Towards a Dataset of Automatically Coded Protest Events from English-language Newswire Documents

Jasmine Lorenzini, Peter Makarov, Hanspeter Kriesi, and Bruno Wueest

Paper presented at the Amsterdam Text Analysis Conference, June 2016

Towards a Dataset of Automatically Coded Protest Events from English-language Newswire Documents

Abstract

Advances in natural language processing (NLP) allow social scientists to design projects that require extensive coding of textual data. In this paper, we present the design of a protest event dataset that we construct by semi-automated coding of English-language newswire documents. We perform automated identification of relevant stories. The coding of protest events is carried out manually. The resulting dataset covers 30 European countries and spans over 10 years. First, we discuss our data collection methodology. We highlight the main difficulties related to the identification of relevant documents, which is challenging for humans, let alone machines. Second, we introduce the tools that we have developed using NLP technology. The use of newswire text and semi-automated techniques introduces bias. We conclude with an evaluation of our protest event data by comparing them to datasets built fully or semi-automatically, as well as manually coded protest event data for single countries. This paper contributes to advancing interdisciplinary research and presents a collaborative effort between computational linguists and social scientists.

Keywords

Protest event; Automated coding; Methods.

Towards a Dataset of Automatically Coded Protest Events from English-language Newswire Documents

In the social sciences, NLP technology has been used to code newspaper articles, tweets, and other on-line textual data. The automated coding of text is promising as it has the potential to increase the amount of processed documents and reduce the costs in terms of time and human work. The development of automated procedures for text analysis for the social sciences requires collaboration between social scientists and computational linguists. This is a challenge that involves a transfer of conceptual and technical knowledge and the development of common standards of data quality assessment. The evaluation of the data produced is critical in these automated approaches (Grimmer and Stewart 2013). Yet, these evaluations are difficult to design to respond to the needs of both social scientists and computational linguists.

In our project, we code protest events in newswire documents published in English. We take protest events to be all politically motivated unconventional actions performed by one or more individuals (Kriesi et al. 1995). Protest includes a broad set of actions by non-state actors expressing political grievances or claims. As we will see below, these three aspects of the definition – the variation in action forms, the kind of actor, and the political meaning attached to protest actions – make automating the task difficult. Many researchers have employed the manual coding of protest events, advancing the method and delivering valuable protest event datasets for multiple countries and years (see Hutter 2014 for a discussion of the history of protest event analysis (PEA)). Yet, the manual coding of protest events does not scale up to large quantities of textual data. In our project, we work with 30 countries: EU-27 countries plus three non-EU members (Iceland, Norway, and Switzerland). We study protest activity over a period of 15 years -- before and after the fall of Lehman Brothers in September 2008. We use multiple English-language newswire agencies as sources. In order to advance on this ambitious task, we have naturally sought to automate the coding of protest events. Our contribution is two-fold: 1) We are through with the construction of a dataset of protest events across Europe in the years 2000-2015, which

we will make publicly available in the future, and 2) We report on our work toward the automated coding of protest event data from news documents. The fully automated coding of events has proven challenging and presently lies well beyond our reach. We have implemented a semi-automated coding approach which we shall present in this paper.

The turn toward big data – which opens the possibility to look at entire populations rather than samples of the populations (Dalton 2016) – brings in new perspectives on the context of individual actions. The opportunity to look at protest across many European countries means that one can now study cross-national variation at the event level, following the cross-national turn in survey data. This opens the way to cross-national research designs that combine individual and event-level data (Dalton 2016). Another opportunity relates to the coding of multiple sources. Jenkins and Maher (2016) urge scholars to build multi-source datasets to address the problem of selection bias. Coding multiple sources allows comparing the news that make it into different sources and to assess biases related to specific news sources. Biases related to the sources selection are a central critique addressed to protest events data (see Earl et al. 2004, Ortiz et al. 2005 for reviews)

Our paper is structured as follows. First, we review the literature on automated protest event analysis. Then, we discuss difficulties associated with the automated coding of protest. We emphasize the importance of conceptual work ahead of automation and the assessment of what is feasible to automate given technological limitations and the limited resources of a project such as time. Second, we discuss the tools that we have built in the automated part of dataset construction. Third, we take a closer look at our sources. The use of documents written in the same style and language allows for the efficient building of automated coding tools, but reduces the coverage of protest events. We compare our data with existing manually and automatically coded datasets highlighting differences and potential biases in our dataset. The assessment of biases is a key issue in the study of protest events (see e.g. Barranco and Wisler 1999, Biggs 2016, Earl et al. 2004, Ortiz et al. 2005). We conclude by considering avenues for further development of semi- and fully automated coding of protest events.

Automated protest event analysis

Previous attempts at automating the coding of protest events have been successful in coding a limited set of events (see Hanna 2014, Yonamine 2013 for literature reviews). Most of this research has been done in the field of peace studies where the focus is on inter- and intra-national military or armed conflict (Raleigh et al. 2010, Schrodt and Gerner 1994, Schrodt, Yonamine and Bagozzi 2013, Yonamine 2013). GDELT primarily addresses conflict and violence. Similarly, ACLED focuses on armed conflicts (Raleigh et al. 2010). Others have moved away from war and armed conflict to include in their analysis non-armed political conflict and cooperation (Shellman 2008). In this way, this line of research comes closer to our topic of interest. Still, the heavy focus is on conflict zones, and whatever tools have been developed to facilitate event coding they are not suitable for our work on European protest.

In the field of PEA, the researchers have used automated approaches to select (and even code) protest events. The most prominent examples of semi-automated procedures are a) the 'European Protest and Coercion Data' (EPCD) collected by Francisco et al. (Francisco 1996, Nam 2006, 2007, Reising 1998, 1999), b) Imig and Tarrow's (2001) study on European protest events, and c) Jenkins et al.'s (2012) project for a new edition of the Handbook for Social and Political Indicators. The first two projects use versions of the Kansas Event Data System (KEDS); the third employs for coding propitiatory VRA-Reader (King & Lowe 2003). For example, Imig and Tarrow automatically code the headlines of Reuters newswire documents to identify protest aimed at EU institutions and policies in 1984-1997. Unfortunately, these projects tend to have important shortcomings in terms of selection of sources and coding procedure (Imig 2001: 256).

In spite of the focus on different action forms in some of the automated event analysis projects, we share with them an interest in key variables associated with events: The date, the location, and the actors involved. In the field of peace studies, early research has had a very high degree of aggregation, focusing geographically on countries and temporally on years (Chojnacki et al. 2012). More recent work has moved toward a more refined geographical and temporal analysis, with event data often taking the form of "day-by-day coded accounts of who did what to whom" (Shellman 2008: 466). In the CAMEO

ontology, each event is a triplet composed of an event code, a source actor, and a target actor (Gerner et al. 2002).

Many widely used systems for automated event coding (e.g. TABARI (O'Brien, 2010) / PETRARCH, VRA-Reader) rely on coding rules generated by humans. Rule-based information extraction systems are known to have low recall; rules are difficult to write and so adapting a rule-based system to a new topic is challenging and time consuming. Statistical learning, on the other hand, provides a principled solution to these problems, which explains the popularity of statistical methods in NLP research. The proprietary system that generates the Integrated Crisis Early Warning System (ICEWS) event data (Boschee, Natarajan and Weischedel 2013) extensively employs sophisticated statistical NLP technology. This event coder has been shown to outperform its rule-based counterparts. Statistical learning has been recently applied to the coding of protest events (Nardulli, Althaus and Hayes 2015, Hanna 2014).

All automated coding approaches face three important challenges (Schrodt 2012). First, events of interest can be easily confused with other types of events reported in the news. For example, commemorations can be taken for protest demonstrations. In both cases, large crowds gather in public spaces with political actors such as parties or trade unions present. However, the political meaning differs: simply put, commemorations are not acts of contentious politics. Another example is sports news which feature a similar vocabulary to protest events (e.g. "attack", "strike"). Second, event coding is challenging for human coders (see also Ruggeri, Gizelis and Dorussen 2011). Low intercoder reliability has been reported for other complex coding tasks like the coding of party manifestos (Mikhaylov, Laver and Benoit 2012). Lastly, the sources used for the coding of protest events introduce bias. Regarding the problems, they relate to the selection of the news into the different types of media: National newspapers, local newspapers, newswire documents, etc. Digital sources often provide numerous versions of the same document (e.g. corresponding to updates), but also the same event is reported by multiple sources. Care should be taken to avoid over-estimation due to such duplicates.

Definition and coding of protest events

Protest events are non-institutional political actions. This understanding is both broad and rather vague. The breadth relates to the set of action forms that are considered as non-institutional, which could probably be addressed by exhaustively enumerating all relevant forms. What is more problematic is the room for interpretation as to whether an action is politically motivated. The seriousness of this problem depends on the action form, and it is the most acute for symbolic actions (van Deth 2012). Franzosi (1989) states bluntly that "a coder should only be an instrument of measurement". This is not possible as human coders interpret the text to identify protest events. Nonetheless, in order to move toward automated coding, we need to put more rigor into coding. A way to achieve this is by using concepts that are as narrow as possible (Koopmans and Rucht 2002). However, studies find inconsistent results with regards to this. Some find that these narrow codes are better for human coders and that they facilitate machine learning and coding (King and Lowe 2003), while others find that intercoder agreement is lowest for the fine-grained categories (Bosche et al. 2013). Our solution to this problem of fuzziness of our object of study is to propose categories of action forms that are more homogeneous: Petitions, Strikes, Demonstrations, Blockades and occupations, and Political violence. For each of them, we provide the coders with detailed descriptions of specific characteristics of the different phenomena that we want to study.

An additional difficulty relates to the differentiation of protest events from one another. Franzosi (1989: 276) defines an event as "a set of actions performed at a particular place and time by some actor(s) against or in favor of some other actor(s)." From this definition, it is clear that the time, location, actors, and their goals are inherent to the identification of protest events. Indeed, protesters often combine different forms of action to create disruption (Tarrow 1989, Tarrow 1999). Actions might take place around the same time in one or more locations. But such actions are not independent events, and some of them might be considered parts of other events – for instance, protesters may be chanting, cheering, booing, carrying banners, or handing out leaflets (McPhail and Schweingruber 1999). We take all these actions to be components of one single event. Therefore, we specify a) not only what does not constitute new protest events, but also b) what events we

consider as part of another event. The core rule is that a change in any of the event variables – action form, location, date, actors, or issues – signifies a new event. For instance, when part of a peaceful demonstration breaks off and turns violent, the change in action form signals that we should count two separate protest events. Similarly, a demonstration and a counter-demonstration sometimes take place on the same date and at the same location, but express different views on a given issue. Again, such events should be considered two different events.

Coders need to take two decisions: First, whether a news story reports about protest, and second, whether it talks about one or more events. In order to be consistent about these decisions, coders need detailed information about distinct protest forms and equally detailed rules on how to treat event characteristics such as time, location, actors, and issues. The difficulty in taking these decisions manifests itself in low inter-coder reliability and translates to challenges for automated coding.

The overall procedure

In a recent piece, Nardulli et al. (2015) advocate for the use of hybrid systems combining the strengths of machines and those of humans in coding a broad set of political events (see also Croicu and Weidmann 2015). They propose automating the extraction of simple factual content from text, and then employing humans to code the content that requires more interpretation. Our approach also combines automated methods and manual coding. We obtained newswire documents from the LexisNexis data service[1]. We followed the multi-source approach in the study of protest events (Jenkins and Maher 2016). For the production of the final dataset, we used 10 English-language newswire agencies[2] and retrieved documents published in the 10-year period from 2005 to 2014. The query

1 http://www.lexisnexis.com/

2 The following news agencies are included in our study: AFP, AP, APA, BBC,BNS, CTK, DPA, MTI, PA, and PAP. Our goal was to include not only the major news agencies (AFP, BBC, PA) but also some regional news agencies that cover eastern or southern Europe more in depth. Unfortunately, we were not able to gain access to Spanish international news agency EFE or any Greek news agencies, which would have been a great source for studying southern Europe.

comprises about 40 keywords and phrases that describe common protest action forms[3]. The query returned the initial set of 4,861,248 documents. The motivation to employ a little constrained query was to obtain as many relevant documents as possible, even at the expense of having to filter out a large number of irrelevant documents.

To go from this amount of documents down to a set of documents that could reasonably be coded manually, we performed the following steps: i) we removed duplicate documents, ii) we used LexisNexis location meta-data to filter out those documents that are clearly not associated with any of our countries of interest, iii) we built a supervised statistical document classifier that learns to distinguish relevant documents from irrelevant ones, iv) to further refine identification of relevant documents, we built a supervised statistical classifier that learns to predict mentions of protest events in the text, v) lastly, we performed manual coding of protest events.

De-duplication

By inspecting LexisNexis document identifiers, we found that only about 82 percent of the obtained documents are unique. Further, we applied our re-implementation[4] of the SpotSigs algorithm (Theobald, Siddharth and Paepcke 2008) that identifies near-duplicate documents. We found that about 6 percent of documents with a unique LexisNexis identifier were near-duplicates of some other document.

Location-based filtering

A vast majority of LexisNexis documents come with rich meta-data which includes information about locations associated with a given document. We filtered out documents that do not feature any of our countries of interest. In case a document is associated with

3 Query string "initiative OR referendum OR petition! OR signature! OR campaign! OR protest! OR demonstrat! OR manifest! OR marche! OR marchi! OR parade OR rall! OR picket! OR (human chain) OR riot! OR affray OR festival OR ceremony OR (street theatre) OR (road show) OR vigil OR strike! OR boycott! OR block! OR sit-in OR squat! OR mutin! OR bomb! OR firebomb! OR molotov OR graffiti OR assault OR attack OR arson OR incendiar! OR (fire I/1 raising) OR (set AND ablaze) OR landmine OR sabot! OR hostage! OR assassinat! OR shot OR murdered OR killed"

4 https://gitlab.cl.uzh.ch/rothenha/nearDuplicateDetection/

more than one relevant location, we assigned it to a single country using an intricate set of rules that prefer countries under-represented in the sample like Portugal, Belgium, or Iceland.

Document classification

We trained a document classifier to distinguish between relevant and irrelevant documents. To this end, we manually labeled a total of 7,447 documents in three coding rounds. The samples were not drawn i.i.d. from the initial document set, but at different stages of the filtering process. The reason for this is pragmatic. We estimated the number of relevant documents in the initial document set at under 5 percent. We over-sampled relevant documents in order to have more relevant documents to train on, ending up with 10 percent relevant documents in the labeled data.

We modeled document contents with a bag-of-words document model – a model that takes into account frequencies of words in a document but not word positions. We added features that signal that a document mentions a relevant location. To this end, we compiled lists of relevant and common irrelevant place names. We found that building a model only over sentences that mention a keyword did not impair prediction quality. We always included the title, byline, and lead sentences into the model. We applied standard *tf-idf* term count scaling, stopword removal, and stemming[5] (Sebastiani 2002).

We trained a logistic regression regularized with the elastic net penalty (Zou and Hastie 2005). We used the implementation of this algorithm from **sklearn**[6], a Python machine learning library. For evaluation, we split the labeled documents 80/20 into the training and test sets. We report precision / recall / F1-score of .52 / .66 / .59 on the test set following a 5-fold cross validation.

The inter-coder agreement for document classification is not high. We have never directly evaluated it – partly because the large number of irrelevant documents makes a direct evaluation prohibitive. On a task of evaluating positive predictions by the classifier (see *Tuning of classifier thresholds*), the inter-coder agreement measured by the pairwise

5 https://gitlab.cl.uzh.ch/makarov/polcon/
6 http://scikit-learn.org/

average F1-score (Hripcsak and Rothschild 2005) was around .84. This figure does not account for potential disagreement on relevant documents not retrieved by the classifier.

The document classifier correctly identifies 94 percent of irrelevant documents. This result should be regarded as a successful application of this technique.

Event mention detection

A bag-of-words document model cannot adequately address the nuances of our definition of relevant protest event. If the dateline says *London* and the body contains a phrase *a terrorist attack in Beirut*, from the perspective of the bag-of-words model both place names could equally likely name the location of the protest event. This clearly is not satisfactory. The problem manifests itself not only in locations – time, actors, and the factual vs hypothetical presentation of an event (e.g. *unions went on strike today* vs *unions threatened to go on strike*) are also common sources of error.

To address this problem, we used text-bound annotations of event mentions which we had collected as part of another line of research. In this project, coders are asked to *annotate*, that is highlight, parts of text that provide information about properties of protest events: location, actors, time, etc. We performed annotation in a browser-based annotation tool called **brat**[7] (Stenetorp et al. 2012). Figure 1 shows an excerpt annotated in **brat**.

<Figure 1>

We identified protest event mentions with action form annotations: The words annotated as an action form typically indicate the protest event most directly. In this way, we reduced the problem of detecting protest event mentions in documents to predicting, for a span of words, whether or not it should be annotated as an action form. Although coders also categorized annotated action forms, we did not use this information for filtering.

Initially, annotation was performed in a content-driven fashion with little constraints on what could and what could not be annotated. In the natural language processing literature, this approach is sometimes advocated for annotation tasks performed by domain

7 http://brat.nlplab.org/

experts rather than linguist annotators (Stubbs 2013). For action form annotations, this resulted in an inter-annotator agreement of .675, measured by the pairwise average F1-score. Two annotations were considered matching if they overlapped.

Due to the free nature of annotations and the rather low inter-annotator agreement, we did not approach the problem of action form prediction as a word sequence classification task. Instead, we decided to simplify the task further into prediction, for a common noun or verb, of whether it is the syntactic head of an action form annotation (e.g. *riots* in *violent riots* or *threw* in *threw stones*). In the information extraction community, such a word is called *event trigger* (ACE 2005). Our reasoning was that even when exact spans differed, all annotators would likely include the event trigger in the annotation.

We used the Stanford dependency parser (Manning et al. 2014) to automatically identify event triggers in the annotations and performed some manual correction of incorrectly identified triggers. We used 340 annotated documents for training and 78 documents for test. Only about 2.5 percent of common nouns or verbs were event triggers, which indicates a highly skewed classification problem.

In our model for event trigger prediction, we only used information from the sentence that a trigger candidate occurs in. We created the following groups of features[8]:

- the lemma, part of speech, stem of the word;
- similarly, for two notional words immediately to the left and to the right, also named entity tags of the context words and whether they occurred in the lists of relevant or irrelevant place names;
- the shortest dependency paths to animate nouns and place names, which models the link to protest actors and locations;
- a match with action form patterns like *take to streets, march through town*, which we obtained by semi-supervised leveraging of unlabeled documents in the spirit of (Huang and Riloff 2013);
- the bag-of-words representation of the sentence.

8 https://gitlab.cl.uzh.ch/makarov/polcon/

Lemmas, parts of speech, and dependency paths were obtained using the Stanford CoreNLP pipeline (Manning et al. 2014). Again, we used **sklearn** to train a logistic regression with the elastic net penalty and performed a 10-fold cross validation. We report precision / recall / F1-score of .40 / .71 / .51. An important use case for subsequent manual coding was prediction of sentences containing an event trigger. For this condition, we report precision / recall / F1-score of .53 / .74 / .62. The event trigger classifier was used for document filtering in the following way: If the classifier found no event trigger in a document, the document was filtered out.

We see a lot of potential in this approach and expect improvement from higher quality training data and a better modeling of the problem.

Tuning of classifier thresholds

We performed a manual evaluation of the predictions by the classifiers for a sample of 14,100 documents that passed de-duplication and location filtering. We re-trained the document classifier with the highest validation F2-score on the entire labeled data. The number of incorrectly classified irrelevant documents was around 50 percent – still too high for manual coding. We re-trained on the entire labeled set the event trigger classifier that had achieved the highest validation F0.5-score. We experimented with the joint prediction by the document and anchor classifiers, and examined threshold values at which the classifiers switch from predicting the negative class to predicting the positive class. We found the thresholds that would result in the best performance on the evaluation set. The combination of the classifiers had a higher precision and recall than the document classifier (Table 1).

<Table 1>

For the production of the document set for manual coding, we set the classifier thresholds at the values that resulted in the highest F1-score on the evaluation set: .61 for the document classifier and .85 for the event trigger classifier.

Manual coding of protest events

In the last step, we move to manual coding of protest event. The document classifier and the event trigger classifier are used to identify relevant documents for manual coding. In other words, they are used to set the boundaries of our universe – the newswires among which we draw our sample for manual coding. This allows reducing the amount of time dedicated to manual coding as coders only see a fraction of all documents retrieved with the use of keywords, those with the higher probability of reporting about protest events. The joint prediction by the document classifier and the event trigger classifier gives the optimal thresholds capturing the point where the document is most likely to report about protest (recall) and least likely to include false positives (precision). Once selected the document with a probability of reporting about protest events that lies above the threshold, we ask manual coders to read and code them using and on-line interface. The interface allows reading and coding the text in a single window. We code a limited set of variables: action form, date, location, actor(s), number of participants, and issue(s).

Events duplicates

A major issue related to the study of protest events through automated coding of multiple sources is the artificial multiplication of events, the duplicate problem (Jenkins and Maher 2016, Schrodt 2012, Ward et al. 2013). When the same event is reported by multiple news agencies, this can give a measure of the salience of an event – its visibility, but this does not mean the event should be counted multiple times. We refer to this part of the problem as the duplicated events. The duplication of events can also result from the fact that news agencies sometimes issue more than one newswire about the same event at different times of the day. This second issue relates to document duplicates or quasi-duplicates that we discussed above. This has a slightly different status and should not even be used to account for saliency as some agency might publish more duplicates and quasi-duplicates than others.

Other automated event analysis projects face the problem of duplicates. While GDELT identifies all the entries reporting about a single event and does not try to match them to solve the duplicated problem, ICEWS does. This appears as an important difference between the two sources (Ward et al. 2013). Thus, GDELT inflates certain

protest and one should be cautious when using it to consider this inflation when interpreting the results (Jenkins and Maher 2016). In our case, we are closer to the ICEWS approach and we aim at solving, or at least handling, the duplicated events problem. In order to address both issues of duplicate events and duplicate (or quasi-duplicate) documents, we used multiple techniques to address this problem.

First, we removed documents which are linguistically very similar to others. More precisely, we compared all documents with the SpotSigs-algorythm (Theobald et al. 2008). Subsequently, we identify news wires with a Jaccard-Coefficient of 0.75 or higher as near duplicates and accordingly keep only one of them in our corpus[9] as presented above. Second, on the event level we deal with the issue of duplicates at the stage of data cleaning. According to our definition, a protest event can be identified based on its action form, location, and date. So we use these three variables to match all entries referring to the same protest events in our dataset in order to keep only one entry for each of them. The distribution of documents is the following: 68 percent contain uniquely identified events while 14 percent contain only duplicates and 18 percent contain no event at all.

The evaluation of our dataset

At the moment, our dataset includes 13'400 protest events retrieved from multiple news agencies and covers 30 countries over a 10-year period. One difficulty in assessing protest event data relates to the fact that we do not know the universe of protest events (Jenkins and Maher 2016). In order to move around this problem, we devise two strategies to assess biases in our corpus. First, we explore the identification of protest events at various thresholds for three countries, namely Hungary, Spain, and the UK, in order to see how this affects the inclusion of various types of protest events across different contexts. As highlighted by Jenkins and Maher (2016), our goal should not be to aim for unbiased data but rather to be aware of the biases built in our data. Second, we use existing manually coded datasets as gold standards and we evaluate how good we are in capturing protest

[9] A brief evaluation of 50 duplicate pairs found no errors in this step.

trends compared to them. Additionally, we also compare our dataset to ICEWS – a high quality automated dataset.

Evaluation of threshold approach

The documents that are hand coded by our team of coders are selected through the automated techniques presented above. In particular, the probability that a document contains a mention of protest event assigned to a document determines whether the document will be included in the universe of documents from which we sample the documents for hand coding. So it is important for us to assess the types of biases that we may introduce with the use of such thresholds. In order to do so, we explore how irrelevant documents are distributed at the different thresholds. Furthermore, we evaluate variations across action forms in terms of how many we find at the different probability levels.

<Table 2>

First in table 2, we present the share of irrelevant documents found at the different probability levels for Hungary, Spain, and the UK. Among the 9'000 documents coded for these three countries, we see that irrelevant documents represent 25 percent and there is a clear divide in the share of irrelevant documents below and above the .61 threshold. Above the threshold the share of irrelevant documents ranges from 3 to 36 percent, whereas below the threshold it jumps up to 80 percent and more. Interestingly, we see that the distribution evolves with our attribution of probability scores – the higher the probability score, the smaller the share of irrelevant documents. This means that when the document-classier is more confident that the document contains protest, the document generally does. For instance, when the classifier is certain that the document contains a protest event (probability score of 1), we find only 3 percent of irrelevant documents. Yet, we find twelve times more at the threshold, the lowest probability score included in the universe from which we sample, at .61 we find 36 percent of irrelevant documents. Importantly, we also see in table 2 that the share of document attributed with high confidence to those reporting about protest event is larger than in any other probability level – almost 2'500

documents have a probability of 100 percent of reporting about protest events. Thanks to these two features, we find 50 percent of all protest events among the documents that get a probability level ranging from .91 to 1. Overall the performance of the classifier used to set the thresholds for the inclusion of documents in the universe from which we sample documents for manual coding is satisfying.

<Figure 2>

In figure 2, we present graphically the distribution of irrelevant documents at the various thresholds. This gives us a summary of the table discussed above and clearly shows that the threshold that we set to delimit the universe of document from which we sample is situated at the point when the share of irrelevant documents drops. This share drops again around .65 and then, more smoothly goes down at the different probability levels until it reaches its lowest point at the highest confidence score (1). The overall trend is similar across the three countries. However, in Spain the share of irrelevant is consistently higher at all probability scores.

<Figure 3>

Another test related to the use of thresholds is presented in Figure 3, this test relates not to irrelevant documents but rather focuses on the protest events that we have coded. More specifically, Figure 3 presents the share of protest events by action form at various probability levels for each of the three countries included in our test. Interestingly, we see that the share of protest events identified at the different probability levels varies according to action forms. Strikes shows the perfect pattern for us, no strikes found below our threshold and a steady increase above the thresholds. Similarly, demonstrations and violent protest events increase as we move up the probability levels. Although the data shows some small ups and downs, the general trend is also that of a steady increase. Below the threshold, we observe to spikes – at .45 and .35 – where we found a few instances of demonstrations and violent events. But in general, the trend for these three forms show that

we can be quite confident about the quality of the data. Yet, for the fourth category of events which includes all the other action forms – like petitions and more symbolic action that are difficult to classify – we find a more hectic trend reflecting the fact that the document-classifier has a more difficult time to identify documents reporting about these events. Looking at this line more closely, we see that the highest share of other protest events is found around .65. But we also see that the share of events found at the different probability levels is rather stable below 10 percent above the .61 threshold. Below the threshold, the picture is not more convincing since we have a few spikes at .55, .40, .20, and even at .10. This suggest that we can be less confident with the other category of protest events, we might have important biases in this category and we need to investigate more the exact types of events that we find in this category above and below the threshold. Tentatively, we conclude that the use of the document-classifier to limit the universe of documents from which we sample did not impair the quality of our data. This is certainly the case for strikes, but also for demonstrations and violence. For protest events taking other forms, we need to evaluate it further by checking also specific action forms, actors involved, and issues addressed.

Capturing trends and validity tests

A major criticism addressed to protest event analysis relates to the validity and reliability of the data. Most criticism relate to the use of newspaper as sources to account for evidence of protest event taking place (see Earl et al. 2004, Ortiz et al. 2005 for reviews). The problems evidenced by this research relate to the news worthiness of protest events which vary depending on political agenda (Oliver and Maney 2000) and on the size of the event (Biggs 2016). Others point at problems related to sampling techniques used to reduce the amount of newspapers to read, for instance using the Monday editions (Earl et al. 2004). Our strategy here aims at assessing what some refer to as "relative bias" (Strawn 2008). We compare our data to two datasets constructed by manual coding of national newspapers. In particular, we compare two countries: Spain and Portugal. The former is a big country, largely covered by international media during the crisis, the other is a much smaller country that received less attention. In so doing, we assess our approach in two rather different

settings. In addition, we propose comparing our dataset to protest events in Spain and Portugal retrieved through ICEWS – that is another data generated through automated coding.

For Spain, we can rely on a dataset established by Martin Portos-Garcia, based on a day-to-day reading of El Pais covering the period 2007-2014, for Portugal, we rely on the dataset produced by Accornero and Ramos Pinto (2014), based on a systematic reading of Diário de Noticías for the period 2010-July 2013[10]. We compare our preliminary results for the .25 sample with these two datasets which serve as our 'gold standards'. At the same time, we also compare the two with the corresponding data provided by ICEWS. For the comparison, we calculate the correlations between the monthly frequencies of the three sets of data. The correlations give us an idea of the extent to which the development over time of our own data covaries with the development of the hand coded data and of the ICEWS data. Depending on the extent to which our series covary closely with the gold standard, we can be more or less confident that they capture the ups and downs of protest in the different countries, even if the levels may not be the same. We calculate two types of correlations in each case – one for the raw frequencies, and one for three-monthly moving averages. Moreover, we calculate separate correlations for the total number of events and for the number of events in different action forms. Table 3 presents the results.

<Table 3>

As the table shows, the overall correlations for the total number of events are at best moderate. In the case of Spain, our series correlate more closely with the gold standard than the ICEWS data, but they are themselves closer to the ICEWS data than to the fully hand coded data. The performance of our data lies somewhere between that of fully hand coded data and that of the ICEWS data. Note that the match is generally closer for moving averages than for the raw data, which suggests that moving averages provide somewhat more reliable data. The performance of our data is better for some specific action forms,

[10] We would like to thank these colleagues for having provided us with their data for the purpose of this test.

such as demonstrations, strikes, and confrontational protest. By contrast, it is really poor for symbolic protests and for violent actions. This is not entirely unexpected, given that the definition of what is symbolic protest and what are violent actions is quite difficult. Figure 4 presents the development over time for pairs of series for the total number of events and for the demonstrations in Spain.

<Figure 4>

In the case of Portugal, the overall matching is worse than for Spain, except for demonstrations for which it is as good, or even better, than for Spain. The match is poor for strikes and other events. The rather poor match for Portugal suggests that for countries with few protest events we should take a larger sample than for countries with a lot of protest in order to get reliable results. We shall pursue this hunch in the next steps of our own work, we propose to increase the sample to .5 to improve our dataset.

Conclusion

We have presented the procedure that we have followed for the semi-automated construction of a large dataset of recent European protest. We start with a large amount of news documents retrieved using a keyword search. We apply document and event mention classification to draw the boundaries of the universe of documents of interest. From this, we draw a random sample of documents for manual coding. We assess biases in our data related to the automated selection of relevant documents and to the use of newswires. Having done that, we have confidence in the quality of our data. We describe the automated coding tools that we have built and evaluate their performance. We also highlight the challenges hoping that researchers interested in furthering automated protest event coding could build on our study. In conclusion, we present our thoughts on future topics for automated protest event analysis.

The importance of sufficiently investing in the conceptual work cannot be stressed enough. Owing to the breadth of action forms, using a very general query for document

retrieval has resulted in a starting corpus with an estimated five percent of relevant documents. As we started to work on document classification, we realized that our definition of protest – which, reflecting the common practice in the field, covers a variety of action forms and includes many exceptions to coding rules – makes the task too hard. One starts to see the circularity of the problem. Thus, we strongly recommend that researchers interested in the use of automated tools, as a first step, invest in the refinement and explication of their object of study. Protest is not the only vaguely defined concept, and others have pointed out similar problems with concepts like policies (Burstein 2014). The blurriness of sociological concepts constitutes a challenge for the empirical study of social and political phenomena, let alone automating their study.

Another piece of advice relates to the evaluation of the performance of tools at each step of the automated procedure. As mentioned in the introduction, evaluation is crucial in the design of automated approaches (Grimmer and Stewart 2013). A careful and forward-looking evaluation design contributes to the development of tools for automated coding by providing material not only for the evaluation but also for future development of the tools. Yet, these evaluations are difficult to design because the evaluation standards are different in the social sciences and NLP. Not only they are different but they are also difficult to reconcile. The field of protest event coding offers the possibility to work on the development of evaluation designs that come closer to the interests of both social sciences and NLP.

Lastly, social scientists should adopt the practice of shared tasks popular in NLP research. A shared task is a challenge in which multiple teams work on solving the puzzle at hand. This kind of work contributes to the intensive exchange of ideas, the formulation of well-founded widely accepted problem statements, the development of common resources and benchmarks for evaluating completing approaches. The amount of work that goes into organizing a challenge should not be underestimated, however the benefits of having conducted one could be profound to a whole field.

References

Accornero, Guya and Pedro Ramos Pinto. 2014. "'Mild Mannered'? Protest and Mobilisation in Portugal under Austerity, 2010–2013." *West European Politics* 38(3): 491-515.

ACE. 2005, "Ace (Automatic Content Extraction) English Annotation Guidelines for Events".

Barranco, José and Dominique Wisler. 1999. "Validity and Systematicity of Newspaper Data in Event Analysis." *European Sociological Review* 15(3): 301-22.

Biggs, Michael. 2016. "Size Matters: Quantifying Protest by Counting Participants." *Sociological Methods and Research*.

Boschee, Elizabeth, Premkumar Natarajan and Ralph Weischedel. 2013. "Automatic Extraction of Events from Open Source Text for Predictive Forecasting." Pp. 51-67 in *Handbook of Computational Approaches to Counterterrorism*, edited by V. S. Subrahmanian. New York: pringer.

Burstein, Paul. 2014. *American Public Opinion, Advocacy, and Policy in Congress. What the Public Wants and What It Gets*. Cambridge: Cambridge University Press.

Chojnacki, Sven , Christian Ickler, Michael Spies and John Wiesel. 2012. "Event Data on Armed Conflict and Security: New Perspectives, Old Challenges and Some Solutions." *International Interactions* 38(4):382–401.

Croicu, Mihai and Nils B. Weidmann. 2015. "Improving the Selection of News Reports for Event Coding Using Ensemble Classification." *Research & Politics* 2(3).

Dalton, Russell J. 2016. "The Potential of Big Data for the Cross-National Study of Political Behavior." *International Journal of Sociology* 46(1):8-20.

Earl, Jennifer, Andrew Martin, John D. McCarthy and Sarah A. Soule. 2004. "The Use of Newspaper Data in the Study of Collective Action." *Annual Review of Sociology* 30:65-80.

Finkel, Jenny Rose, Trond Grenager and Christopher Manning. 2005. "Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling." in *43rd Annual Meeting on Association for Computational Linguistics*, edited by A. f. C. Linguistics.

Francisco, Ron. 1996. "Coercion and Protest: An Empirical Test in Two Democratic States." *American Journal of Political Science* 40(4):1179-204.

Franzosi, Roberto. 1989. "From Words to Numbers: A Generalize and Linguistic-Based Coding Procedure for Collecting Textual Data." *Sociological Methodology* 19(1990):225-57.

Gerner, Deborah J., Philip A Schrodt, Omur Yilmaz, and Rajaa Abu-Jabr. 2002. "Conflict and Mediation Event Observations (cameo): A New Event Data Framework for the Analysis of Foreign Policy Interactions." in *International Studies Association*. New Orleans.

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3): 267-97.

Hanna, Alex. 2014. "Developing a System for the Automated Coding of Protest Event Data." Vol. SSRN 2425232.

Hripcsak, George and Adam S. Rothschild. 2005. "Agreement, the F-Measure, and Reliability in Information Retrieval." *Journal of the American Medical Informatics Association* 12(3):296-98.

Huang, Ruihong and Ellen." . 2013. Riloff. 2013. "Multi-Faceted Event Recognition with Bootstrapped Dictionaries." in *HLT-NAACL*, edited by A. f. C. Linguistics. Atlanta, Georgia.

Hutter, Swen. 2014. "Protest Event Analysis and Its Offsprings." Pp. 335-67 in *Methodological Practices in Social Movement Research*, edited by D. della Porta. Oxford: Oxford University Press.

Imig, Doug. 2001. "Methodological Appendix: Building a Transnational Archive of Contentious Events." Pp. 253-59 in *Contentious Europeans: Protest and Politics in an Emerging Polity*, edited by D. Imig and S. Tarrow. Lanham: Rowman & Littlefield Publishers.

Imig, Doug and Sidney Tarrow. 2001. "Mapping the Europeanization of Contention: Evidence from a Quantitative Data Analysis." Pp. 27-49 in *Contentious Europeans: Protest and Politics in an Emerging Polity*, edited by D. Imig and S. Tarrow. Lanham: Rowman & Littlefield Publishers.

Jenkins, J. Craig and Thomas V. Maher. 2016. "What Should We Do About Source Selection in Event Data? Challenges, Progress, and Possible Solutions." *International Journal of Sociology* 46(1):42-57.

Jenkins, J. Craig , Charles Lewis Taylor, Marianne Abbott, Thomas V. Maher and Lindsey Peterson. 2012. "The World Handbook of Political Indicators Iv." Vol. Columbus, OH: Mershon Center for International Security Studies, The Ohio State University.

King, Gary and Will Lowe. 2003. "An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design." *International Organization* 57(3):617-42.

Koopmans, Ruud and Dieter Rucht, eds. 2002. *Protest Event Analysis*, Vol. 16, Edited by B. Klandermans and S. Staggenborg. Minneapolis, London: University of Minnesota Press.

Kriesi, Hanspeter, Ruud Koopmans, Jan Willem Duyvendak and Marco G. Giugni. 1995. *New Social Movements in Western Europe*. Minneapolis: University of Minnesota Press.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard and David McClosky. 2014. "The Stanford Corenlp Natural Language Processing Toolkit." Pp. 55-60 in *ACL (System Demonstrations)*. Baltimore, Maryland USA.

McPhail, Clark and David Schweingruber. 1999. "Unpacking Protest Events: A Description Bias Analysis of Media Records with Systematic Direct Observations of Collective Action - the 1995 March for Life in Washington, D.C." Pp. 164-98 in *Acts of Dissent. New Developments in the Study of Protest*, edited by D. Rucht, R. Koopmans and F. Niedhardt. Lanham, MD: Rowman & Littlefield Publishers.

Mikhaylov, Slava, Michael Laver and Kenneth R. Benoit. "Coder Reliability and Misclassification in the Human Coding of Party Manifestos." *Political Analysis* 20(1):78-91.

Nam, Taehyun. 2006. "What You Use Matters: Coding Protest Data." *PS: Political Science & Politics*:281-87.

Nam, Taehyun. 2007. "Rough Days in Democracies: Comparing Protests in Democracies." *European Journal of Political Research* 46(1):97-120.

Nardulli, Peter F., Scott L. Althaus and Matthew Hayes. 2015. "A Progressive Supervised-Learning Approach to Generating Rich Civil Strife Data." *Sociological Methodology* 45(1):148-83.

Oliver, Pamela E. and Gregory Maney. 2000. "Political Processes and Local Newspaper Coverage of Protest Events: From Selection Bias to Triadic Interactions." *American Journal of Sociology* 106:463-505.

Ortiz, David G., Daniel J. Myers, Eugene N. Walls and Maria-Elena D. Diaz. 2005. "Where Do We Stand with Newspaper Data?". *Mobilization: An International Quarterly* 10:397-419.

Raleigh, Clionadh, Andrew Linke, Håvard Hegre and Joakim Karlsen. 2010. "Introducing Acled: An Armed Conflict Location and Event Dataset." *Journal of Peace Research* 47(5):651–60.

Reising, Uwe K. H. 1998. "Domestic and Supranational Political Opportunities: European Protest in Selected Countries 1980-1995." *European Integration online Papers (EIoP)* 2(5).

Reising, Uwe K. H. 1999. "United in Opposition? A Cross-National Time-Series Analysis of European Protest in Three Selected Countries, 1980-1995." *The Journal of Conflict Resolution* 43(3):317-42.

Ruggeri, Andrea, Theodora-Ismene Gizelis and Han Dorussen. 2011. "Events Data as Bismarck's Sausages? Intercoder Reliability, Coders' Selection, and Data Quality." *International Interactions* 37(3):340-61.

Schrodt, Philip A. and Deborah J. Gerner. 1994. "Validity Assessment of a Machine-Coded Event Data Set for the Middle East, 1982-92." *American Journal of Political Science* 38(3):825–54.

Schrodt, Philip A. 2012. "Precedents, Progress, and Prospects in Political Event Data." *International Interactions* 38(4):546-69.

Schrodt, Philip A., James Yonamine and Benjamin E. Bagozzi. 2013. "Data-Based Computational Approaches to Forecasting Political Violence." Pp. 129–62 in *Handbook of Computational Approaches to Counterterrorism*, edited by V. S. Subrahmanian. New York: Springer.

Sebastiani, Fabrizio. 2002. "Machine Learning in Automated Text Categorization." *ACM computing surveys (CSUR)* 34(1):1-47.

Shellman, Stephen M. 2008. "Coding Disaggregated Intrastate Conflict: Machine Processing the Behavior of Substate Actors over Time and Space." *Political Analysis* 16(4):464-77.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii. 2012. "Brat: A Web-Based Tool for Nlp-Assisted Text Annotation." in *13th Conference of the European Chapter of the Association for Computational Linguistics*, edited by A. f. C. Linguistics.

Strawn, Kelley D. 2008. "Validity and Media-Derived Protest Event Data: Examining Relative Coverage Tendencies in Mexican News Media." *Mobilization an International Quarterly* 13(2):147-64.

Stubbs, Amber C. 2013. "A Methodology for Using Professional Knowledge in Corpus." Brandeis University.

Tarrow, Sidney. 1989. *Democracy and Disorder. Protest and Politics in Italy 1965-1975*. New York: Oxford University Press.

Tarrow, Sidney. 1999. "Studying Contentious Politics: Frome Eventful History to Cycle of Collective Action." Pp. 33-64 in *Acts of Dissent. New Developments in the Study of Protest*, edited by D. Rucht, R. Koopmans and F. Niedhardt. Lanham, MD: Rowman & Littlefield Publishers.

Theobald, Martin, Jonathan Siddharth and Andreas Paepcke. 2008. "Spotsigs: Robust and Efficient near Duplicate Detection in Large Web Collections." in *31st annual international ACM SIGIR conference on Research and development in information retrieval*, edited by ACM.

van Deth, Jan W. 2012. "Is Creative Participation Good for Democracy." Pp. 148-72 in *Creative Participation: Responsibility-Taking in the Political World*, edited by M. Micheletti and A. S. McFarland. Boulder, CO: Paradigm Publishers.

Ward, Michael D., Andreas Beger, Josh Cutler, Matt Dickenson, Cassy Dorff and Ben Radford. 2013. "Comparing Gdelt and Icews Event Data." *Analysis* 21:267-97.

Yonamine, James E. 2013. "A Nuanced Study of Political Conflict Using the Global Datasets of Events Location and Tone (Gdelt) Dataset." Ph.D., The Pennsylvania State University.

Zou, Hui and Trevor Hastie. 2005. "Regularization and Variable Selection Via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301-20.

Table 1: Classifier performance. The document classifier is set at the optimal threshold of .69. The combination of classifiers is set at thresholds .61 for the document classifier and .85 for the anchor classifier. The combination predicts that a document is relevant if both classifiers do so, and irrelevant otherwise.

	document classifier	combination
Precision	.708	**.723**
Recall	.779	**.815**
F1-score	.742	**.766**

Figure 1: A sample of annotations in **brat**: Locations, dates, and actors come partially pre-annotated with Stanford named entity recognizer (Finkel, Grenager and Manning 2005). The coder indexes action forms that refer to the same event.

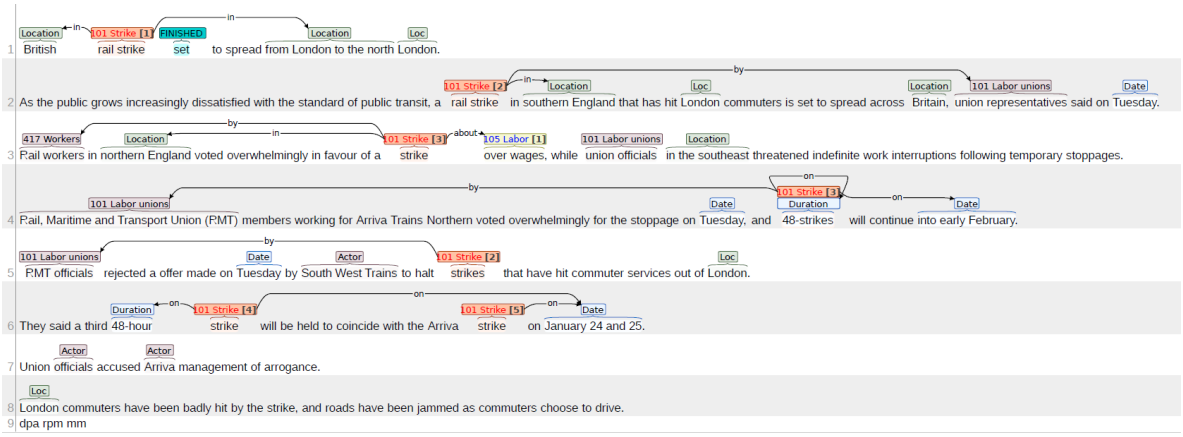


Table 2: Documents reporting or not about events and duplicate events at various probability thresholds, for Hungary, Spain, and UK

Probability level	Irrelevant doc. (%)	Events found in doc. (%)	Coded doc. (n)	Universe of doc. (n)	Doc. only duplicates (%)	Estimated doc. with events (n)
0.05	100	0	3	2019	0	0
0.1	94.44	0.02	18	12,003	0	120
0.15	100	0	72	25,775	0	0
0.2	97.4	0.08	154	28,356	0	582
0.25	99.39	0.02	165	24,526	0	153
0.3	98.73	0.06	157	19,458	0	277
0.35	95.65	0.13	138	14,896	0	750
0.4	96.19	0.06	105	11,365	0	498
0.45	96.63	0.06	89	8,999	0	336
0.5	97.18	0.04	71	6,808	0	218
0.55	88.33	0.13	60	5,555	0	673
0.6	86.79	0.13	53	975	0	616
0.65	36.33	5.13	578	3,585	16.61	1,909
0.7	28.36	5.94	617	3,894	20.58	2,325
0.75	24.3	6.79	642	3,480	19.63	2,213
0.8	25.08	6.32	622	3,109	21.06	1,979
0.85	15.25	9.07	754	3,156	20.95	2,293
0.9	17.52	9.48	799	3,532	19.52	2,555
0.95	13.69	14.86	1,176	4,386	19.3	3,419
1	3.25	41.7	2,524	5,413	9.11	4,705
Total	25.50%	100%	8,797	191,290	14.22%	25,623

Figure 2. Percentage of irrelevant documents at set probability levels

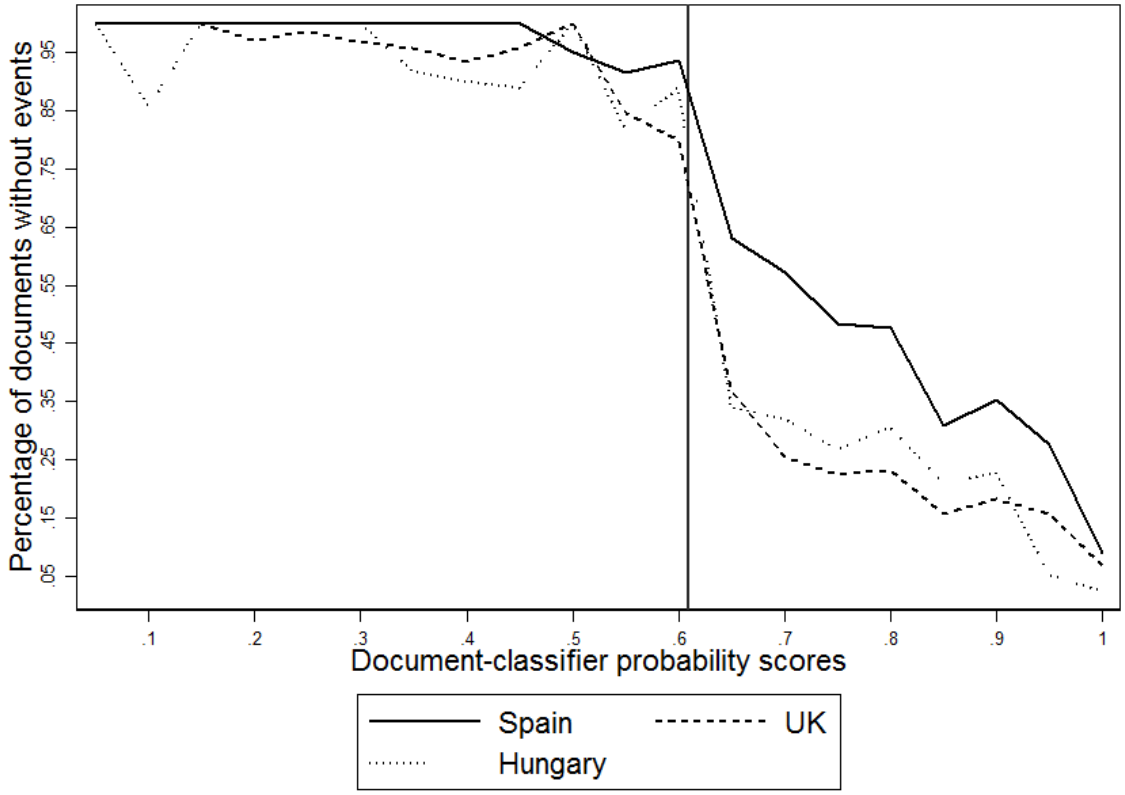


Figure 3. The percentage of protest events found at different thresholds by action forms

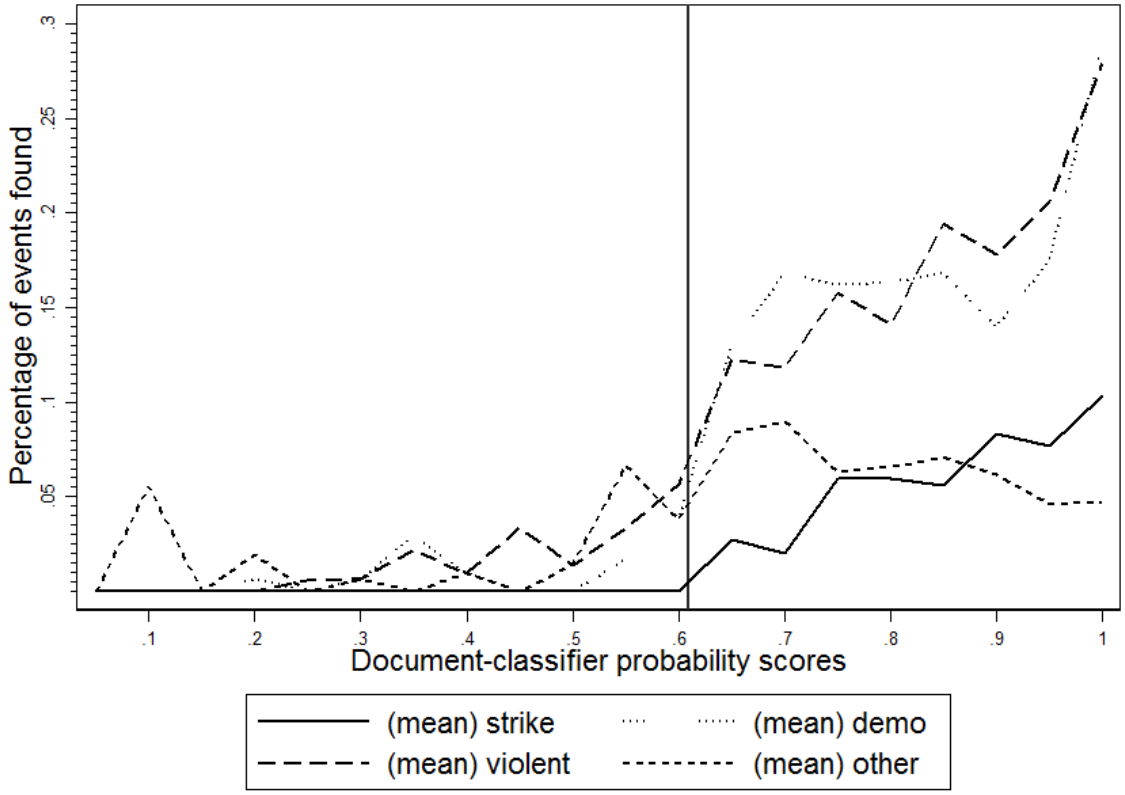
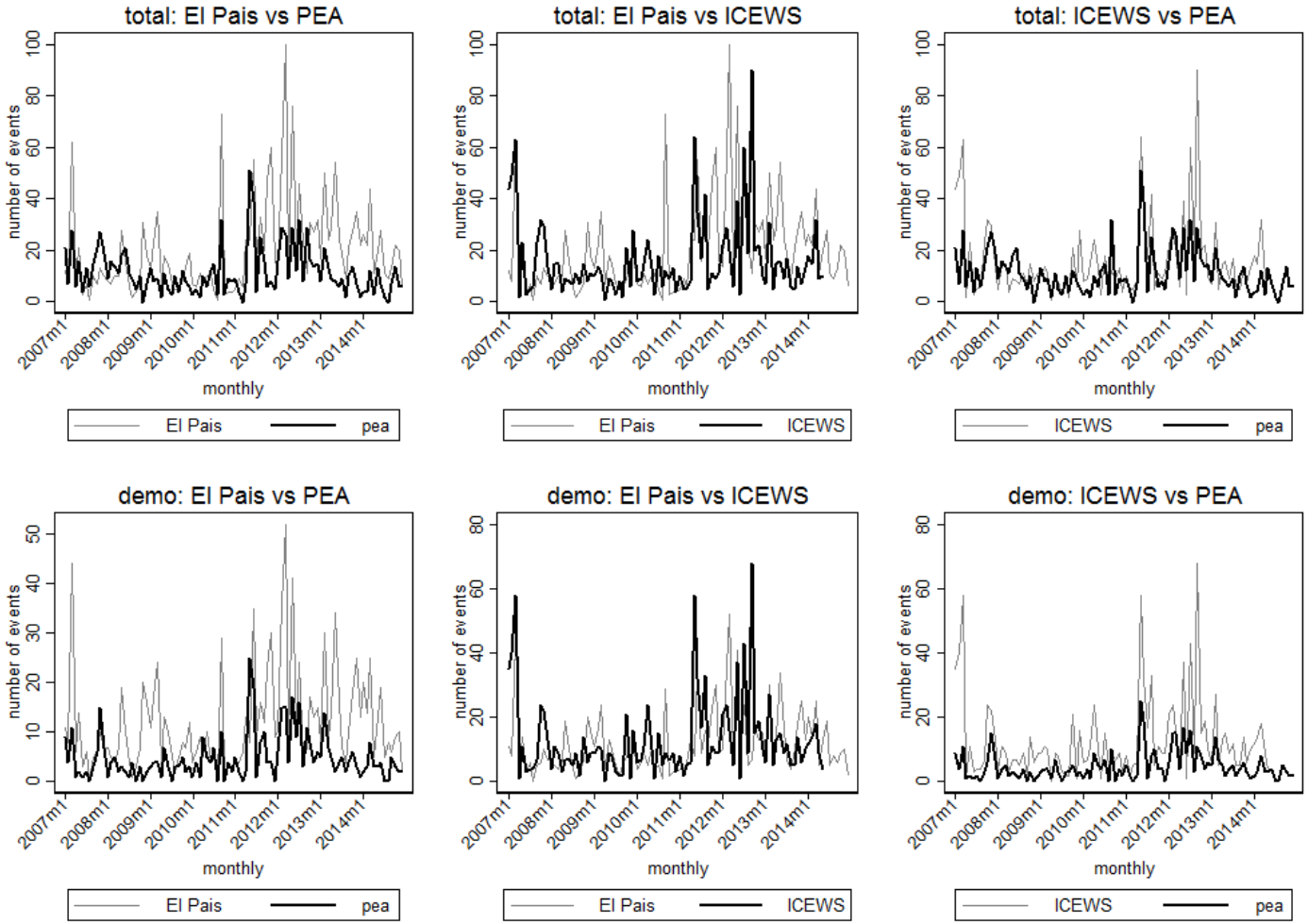


Table 3: Correlations between monthly frequencies of three sources: gold standard (El pais and Diário de Noticías), our protest event analysis (PEA) and ICEWS data

Spain	raw frequencies			moving averages ('-1/0/+1)		
	ElPais/ ICEWS	ElPais/pea	pea/ ICEWS	ElPais/ ICEWS	ElPais/pea	pea/ ICEWS
Total	0.32	0.50	0.69	0.44	0.46	0.72
Demonstrations	0.39	0.53	0.70	0.51	0.60	0.80
Strikes		0.28			0.69	
Symbolic actions		0.24			0.34	
Confrontation[1)]	0.27	0.53	0.24	0.17	0.66	0.16
violence[1)]	0.33	0.05	0.26	0.44	-0.26	0.13
Portugal	diário de notícias/ ICEWS	diário de notícias/ pea	pea/ ICEWS	diário de notícias/ ICEWS	diário de notícias/ pea	pea/ ICEWS
Total	0.30	0.34	0.51	0.05	0.33	0.39
Demonstrations	0.36	0.51	0.53	0.08	0.71	0.47
Strikes		0.20			0.16	
others	0.00	-0.22	.14	-.38	-0.65	0.26

Notes: [1)]The Spanish gold standard does not distinguish between confrontational and violent actions, which is why we use the combined category both for the calculation of correlations with violent and confrontational forms in the other sources

Figure 4: Comparison of time series for gold standard (El Pais), our own data (pea) and ICEWS: total number of events and demonstrations, for Spain



Appendix 1: Documents and event distribution, for Hungary, Spain, and UK

	< .61	> .61
Universe of newswires published in English		
Number of documents in the universe	160,735	30,555
Estimated number of documents with events in the universe	4,155	24,646
Estimated number of unique events (e.g. without duplicates)	4,155	22,653
Sample of documents coded		
Number of documents coded	1,085	7,702
Number of documents with events	37	5,264
Number of documents with duplicate events	-	1,251
Number of irrelevant documents	1,048	1,187

Appendix 2. The percentage of protest events found at different thresholds by action forms, for Hungary, Spain, and UK separately