Towards Building a Political Protest Database to Explain Changes in the Welfare State

Çağıl Sönmez
Department of Computer Engineering
Boğaziçi University
Istanbul, Turkey
cagil.ulusahin@boun.edu.tr

Arzucan Özgür
Department of Computer Engineering
Boğaziçi University
Istanbul, Turkey
arzucan.ozgur@boun.edu.tr

Erdem Yörük
Department of Sociology
Koç University
Istanbul, Turkey
eryoruk@ku.edu.tr

Abstract

Despite considerable theoretical work in social sciences, ready to use resources are very limited compared to digitally available mass media resources. Thus, this project creates a political protest database from online news resources in Brazil that will be used to explain Brazilian welfare state policy changes. In this paper we present the preliminary results of a system that automatically crawls digital resources and produces a protest database, which includes events such as strikes, rallies, boycotts, protests, and riots, as well as their attributes such as location, participants, and ideology.

1 Introduction

Social assistance programs in Brazil have largely expanded during the last two decades. The work presented in this paper is part of a project, which hypothesizes that this social assistance expansion in Brazil is a political response of the Brazilian state to the changes in social movements, particularly to the growing political radicalism of the poor and ethnic/racial minorities. Demonstrating a causal chain between social movements and social welfare outcomes in a systematic way has often been a difficult task. This is partly because of the lack of quantitative data on social movements beyond labour strike statistics and the field is marked by more or less informed speculation (Hutter, 2014). Using computational linguistics based methods and online newspaper archives, this study will create a holistic protest event database for Brazil for the period since the mid-1980s, when new social assistance programs began to emerge. This database will be used in pooled cross-sectional time-series regression analysis to explain welfare policy changes.

The protest database will count the number of events such as strikes, rallies, boycotts, protests, riots, and demonstrations, i.e. the "repertoire of contention" (Tarrow, 1994; Tilly, 1984). It will also indicate the location, city, neighbourhood of the event, ethnicity, religion, political identity of participants and organizers, the number of participants, death and casualty if occurred. We will collect data on all protest events and operationalize protest events of the poor by including (i) spontaneous or organized protests that take place in poor urban and rural areas, (ii) protests led by organizations (political, ethnic, religious or criminal) that work among the poor, independently of the location of the protest event.

The research does not intend to produce an exhaustive count for all, or for even most incidences of political events, since newspapers report on a fraction of the events that occurred (Davenport, 2009; Earl et al., 2004; Ortiz et al., 2005). The assumption is that during times of strong social movements, newspapers report social events more than usual (Silver, 2003). Therefore, the database will count each time that an event is reported in order to differentiate events in terms of their importance. It intends to create a measure of the changing levels of grassroots politics events over time and space during the welfare transformation. It is interested in the waves of contentious political activities with a comparison between the poor and other social groups.

Newspaper archives are the most reliable source from which to create protest databases, i.e. to

Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 106–110,
Berlin, Germany, August 11, 2016.

transform "words to numbers" as they provide access, selectivity, reliability, continuity over time and ease of coding (Hutter, 2014; Franzosi, 2004). International news wires and newspapers are not the best source in cross-national research because of the low level of incidence reported on each country, undermining the representativeness of each case (Imig, 2001). Yoruk (2012) has already created a protest database for Turkey that records and classifies protest activities spanning the whole period from 1970 on by leading a research team that manually surveyed microfilm archives. This database shows that grassroots politics in Turkey has shifted from the formal working class to the informal working class and from Turks to Kurds, which explains the shift in Turkish welfare policies from social insurance to social assistance and the disproportional targeting of the Kurdish poor in social assistance provision.

The protest database, the initial phase of which is introduced in this paper, will be the first comparable protest event database on emerging markets, created using local news sources and, ambitiously, using computational methods of natural language processing and machine learning.

The protest database includes events and event properties (Table 1).

Event type	protest, strike, armed struggle, occupation, rebellion
Participants	workers, teachers, poor, peasant, favelado, student, women, youth, environmentalist
Organizer	labor Union, political party, illegal party, student organization, NGO, religious organization, occupational organizations, drug traffic, peasant organization
Neighbourhood or District	centro, burantan, zona norte, zona sul, Jardim Educandário, Liberdade, etc.
Place	factory, street, university, neighborhood, courthouse, political party, public office, theatre, association, workplace, square, building, fazenda
City	Sao Paulo, Ribeirão Preto, São José, Araçagi, etc.
Participants ethnicity	mix, white, black, Indigenous, Asian
Participants ideology	left-wing, right-wing, religious, feminist, environmentalist, uncertain

Table 1: Event Attributes

In this paper, we present the article classification and entity tagging results of a system that targets producing a protest database automatically, using newspaper articles/archives from previous decades. We develop a classification module that classifies newspaper articles as reporting or not reporting a protest event. The articles that are classified as reporting a protest event are further processed and the entity mentions are extracted using our supervised maximum entropy tagger. The classification and entity tagging methods are evaluated using a manually annotated data set. In addition, the results of running the classification method on 200k newspaper articles are reported.

2 Methodology

First, we compile a newswire data set that includes daily news articles in textual form from a local newspaper. Next, we develop a classification system that filters out news articles that do not include any protest events. Lastly, we build an entity extraction system that identifies entity mentions such as the location or participants of an event.

2.1 Newswire Data Set

In the manually produced Turkish protest database (Yoruk, 2012), an average of three protest events per day for 365 days during the last 30 years, yielded a 30 thousand entry database.

We collected publicly available news articles that had been digitized and are available at the newspaper archives from Brazilian daily Folha de São Paulo[1]. The Folha Digital News Archives are available beginning from early 1920s. However, only after 1994 articles are available in text format, older archives are only available in pdf (of image) format.

We collected 200 thousand news articles in Portuguese, published between 2004 and 2015 at Folha de São Paulo. The number of articles between the years 2007 and 2011 are shown in Table 2. We only collected the articles from specific categories such as daily and politics. Our Portuguese Newswire data set is publicly available[2].

Year	2007	2008	2009	2010	2011
News count	18579	19281	16337	24062	22372

Table 2: Number of news articles per year between 2007 and 2011

[1] www.folha.uol.com.br/

[2] mann.cmpe.boun.edu.tr/folha_data/

2.2 Classification

Classification is an important step in our system. Newspaper archives include several news articles, and keyword based search yields thousands of irrelevant articles besides the few relevant ones. Given the news articles, we trained a binary classifier to differentiate protest-related news articles from others.

We converted the data into feature vectors using Weka "StringToWordVector" function and selected top 50 words for each class using tf and idf transformations on word count[3].

We compared different classifiers using our manually annotated newswire data set, namely, Random Forest (RF), Support Vector Machines (LIBSVM) (Chang and Lin, 2011), John Platt's sequential minimal optimization algorithm for training a support vector machine classifier (SMO) (Platt, 1999), Multilayer Perceptron (MLP), C4.5 decision tree (DT), Voted Perceptron (VP), Naive Bayes (NB), and Naive Bayes with kernel estimator (NB-K). The performance results of each classifier are available in Section 3.

2.3 Data Set Annotation

The system first classifies protest related news and secondly extracts components of protest information (participants, place, ethnicity etc.) via entity tagging.

For news article classification, 1000 news articles (500 reporting protest events, 500 not reporting protest events) are manually annotated and used for training and evaluation.

For entity tagging, 500 news articles are manually annotated following the ACE 2005 annotation guideline (Consortium and others, 2005). ACE is a comprehensive annotation standard that aims to annotate entities, events, and relations within a variety of documents in a consistent manner (Aguilar et al., 2014). We used the BRAT annotation tool (Stenetorp et al., 2012) for annotating the corpus (See Figure 1). Brat[4] is based on a visualizer and was initially developed to visualize BioNLP'11 Shared Task data.

[3]Configuration used to compute feature vectors: *weka.filters.unsupervised.attribute.StringToWordVector -R first-last -W 50 -prune-rate -1.0 -C -T -I -N 0 -L -M 1*

[4]brat.nlplab.org/

2.4 Entity Tagging

For entity tagging we used a maximum entropy model (Berger et al., 1996). We used the maxent[5] (Maximum Entropy Modeling Toolkit) library to built our entity tagger with BIO scheme and textual features.

3 Preliminary Results

The results of each article classifier computed using the Weka tool (Hall et al., 2009) are shown in Table 3. These results are obtained using 10-fold cross-validation over the 1000 manually annotated news articles described in Section 2.3. The best performance with an F-measure of 95.4% is achieved by the Random Forest model.

Classifier	Precision	Recall	F-Measure	TP	FN
RF	95.4	95.4	95.4	461	19
SMO	95.4	95.2	95.2	450	10
MLP	94.2	94.2	94.2	455	25
DT	93.8	93.7	93.7	448	23
VP	92.4	92.4	92.4	449	37
LIBSVM	92.4	92.4	92.4	440	37
NB	91.6	91.4	91.4	461	59
NB-K	91.5	91.1	91.1	465	66

Table 3: Comparison of different classifiers on the news article classification data set

We ran the Random Forest classifier over the 200 thousand news articles that we compiled from Brazilian daily Folha de São Paulo. The classifier identified 20 thousand articles as reporting protest events. Figure 2 shows the first tentative results of our analysis, indicating the changes in the number of total monthly protest events in Brazil between 2004 and 2011.

We used 10-fold cross-validation over the 500 news articles manually annotated for events to evaluate our entity tagger. The accuracy obtained is 76.25%.

4 Discussion and Future Work

The focus in this paper is Brazil and Brazillian Portuguese newswire text. However, our ultimate goal is to build our system in a way that will produce protest databases for other emerging countries using local newspaper archives.

The future work will be a further modification, where we will form a language independent

[5]http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

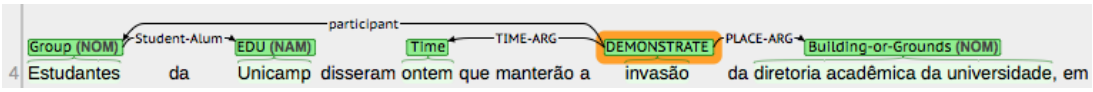

Figure 1: Annotation of a sentence in Brat

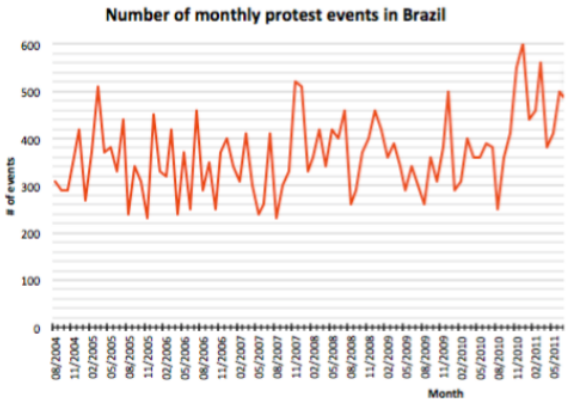

Figure 2: Changes in the number of total monthly protest events in Brazil between 2004 and 2011

tool. Then, we will use the language independent tool on news sources in English and Spanish languages, for which state-of-the-art in language processing and language resources is much more developed than for Portuguese. A tool for Turkish will also be produced by utilizing the manually created protest database in (Yoruk, 2012) for training and evaluation.

A comparative analysis of protest behaviour using quantified indicators from newspaper archives from each country will be a novelty in the literature. The collected data will be analyzed both as time-series indicator and independent variable in a pooled cross-sectional time-series multivariate regression analysis to establish causal relations between protest waves and welfare policy changes.

Acknowledgments

This research was supported by Marie Curie FP7-Reintegration-Grants within the 7th European Community Framework Programme. We would like to thank Duru Ors and Atılberk Çelebi for their contributions during the data set corpus collection and annotation process. We would also like to thank Deniz Yuret who provided insight and expertise that greatly assisted the research.

References

Jacqueline Aguilar, Charley Beller, Paul McNamee, and Benjamin Van Durme. 2014. A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. *ACL 2014*, page 45.

Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Linguistic Data Consortium et al. 2005. ACE (automatic content extraction) english annotation guidelines for events.

Christian Davenport. 2009. Regimes, Repertoires and State Repression. *Swiss Political Science Review*, 15(2):377–385, June.

Jennifer Earl, 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.

Roberto Franzosi. 2004. *From words to numbers: Narrative, data, and social science*, volume 22. Cambridge University Press.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Swen Hutter, 2014. *Methodological Practices in Social Movement Research*, chapter Protest Event Analysis and Its Offspring, pages 335–367. Oxford University Press, Oxford.

Douglas R. Imig. 2001. *Contentious Europeans: Protest and Politics in an Emerging Polity*. Rowman & Littlefield, Lanham - Boulder - New York - Oxford, January.

David Ortiz, Daniel Myers, Eugene Walls, and Maria-Elena Diaz. 2005. Where Do We Stand with Newspaper Data? *Mobilization: An International Quarterly*, 10(3):397–419.

John C. Platt, 1999. *Advances in Kernel Methods*, chapter Fast Training of Support Vector Machines Using Sequential Minimal Optimization, pages 185–208. MIT Press, Cambridge, MA, USA.

Beverly J. Silver. 2003. *Forces of Labor: Workers' Movements and Globalization Since 1870*. Cambridge University Press, April.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April. Association for Computational Linguistics.

Sidney Tarrow. 1994. *Power in Movement: Social Movements, Collective Action and Politics*. Cambridge University Press, Cambridge, UK, July.

Charles Tilly. 1984. *Big Structures, Large Processes, Huge Comparisons*. Russell Sage Foundation, New York, December.

Erdem Yoruk. 2012. *The Politics of the Turkish Welfare System Transformation in the Neoliberal Era: Welfare as Mobilization and Containment*. Ph.D. Dissertation, The Johns Hopkins University, Baltimore, Maryland.