions of the data. We take half or one-third of the random dataset, which makes the data contain much less positively labeled data than the harshest filtering method. Performance remains around .82 F1-macro. The use of one fifth of the data as training data decreases the performance scores slightly. Nevertheless, these scores are comparable with the best performing filtered version. Hence, the random sample performs comparably well with the keyword list even with reduced data size. These results show that the outperformance of the random sample is not due to the larger data size. As a side result, we also observe that the filtering suggested by Kriesi et al. (2019) is significantly affected by decreasing the size of the data set, which indicates that the random sample method is

<table>
<thead>
<tr>
<th>Sampling method</th>
<th>Support Vector Machine</th>
<th>Naive-Bayes</th>
<th>Random Forest</th>
<th>Recall Biased Voting</th>
<th>BERT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Wang KL</td>
<td>.75</td>
<td>.68</td>
<td>.71</td>
<td>.62</td>
<td>.73</td>
</tr>
<tr>
<td>PolDem KL</td>
<td>.82</td>
<td>.68</td>
<td>.73</td>
<td>.63</td>
<td>.74</td>
</tr>
<tr>
<td>MMAD KL</td>
<td>.67</td>
<td>.71</td>
<td>.69</td>
<td>.62</td>
<td>.69</td>
</tr>
<tr>
<td>EventStatus</td>
<td>.61</td>
<td>.72</td>
<td>.63</td>
<td>.58</td>
<td>.69</td>
</tr>
<tr>
<td>All Keywords</td>
<td>.82</td>
<td>.66</td>
<td>.71</td>
<td>.64</td>
<td>.75</td>
</tr>
<tr>
<td>Random</td>
<td>.88</td>
<td>.65</td>
<td>.71</td>
<td>.61</td>
<td>.77</td>
</tr>
</tbody>
</table>
much more immune to data size reductions than the keyword method. Although the amount of the data that should be annotated in the random sampled data is more than in the case of keyword filtering, we observed that the random sample yielded higher performance as the size of the sample increases in Table 5b. However, the performance gains as the data size increases stops at some point for the Kriesi et al. (2019) keyword set. The column names in Table 5b denote units of increase in size starting from 500 instances and increases by 125 for Kriesi et al. (2019) and by 500 for the random sample. The data sizes in Tables 5a and 5b do not match the total counts in the datasets due to splitting of the data for K-fold-cross-validation.

**Table 5a. Performance on Random and Filtered Data Across Different Sizes of Training Data.**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Size</td>
<td>Precision</td>
</tr>
<tr>
<td>1</td>
<td>5000</td>
<td>.83</td>
</tr>
<tr>
<td>½</td>
<td>2,400</td>
<td>.84</td>
</tr>
<tr>
<td>⅓</td>
<td>1800</td>
<td>.81</td>
</tr>
<tr>
<td>¼</td>
<td>1,200</td>
<td>.79</td>
</tr>
<tr>
<td>⅕</td>
<td>1125</td>
<td>.78</td>
</tr>
</tbody>
</table>

**Table 5b. Saturation of the Performance in Relation to the Data Size**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>.6754</td>
<td>.7581</td>
<td>.8054</td>
<td>.8067</td>
<td>.8027</td>
<td>.8450</td>
<td>.8401</td>
<td>.8543</td>
</tr>
</tbody>
</table>

Tables 4, 5a and 5b show that our random sample methodology is the best performing approach in a random evaluation setting from all country sources. In the next step, we continue to compare performances of random and keyword based models in a cross-context setting. We do this by comparing our model with the keywords based model of the best performing keyword list, i.e. Kriesi et al (2019). As seen in Table 4, in the previous step, Kriesi et al (2019) and our model yielded 0.84 F1-macro using the BERT model.

In the fifth step, we perform a cross-context evaluation of the random sampling and keyword filtering methods using three different sources of data from India, China, and South Africa by treating all batches from a particular country as a single batch. In our evaluation, we train our model on every combination of these resources, leaving one or two resources out for testing. As can be seen in Tables 6 and 7 below, every column represents a different resource that is used as a test set. We continue using the best performing ML approach, which is BERT. In Table 6 we first filter the training data using the keyword list provided by Kriesi et al (2019). Table 7 is based on random
training data. The evaluation data for both tables, which is held-out from the training data and used to generate these scores, is from the random set for each country.

**Table 6. The performance of the Kriesi keyword list in a cross-context setting.**

<table>
<thead>
<tr>
<th>Training Set</th>
<th>India</th>
<th>China</th>
<th>South Africa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>India</td>
<td>-</td>
<td>.66</td>
<td>.56</td>
</tr>
<tr>
<td>China</td>
<td>.73</td>
<td>.89</td>
<td>.77</td>
</tr>
<tr>
<td>South Africa</td>
<td>.79</td>
<td>.72</td>
<td>.75</td>
</tr>
<tr>
<td>India + China</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>India + South Africa</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>China + South Africa</td>
<td>.78</td>
<td>.86</td>
<td>.81</td>
</tr>
</tbody>
</table>

The cross-context performance of the random training data, as presented in Table 7 is significantly better than the filtered training data, as presented in Table 6. The random sample always performs better than the filtered version. The filtered data performs slightly better when data from China is used for training and data from South Africa is used for testing. The significant differences such as .58 vs .76 (India-China) and .68 vs .82 (India-South Africa) F1-macro scores between the filtered and random training data versions in Tables 6 and 7 demonstrate the superiority of the random sample as training data.

**Table 7 The performance of the random training data in a cross-context setting**

<table>
<thead>
<tr>
<th>Training Set</th>
<th>India</th>
<th>China</th>
<th>South Africa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>India</td>
<td>-</td>
<td>.83</td>
<td>.72</td>
</tr>
<tr>
<td>China</td>
<td>.77</td>
<td>88</td>
<td>.81</td>
</tr>
<tr>
<td>South Africa</td>
<td>.84</td>
<td>.80</td>
<td>.82</td>
</tr>
<tr>
<td>India + China</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>India + South Africa</td>
<td>-</td>
<td>.84</td>
<td>.80</td>
</tr>
<tr>
<td>China + South Africa</td>
<td>.83</td>
<td>.83</td>
<td>.83</td>
</tr>
</tbody>
</table>

Finally, we performed a qualitative analysis of false negatives of the model that was trained with the aggregated keyword list, finding a systematic bias against kidnapping events alongside
misses of extreme or rare cases. None of the keyword lists contain words associated with kidnappings (kidnap, abduct, hostage etc.), which leads to this outcome.8

Part of the rationale of using keyword filtering in developing automated document classifiers is to decrease data imbalance that would exist in a random samples by oversampling positive cases, in our case documents that contain protest event information. Traditional machine learning models yield lower performance with imbalanced training data, when protest event articles, for instance, comprise 5% of the sample. Advanced deep learning models like BERT remedy this problem in ML, as they are better able to handle low frequency of positive cases, thus making up for the difficulties of working with unfiltered raw data. This convenience provides the opportunity to utilize random samples, and thereby, achieve higher levels of data validity through increased representation of the actual data universe. Our results show that a random sample of news articles, as compared to keyword filtering, constitutes a much more reliable source as the training and test dataset for ML systems that are commonly used for automated protest event collection efforts. This random sample method is superior, especially when state-of-the-art deep learning methods are utilized both for in-context and cross-context evaluation settings.

**Ethical Dimensions and Data Sharing**

As in all studies of political contention and social movements, protest database building demands special attention to ethical dimensions of research because of the risks associated with political dissent. Compiling information about the contentious activity may disclose some key actors and dynamics of political dissent, rendering activists or insubordinate populations vulnerable to further surveillance and repression (Milan 2014, Gillan and Pickerill 2016). This is the case even in protest databases which are created by retrieving information from publicly available news sources (newspapers or news agencies) because these databases compile information about many actors, events and other features of events in a systematic fashion, making protest trends visible and allowing predictions of future protest activity. This risk, which also applies to projects using manual methods, is aggravated by automated methods which allow for the processing of large amounts of information quickly and enable the production of near real-time estimates of contentious event activity from online news sources. In developing new methodological approaches that would significantly improve the performance of automated protest event classification and extraction tools, we take these risks seriously and strive to develop strategies to mitigate these risks and address ethical concerns about social movements research as much as possible.

These questions of ethics and risks to social movement actors inform all stages of our research, including research design, data collection, analysis, and dissemination. The first question we address in this regard is about the relevance of our research to research subjects (social movements, activists and precarious communities that engage in contentious activities) as well as the scientific community. The GLOCON Project is part of a larger research project, Emerging Welfare, which investigates the causal relations between grassroots political activity and welfare state expansion in emerging market countries. We collect and analyze protest data from our case countries (Argentina, Brazil, China, India, South Africa, and Turkey) for the purpose of revealing dynamics that inform social policy-making processes and assessing the impact of grassroots

---

8 We performed a similar analysis of false negatives generated by the model trained with data filtered by the Kriesi keyword list in the part of our corpus that we could not incorporate into the experiments here, due to the difference of their initial sampling method which is potentially distinct from the definition we adopt in this article. In this analysis we found a similar systematic bias against dharna events (a form of sit-in specific to India) in one of our sources from India, which unlike New Indian Express, somehow writes the articles in a way which excludes the otherwise included keywords such as protest, etc.
activism and/or organized politics on government policies, specifically the provision of socio-economic entitlements. This is to test our argument that most governments use and expand social welfare programs in their efforts to contain social unrest and establish political stability. In this regard, our research produces data and knowledge of relevance to both scholars and social movements themselves, as it provides them with intellectual resources and data that they might not obtain otherwise about the impact of social movement and grassroots activities on social justice.

The second ethical concern we address in building the GLOCON Project is about the protection of research subjects in the processes of both data collection, analysis, and dissemination. As we discussed above, automated protest data collection may pose significant risks for research subjects. The methodology we develop in the GLOCON project allows us to compile fine-grained protest event information (including organizers, participants, place, facility names, time, and event type) more accurately than other automated event databases and more easily and cost-effectively than manual methods. Therefore, it requires us to take further precautions to prevent access to and use of these data by malicious actors who may target resistant groups or individuals.

As a first step to protect the identity and activities of research objects (be it specific populations, groups, organizations or individuals), we store our GSC, automated tools and protest datasets in encrypted hard drives at our institution. Secondly, in data processing and analysis, including the creation of ontologies, semantic categories as well as the production of social science articles as part of the Emerging Welfare Project, we refrain from disclosing the identity of any specific individual and from using any derogatory, stigmatizing or criminalizing language (such as “terrorist”, “bandit”, etc.) in our categorization of contentious activity or actors. Instead, we use value-neutral and factual terms in categorizing protest participants, organizers or event type data (e.g. “militant”, “armed organization”, “armed militancy”).

In addition, we strive to develop specific protocols of data sharing to minimize the risk of the use of our data for malicious purposes against activists or resistant groups. Our data sharing policy is shaped by our belief in the power of scientific collaboration and research transparency for theory and method development as much as our concern about the well-being of precarious populations, groups and/or individuals that are engaged in political activities for social change. In order to balance these sometimes-conflicting priorities, we have embraced a two-tier data sharing method. At the first layer, we share a processed and limited version of our data in a visualized form on an open-access website of the GLOCON Project (glocon.ku.edu.tr). The protest data visualized on country maps on this website provide only macro-level information about the main protest categories, such as year, event-place (province or city), event-type (five major aggregated event-type semantic categories we have produced: demonstrations, industrial actions, group clashes, armed militancy, electoral politics), urban/rural locations, ethnicity, and ideology. At the second layer, we share our detailed protest datasets (including detailed event, organizer, participant, place, facility and time information) and our GSC and computational tools only with researchers and parties (including social movements themselves) who demonstrate compliance with ethical standards of social movements research and norms for the protection of research subjects. For instance, we do not share our detailed data with government or law enforcement agencies or researchers who collaborate or receive funding from intelligence or defense agencies.

In addition to protecting research subjects, we also develop strategies of sharing our corpus without violating the copyrights of news sources that we use in developing our tools and building our protest database – an unresolved issue in the automated protest building business addressed by Wang et al. (2016a). For instance, in the shared tasks we organized, we engineered methods of sharing part of our GSC with other participants to run their tools without causing any copyright
violation. This involved only sharing information that is needed to reproduce the shared part of
the corpus from the source for classification tasks (sharing only the news article URLs) and only
relevant snippets from annotated news articles needed for information extraction tasks. For the
latter, we released a Docker image that contains the toolbox required to reproduce the news articles
on the computer of a participant. The toolbox generates a log of the process that reproduces the data
set and we have requested these log files from the participants. The toolbox is a pipeline that scrapes
HTMMLs, converts HTMLs to text and finally performs specific filling operations for each of the
tasks. To the best of our knowledge, the toolbox has allowed participants to create the shared portion
of the dataset on their computers.

Conclusions

Do automated text processing tools match human coding in terms of validity, accuracy,
completeness, and reliability when faced with large varieties of event types, actors, and targets
across different socio-political contexts? This article advances a new bottom-up approach by using
random sampling and deep learning models. Our training dataset for ML is a gold standard corpus
based on random sampling, double annotation, and multiple context evaluation. We use unfiltered
random samples of national and international news articles from three countries (India, China, S.
Africa) and we show that this random sampling method has much better performance than keyword
filtering in achieving cross-country reliability and generalizability.

Our experiments yielded results that are in line with our argument that in creating a GSC
for training of ML models, initial selection of relevant documents by keyword filtering introduces
bias that distorts the training data. This distortion is exacerbated in the multi-context, multi-country
settings that are typical in comparative research. Most likely this is because fixed keyword lists are
not context sensitive, that is, they are biased in different ways in different countries. This is shown
by our primary finding that keyword lists perform differently in different countries (Table 2).
Models we trained with unfiltered, full-text annotated data, on the other hand, outperformed filtered
data models significantly in cross-context models. This means that automated text classification
efforts for large multi-context, multi-country settings should rely on random sampling to build their
GSC data. The lower performance of keyword filtered data likely stems from the fact that the biases
introduced by predetermined keywords distorts the data in different ways, which makes comparing
different countries in terms of the same event ontology even more difficult. Our overall workflow,
which integrates context variability into training by allocating training and development data to
different countries, is better served by randomly sampled data, compiled through full-text reads by
trained annotators who apply an event ontology which is conceptually well circumscribed yet open-
ended in the sense that it is not restricted by fixed terminologies and generic event identifiers.

Compiling training and development data through keyword filtering is not a design choice
that is valued in itself but rather a procedure driven by concerns over resource efficiency and data
access. Our results also show that application of novel deep learning models, such as BERT,
diminishes resource restriction problems significantly as, unlike traditional ML models, it handles
data imbalance and sparsity much better. The efficiency gains of oversampling relevant documents
via keyword filtering becomes much less pronounced as full-text reading of comparable amounts
of documents yields comparable recall and precision performance. Thus, with use of advanced deep
learning models, the need to sacrifice data validity for gains in efficiency diminishes greatly. As to
data availability, the widespread on-line presence of news sources from an ever growing number of
countries offers new opportunities to gather raw data without the necessity of using search platforms
that require keyword searches. In any case, we believe that access to unfiltered data is essential in
securing data validity.
With this last point, it must be emphasized that using random samples offers the best possible method of approximating the universe of actual events (i.e. the entire set of contentious politics events described by news documents in this case) and in ways that can actually be measured. In the absence of validity tests of keyword lists performed on unfiltered data, like Strawn (2010) has provided, it is impossible to know what is left behind after the initial batch of potentially relevant documents are collected. In developing automated text processing models, this deprives subsequent phases of development of the possibility of testing on ground truth data and therefore to assess the validity of the final database vis-a-vis the actual events. This is illustrated in our experiment which showed that the models trained on keyword filtered data performed worse when tested on randomly sampled data. In the absence of unfiltered data, it will not be possible to detect, or quantify this error.

Our study is the first genuine recall and precision quantification analysis to compare keyword-filtered collections and a random sample. As such, the model we outline above not only promises higher reliability than many existing databases in such a multi-context setting, but it also presents a more realistic and transparent process of testing reliability. The latter is thanks to the bottom-up approach and our insistence on working with unfiltered data. The completeness and accuracy of protest event databases are of utmost significance for comparative social science studies. The challenges and complexity that social scientific research presents to the field of natural language processing and ML is a unique opportunity for breakthroughs in the latter. The cross-context generalizability of the automated systems is already a challenging task (Hürriyetoğlu et al. 2019b). Our results show that maintaining a gold standard corpus should be one of the most critical concerns for researchers. We also emphasize the centrality of the ethical concerns about the protection of subjects as well as practical ways of data sharing that would not violate copyrights. We aim our project to be exemplary in terms of close and persistent multidisciplinary collaboration to improve methods and expand research horizons.
References


Appendix: Keyword lists

<table>
<thead>
<tr>
<th>EventStatus</th>
<th>PolDem</th>
<th>MMAD</th>
<th>Wang</th>
</tr>
</thead>
<tbody>
<tr>
<td>protest</td>
<td>initiative referendum petition* signature* campaign* protest* demonstrat* manifesto* marche* marchi* parade rall* picket* human chain riot* affray festival ceremony street theatre road show vigil strike* boycott* block* sit-in squat* mutin* bomb* firebomb* molotov graffiti assault attack arson incendiari* fire.* raising set.* ablaze landmine sabot* hostage* assassinat* shot</td>
<td>protest* demonstrat* rall* campaign riot* picket*</td>
<td>demonstrators civil unrest civil disturbance civil disturbances civil disruption civil disruptions civil disorder civil strife civil disobedience mob mobs protesters tear gas strikers marchers parade parades march marches marched marching protest demonstration demonstrations protests protested protesting unrest disturbance disruption disrupted disorder strike strikes striking strife blockade blockade blockades blockaded rally rallies sit-in</td>
</tr>
<tr>
<td>murdered, killed</td>
<td>sit-ins, clashed, clashes, attacks, conflict, barricade, barricades, roadblock, roadblocks, demonstrate, demonstrated, demonstrating</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>