Development of a methodology to measure media context in the European Social Survey

Natalja Menold, Johann Schaible, Theoni Stathopoulou, Cornelia Zuell

Please cite as: Menold, N., Schaible, J., Stathopoulou, T. And Zuell, C. (2018) *Development of a methodology to measure media context in the European Social Survey*, London: publisher- European Social Survey ERIC.

Development of a Methodology to Measure Media Context in the European Social Survey

Natalja Menold, Johann Schaible, Theoni Stathopoulou, Cornelia Zuell

Content

1. Aim of the Project

Media context has been discussed in the European Social Survey (ESS) as a relevant part of cultural or societal context, which should be considered in cross-cultural research (Fernee, Stoop & Harrison, 2012). According to the authors, in the cross-national studies it is relevant to have context information to understand and to correctly interpret cross-cultural differences with regard to attitudes and behaviors of individuals. This cultural context can be influenced by important national or international events, reported in mass or discussed in social media, so that there is a notion to systematically record and analyze the media context and to produce a corresponding database to the researchers who use the data of the ESS.

The ESS set the aim and the task of the project as follows: "The ESS ERIC wishes to commission a scoping report on potential methodologies for measuring media context on the European Social Survey. The report should review past approaches to this problem within the European Social Survey, evaluate alternatives where they already exist, and explore the merits and costs of developing a new methodology that reduces the burden on national coordinators and, if possible, eliminates the need for real-time selection and coding of media material during fieldwork periods. It should make recommendations for how the ESS ERIC should tackle the measurement of the fieldwork context henceforth, and produce a draft work package to implement these recommendations as part of the next ESS ERIC work programme in 2019-21." (Invitation to Tender 02/2017).

The aim of the project, GESIS Leibniz Institute for the Social Sciences conducted in cooperation with the National Centre for Social Research in Greece (EKKE) was, therefore, to provide a scoping report on potential methodologies for measuring media context in the ESS and to conduct a feasibility study about the methodology to detect claims or events in different relevant media in a cross-cultural context. The project was conducted from the January, 1st, 2018 to the June, 30th, 2018.

The working plan included the following work packages:

1) Evaluation of past approaches in the ESS to document media claims (or relevant events), as well as a methodology for their extraction and analysis.

2) Evaluation of alternative methods to document media context. This was supposed to be comprised of the three different elements. The first element was to examine the applicability of new methods across countries. The second was to compare pros and cons between different media context sources: traditional media (newspapers) and social media (twitter). The third element was to review existing tools to extract media context and to analyze the extracted data.

3) Analysis of merits and costs for the ESS.

On the basis of the results obtained in steps 1) and 2) we evaluated which approaches would be successful and usable for the analysis of media context in the ESS.

During this work package, the following questions were addressed.

a. Selection of media: With this, we evaluated which are the suitable media to search for claims for the ESS.
b. Definition of criteria for selection of tools suitable for the ESS.
c. Providing overviews of the available tools for the automatic detection of events/claims (for example, event registry, EL:DIABLO, Python Engine for Text Resolution And Related Coding Hierarchy (PETRARCH), PHOENIX or Protest Mapping (PROMAP)) and tool selection.
d. Study to test the selected tools to assess their usability for the ESS. The goal of this study was to gain knowledge about the accuracy and relevance of the outcome of different tools for measuring media context in the ESS.

4) Methodology to measure media context in the ESS was developed and evaluated using previous analysis with respect to the merits and costs for the ESS.

In the following, we provide a description of the project results along the presented project steps.

2. Past Approaches of Event Reporting/Claim Coding within the ESS

Description of Past Approaches

The interpretation of findings in cross-national surveys has to take into account the diverse contexts (Kish, 1994) in which respondents perceive and answer the questions asked. The ESS has developed a corollary tool alongside the main questionnaire to measure the potential impact of significant events, both national and international, on survey outcomes starting at the first round in 2002. In the past the ESS addressed a series of methodological challenges apparent in cross-national research, such as linguistic, conceptual, and sampling equivalence, and has created a separate data set alongside the main ESS data set. In the following we provide an overview of the different methods applied in the ESS.

In the first rounds of the ESS National Coordinators (NCs) in each country were asked to report on national and international events as reported by newspapers. The ESS has created a data set where major events were reported that took place in each participating country during the field work and in the period immediately prior to the fieldwork. The rules for events which should be reported were "An event may have a clear location in time and space" (e.g., elections, crimes, disasters, etc., Stoop 2004, p. 3). Although the systematic cross- national recording of events has shown significant reporting differences among the participating countries, it has paved the way to more rigorous approaches of measuring "social, cultural and attitudinal climate changes across Europe, rather than transitory changes in the attitudinal weather" (Stoop, 2007, p. 96). Changes in social and political climate prior to and during the fieldwork might have implications for survey quality as well, affecting survey cooperation and satisficing (Bantucci & Stevens, 2015). In addition survey fatigue and high non-response rates endanger the future of "traditional" surveys. Experimentation with new modes of data collection (i.e. mobile) or web push surveys has shifted the focus to new methods and techniques. Moreover, the so-called advent of Big Data has opened up new opportunities for survey research (Alvarez, 2016, Japec et al., 2015, Strohmaier & Wagner, 2014) linking or even replacing surveys with user-generated online behavior like social media posting (Callegaro, 2016).

However as Schober et al. (2016) have shown several methodological limitations are posed when survey data are compared to media data either traditional (i.e. newspapers) or "new" (i.e. social media). In their own words, "survey and social media data differ in a number of additional ways relevant to the larger question of population versus topic coverage: in the sampled units (individuals versus posts), the sampling frame (population versus corpus), the sampling procedure (probability versus non-probability), the sample size (typically, smaller versus much larger), and the relevance to the research topic of the survey responses or social media posts (directly relevant by definition versus not necessarily relevant)." (Schober et al. 2016, p. 193).

In the context of Framework 6 Programme (ESSi), first attempts used text mining techniques to record event data (Landmann & Zuell, 2008; Stathopoulou, 2005). Landmann and Zuell (2008) developed an approach based on factor analysis. As a result, each factor represented a significant event during a specific time period.

Also as part of ESSi, an expert level symposium was organized by City University to address the need for a more systematic recording of events through political claims analysis (PCA) (Koopmans & Statham, 1999; Statham & Howard, 2013). The implementation of this media claim coding was described in the "Guidelines on media claims in Round 6" as follows: "From the range of events that occur in the world every day or week, only a small selection becomes salient to the public. […] By coding such events from contemporary news sources we not only have a record of the events themselves but also of the contexts, issues and values – public discourses – through which they are attributed meaning and become publicly meaningful." (Fernee, Stoop, & Harisson, 2012, p. 14). The implementation of the new tool (media claims) was tested in a pilot study conducted among eight ESS participating countries: United Kingdom, The Netherlands, Spain, Greece, Portugal, Lithuania, Slovakia, and Poland (Fernee & Stoop, 2013; Stathopoulou, 2012).

Even though the approach of media claims was more systematic than event reporting (event data), it entailed a similar drawback to the previous approach: the salience of an issue in the media could reveal more about the agenda setting in the media than the effect of this issue on the attitudes and beliefs of survey respondents. In other words, it was based on the assumption that the respondents were proportionally interested in politics thus affected by the political discourse in the public sphere and were exposed to media context. Past research has shown that a) national context is not restricted to measuring political climate as a non-political event may have a political effect, b) the impact of events does not follow the issue attention cycles in the media, events that are underreported or not reported at all may have greater impact in the long term, and c) events follow detectable patterns across counties (Stathopoulou 2007a, 2008, 2011a, 2011b, 2018).

Evaluation of Past Approaches in the ESS

We conducted four interviews to evaluate the claim analysis approach implemented in the last waves of the ESS. Three interview partners were colleagues participating in the claim analysis pilot project and one participant was responsible for coordinating and assisting with the collection of the media claim data in the ESS. The guideline for the interviews can be found in Appendix A.1. We provide a report of the results along with this guideline, and describe main findings for each of the issues addressed in the interviews.

The findings in detail are:

1) General experiences with coding

The interviewees reported that the weakness of the conception of the issue "claim" made the application of the provided scheme difficult. In addition, for the claim analysis some categories were missing: The claim analysis focused on political claims, while a range of relevant events like terror attacks, political scandals, or financial crises were not codable. These significant events not covered by the analysis, might influence the respondents at least as much as political events like elections do. The interviewees mentioned that the application and documentation of the claim coding need further development. Although the coding rules were quite elaborated, they might still be improved in terms of comprehensiveness.

2) Usability of the generated tool for monitoring national context

The Monitoring of National Contexts tool (MNC, Fernee, Stoop & Harrison, 2012) developed so far in the ESS was referred to as an important tool. The tool was evaluated by the interview participants to be suitable as a reference book if the analysis showed saliences in one specific country. Tracking country-specific effects of major events was helpful. However, it was also stated that it was difficult to achieve cross-country comparability in the claims data due to differences in media landscapes, chosen journals, coder practices, length of articles, etc. There were also problems in comparability within countries over time. One advantage of the tool in comparison to earlier event reporting was that the comparability of the coding variables was achieved.

3) Coding procedure

a. Experience with coding

The coding process led to some problems because the theoretical concept of claim coding was unclear. Often the coders were confronted with codings which were ambiguous in the sense that it was unclear whether to code something or not to code. The coding rules were sometimes not clear, which resulted in substantial differences in coding practices between coders in some countries. In sum, the documentation of the claims and coding procedure (Fernee, Stoop & Harrison, 2012) was suggested to be revised and rules of coding to be improved. All in all, the coding scheme was too detailed and therefore time consuming and difficult to use.

b. Problems accessing media content (newspapers; and/or online material)

Concerning the selection and accessing of media content no concrete problems were reported. Nevertheless, the main idea of the claim coding was the comparability of the results, but the text base used in different countries was not comparable. The criteria in choosing the media generated no problems in general. But there could be country-specific problems. To give an example, one interviewee reported that left-wing media in his/her country does not really exist. Another important recommendation was for some countries to use Internet media instead of printed media due to availability and costs.

c. Cost and time

The answers varied between "no high costs" to "too expensive" (including one answer that only NCs' working hours were required). Most time consuming was the translation of the texts into English.

d. Homogenization of claims

The answers of interview partners about homogenization were vague like "is always desirable", but it was also stated that due to the ambiguity of the approach, homogenization was impossible.

e. Visualization of claims

Visualization of claims was seen as an interesting idea, but evaluated as being not really important. If it were available it would enable to "build external dashboards (see examples in election studies, protest mappings (socioscope.gr) – nice charting on demand). But as long as theory is weak it will be very risky to invest in visualization of current MNC".

4) Suggestions for future development of the MNC

a. Automatization

The participants discussed using machine learning tools for topic selection.

b. Use of social media content to retrieve claims/data for MNC

The use of social media was viewed critically because a) the selection/retrieval process was evaluated to be difficult; b) a selection bias in the use of social media was expected, and c) it was evaluated to be hard to generalize results from social media to the general population. However, the interview partners stated that there would be some countries, in which the usage of social media would be more appropriate that the usage of conventional media.

3. Alternative Approaches

Besides the approach used in the ESS, there are approaches that have been implemented to analyze the content of mass and social media. In this section, an overview of these alternative approaches will be provided. Media content is primarily text content, and the suitable methods for analyzing it derive mainly from language technologies. Social media content and online activities (e.g. views, edits, likes) provide new opportunities to measure the salience of events within specific context. Next, an application in Greece, which used both sources, mass and social media, is presented. The existing tools, which can be used to pursue these alternative approaches, are described in section 4.

Mass Media Context Measurements

The field of computational linguistics provides a series of innovative techniques to discover, explore, exploit, annotate, and analyze large corpora like the ones used in the media (printed and online). There are various views regarding what exactly constitutes an event and how, through which linguistic structures, it is expressed in data (Pustejovsky, 2000). Moreover, many research efforts have focused on detecting and identifying events (Filatova & Hatzivassiloglou, 2003, Filatova & Hovy, 2001). The establishment of Research Infrastructures likes Common Language Resources and Technology (CLARIN; https://www.clarin.eu) or Digital Research Infrastructure for Arts and Humanities (DARIAH; http://www.dariah.eu/about/dariah-in-a-nutshell.html) facilitated the development of advanced tools for automated information extraction from texts. All the above-mentioned approaches aim at developing applications which can automatically extract events and recognize the spatial and temporal relations connecting them.

Event extraction for political and social science has been a long-standing topic, dating back to hand coding data. Work on automatic annotation started within the KEDS/TABARI (Kansas Event Data System / Textual Analysis By Augmented Replacement Instructions) project (Schrodt, Shannon & Weddle, 1994, Stathopoulou, 2007b). Evaluations have shown that hand coded and automatic events coding show comparable performance (King & Lowe, 2003). Several coding schemes have been developed since, e.g. ICEWS (Integrated Conflict Early

Warning System) (O'Brien, 2012). One of the most renowned and influential frameworks for event extraction is CAMEO (Conflict and Mediation Event Observations coding ontology; Gerner et al., 2003), which is still used by the ongoing Global Data on Events, Language and Tone (GDELT) project[1] (Leetaru & Schrodt, 2013). All of these efforts have focused on news data, which have traditionally been the main data source. Protest Events Analysis has been a central issue in the context of Political and Social sciences (Wuest, Rothenhäusler & Hutter, 2013).

In the context of Information Extraction (IE), several frameworks have been proposed for Event Extraction. Two of the most influential are the Probabilistic Event Model and the Structural Event Model. The former was put forward within the framework of the Topic Detection and Tracking (TDT) study (Allan et al., 1998) and especially the Retrospective news Event Detection (RED) task (Li et al., 2005). This study defined an event as "some unique thing that happens at some point in time". Four types of information are used to represent events within news articles: *who* (person), *when* (time), *where* (location) and *what* (keywords) (Wang & Zhao, 2012). Event extraction is regarded mostly as a clustering problem. Given a stream of news data, the task is to segment the stream into stories and group the stories into clusters, where each cluster represents an event. This is considered an unsupervised learning task, as it uses no labeled training examples. The Structural Event Model (SEM) was proposed in the context of Message Understanding Conferences (MUC). The aim of SEM was to identify different types of events and to develop corresponding specific templates to categorize them. Thus, an event is considered as a structure of certain information types. Consequently, specific event (scenario) templates (Chinchor & Marsh, 1998) were developed.

Social Media Context Measurements

The aforementioned approaches have been implemented on news data, which were considered for a long time the main source of events and political claims. However, the advent of social media (SM), especially Twitter during the last years, has shifted the focus of research. Hence, SM have also been extensively used for event extraction on many different topics and for various purposes such as climate change (Olteanu et al., 2015) as well as for measuring collective patterns of user behavior during major events (Lin et al., 2014, Keegan, Gergle & Contractor, 2013).

With respect to SM, some authors (Wang, Fink & Agichtein, 2015) discover social events, and their structural components, i.e. location, time and title of the event, by linking the information from the tweet and the embedded URL. In Becker, Naaman and Gravano (2011), tweets are distinguished as event and non-event messages, with the first being clustered into topic categories. Ritter et al. (2012) extract event tuples from Twitter streams and classify them into topics, using an unsupervised approach. In Popescu, Pennacchiotti and Paranjpe (2011), events concerning specific known entities are discovered and structured, using a supervised method to decide on the relevance of tweets. Li, Sun and Datta (2012) and Qin et al. (2013) both rely on text segments in order to detect and classify events in Twitter, with the first making use of Wikipedia for the filtering of real events and the second implementing feature clustering.

Temporal and spatial information has also been used for identifying and categorizing events in Twitter, as in Parikh and Karlapalem (2013) and Walther and Kaisser (2013). In Weng and Lee (2011), signals are built for each word in a tweet and then correlated to form a distinct event. Finally, in the context of protest events extraction, Zachary et al. (2015) examine mass protest

[1] http://gdeltproject.org/

that can lead to political changes at a national level, using popular hashtags and measuring the extent to which Twitter users' coordination within social networks can cause collective action. The different scope and evaluation methodology of the above systems make it difficult to compare their performance. However, most of them report a precision ranging between 70 – 85%. All of the above mentioned projects implement methodologies varying from completely unsupervised, purely data-driven approaches, to knowledge-driven methods based on domain experts. Hybrid frameworks have also been used (Hogenboom et al., 2011).

Applying a Computational Method for Media Measurement: Protest Mapping in Greece

In the context of the PROMAP[2] project, which was carried out by the EKKE, in collaboration with the "ATHENA" Research and Innovation Centre, the transformations of protest in Greece for the period 1996-2014 were examined. Moving beyond classical Protest Event Analysis (PEA), PROMAP (Stathopoulou, 2015, 2018, Stathopoulou et al. 2018) constitutes a computational social science approach applied in the framework of a cross-cutting interdisciplinary work between social scientists and computer scientists. PROMAP forms part of a platform for the visual analysis and cartographic representation of social and political data disseminated as Linked Open Data (available at: www.socioscope.gr) (Papastefanatos & Maroulis, 2018)

Based largely on the theoretical scheme of claims analysis, the project used newspapers as a primary source in order to map, document and analyse the dynamics of protest and mobilization in Greece in a longitudinal perspective. Having as the primary unit of analysis the claim, the evolution of protest events (Diani & Kousis 2014) across the specific period studied, was detected and mapped.

Under PROMAP, an event analyzer (PALOMAR platform) for news data (Papageorgiou & Papanikolaou, 2018) was developed, which has been expanded to Twitter (Papanikolaou et al., 2016), following a data-driven, yet linguistically based approach. Twitter as a dataset exhibits some entirely different features compared to news. One of the most important is the limit of 140 characters, which enforces users to condense the message they want to communicate. In order to do this, they use special elements such as hashtags, user mentions or URLs. Additionally, words that do not influence the general meaning (e.g. function words) are often omitted, making it hard for the standard Natural Language Processing (NLP) tools to process.

4. Software Selection

A method to extract media context and to perform event coding in a (semi-) automatic manner, is to use software solutions specifically designed to retrieve information from the web. Such a software solution, or tool, can harvest information from established and well-known media/news websites and detect articles that are about (major) events. For example, a tool can

[2] The PROMAP project was funded by the General Secretariat for Research and Technology for the programming period, 2013-2015, within the framework of the "So.Da.Map: Dynamic Management of Social Databases and Cartographic Representations-KRIPIS" Program.

harvest BBC news and detect articles about Brexit that had and still has a major impact on the society's reactions.

The review of such tools for automated event coding reveals the wealth of existing approaches (e.g. TABARI[3], PETRARCH[4], JabariNLP, BBN Accent, PALOMAR, Virtual Research Associates (VRA) Reader, GDELT, EventusID, etc., see Appendix A.2). The approaches have different requirements for providing input texts and they are different in how they analyse the texts as well as how the results are presented to the end user. Despite the advantage of automatically analysing texts and detecting events, each of the reviewed tools, however, has several disadvantages and shortcomings. In order to perform a valuable evaluation of the tools with respect to their using in the ESS, we defined a set of criteria that specify usability and technical requirements.

Criteria of Tool Selection

At the beginning of our software evaluation we defined criteria which are relevant for measuring media context in the ESS. We divided the criteria in two main groups: the usability of the tools and technical requirements.

Usability

1) Event/claim definition

This criterion regards the type and relevance of claims or events a tool is designed to extract. The topics or events extracted should have a relevance for the survey context in a country. There would be issues or topics like actions of country-relevant actors (political, financial, economic, religious) or major events (e.g. heavy weather events, terror attacks, scandals) including location and date. The typical content analytic information, which has been used in the claim analysis paradigm, is a "Who-did-what-to-whom" structure, which can help to describe a claim more precisely.

An additional requirement was whether a tool supports the analysis of the relevance of the extracted events or claims. One possibility would be to analyse the frequency of reporting in a specific time period. Another concept of relevance could be defined by the "fuzz"; an event produces within the population. This could be measured by the number of retweets (although there are many issues (in terms of biases) with retweeting to resolve) or by the number of discussions in online news sites.

2) Languages supported

The selected tool should be applicable for the languages of countries included in the ESS. Therefore, a tool should support as many languages as possible. Alternatively, codebooks which had to be adapted by the countries (and kept up-to-date) can be used. However, for each certain topic (e.g., protest, religious activities), codebooks in each language should be developed and updated on a regular basis, which seems to be time consuming and expensive.

3) Handling

[3] http://eventdata.parusanalytics.com/software.dir/tabari.html
[4] https://github.com/openeventdata/petrarch2

This criterion addresses an easy handling for non-informatics. A tool should be simple to use, e.g. applicable by NCs or – centralized – by social scientists at the ESS. At least an easy to use manual should be available for the selected tool.

4) Media supported (e.g. newspapers, internet, and social media like Twitter)

The usage of print media should be avoided because machine-readable texts are needed, and downloading newspaper texts requires extensive manual work. Therefore, the use of news from online sources of mass media is preferred. Main newspapers in each European country have an internet version. In Germany, for example, Spiegel.de or tagesschau.de can be used. Because different media repeat important news, the issue of the selection of certain medium is of a lower importance. In addition one could use real-time user-generated content (e.g. twitter).

Regarding the type of the input, it should also be considered how the tools/methods receive the input texts. For example, does a tool provide the possibility to use application programming interfaces (APIs) of news media web sites or does it need an explicit link for web scraping? This can affect which news media sites can be analysed.

Technical Criteria

1) Maintenance

This criterion regards the question whether a tool is continuously maintained and further developed by some community, e.g., Open Source community, or an organization. Of advantage are tools that have been continuously maintained.

2) Evaluation of the tool

With respect to the quality and suitability of a tool for use in the ESS, it would be of advantage if a tool has been evaluated, e.g., within a research project, where the results would point out the usability for the purpose of the media analysis in the ESS. The relevant points would be the amount of texts necessary/usable to get meaningful results, or the number of claims/events identifiable.

3) Costs

The cost to use a tool should be kept low and using media texts that are not free of charge should be avoided. Next, running costs for linguists and IT specialists when the tools are implemented should be minimised.

4) Platform

The platform a tool can be used with should be documented. Windows is preferred.

5) Use of (external) libraries

It should be documented, whether and which external libraries are used by a tool (for example, Stanford's CoreNLP, PETRARCH2, or ELK (Elasticsearch, Logstash, and Kibana stack)) and whether the libraries are of open-access.

6) Memory and data space usage

Using text mining tools often requires a high capacity of memory and data space, which may be not available on a standard user PC. Instead a server for computer power and data space is required, which should be avoided.

7) Proprietary issues

This criterion regards other issues, such as who owns the platform, security and customization issues.

Available Tools and Selection Process

A large variety of tools that can be used for (semi-) automatic claim/event detection was considered. Appendix A.2 shows a table comprising all investigated tools as well as a short description of the benefits/shortcomings of each tool. In general, one can distinguish tools between those which use reports from online news sites and those which use user generated content, like Twitter or Wikipedia.

The former group comprises most of the tools investigated in this project. Tools like TABARI, PETRARCH, PETRARCH 2, and others, typically use three types of information "actors", "verbs", and "phrases" and follow the Claims analysis paradigm (using the CAMEO coding ontology). Analyzing text based on this paradigm enables NLP libraries to detect specific elements in the text that can determine an event. Each investigated tool is somewhat different to its counterparts, but many tools are either follow-up projects of other tools or they are based on the same or similar method. For example, PETRARCH is the follow-up tool of TABARI with various improvements, such as being able to use entire news summaries instead of preprocessed text in tabari-specific format. PETRARCH 2 is the successor of PETRARCH, which mainly uses TreeBank data from the CoreNLP library as a tree structure of linked nodes, where each node is a Phrase object. Other tools like Jabari NLP (Java Implementation of TABARI), W-ICEWS iTrace (based on JabariNLP), and Phoenix Pipeline (a pipeline existing of a web scraper, CoreNLP, and PETRARCH) are yet other versions of the PETRARCH line of tools, like TABARI, PETRARCH (1-2), and Jabari NLP.

Other approaches like PALOMAR, BBN Accent, EL:DIABLO (or hypnos), GDELT, Virtual Research Associates (VRA) Reader, EventusID 2.0, or Lydia are not follow-up projects of the before mentioned tools. However, they all apply the claim analysis paradigm as well. This means, they also highly depend on the CoreNLP library to extract events.

Twied and Wikipedia Live Monitor rely on user generated content. Twied collects tweets from the Twitter API, infers the location based on information from gazetteers and clusters the tweets. Finally, it classifies the clusters based on relevancy (number of retweets and likes) and stores relevant events in an event database. The Wikipedia Live Monitor monitors article edits on different language versions of Wikipedia. The idea is that a high number of edits of Wikipedia's articles covering the same topics indicates potential breaking news, which can then be investigated more closely whether the breaking news comprise a specific event or not.

PALOMAR is an NLP platform that was developed within the context of the PROMAP project, which was carried out by the EKKE, in collaboration with the "ATHENA" Research and Innovation Centre. PROMAP forms part of a platform for the visual analysis and cartographic representation of social and political data disseminated as Linked Open Data (available at: www.socioscope.gr) (Papastefanatos & Maroulis, 2018).

For the evaluation, we selected tools on the criteria described above. No tool was able to meet every single one of the stated criteria. Hence, we only considered the most important criteria, which were 1) multi-language support, 2) open source, 3) maintained by developer, and 4) easy to use for non-informatics. Based on this, only PALOMAR, GDELT and the Wikipedia Live Monitor seemed as suitable candidates. Unfortunately, at the time of the evaluation,

PALOMAR was not production ready, so that it was not possible to consider it in the study. In addition to GDELT and the Wikipedia Live Monitor, we evaluated an R Topic Modeling package (LDA, Blei, Ng., & Jordan, 2003) for detecting events, because this fulfilled the predefined criteria.

5. Software Tests

GDELT

General Description

The GDELT Analysis Service is based on the database provided by the GDELT project (https://www.gdeltproject.org/). The database aims to provide texts of the world's broadcast, print, and web news which are stored from nearly all countries around the globe in over 100 languages. Texts written in languages other than English are translated automatically. Search tools allow searching for different issues, e.g. people, locations, organizations, themes, and other. To search the database for event records the user may specify a set of criteria for the event type and actors involved, along with an optional date range. The events identified here are of the type "Who-did-What-to-Whom" (claim analysis paradigm). In addition to the database, some tools are provided to allow visualization and explore events. The tools are, for example the Event Record Exporter (http://analysis.gdeltproject.org/module-event-exporter.html). Other tools allow creating geographic networks, geographic heat maps and other geographical outputs.

Some Technical Specifications

- Costs: The tools are free to use and supported by Google Jigsaw
- Platform: Cloud based
- Maintainability: New files are created daily
- Memory/data space: Runs in the cloud, query results are delivered via email; only one submission at a time allowed (e.g. if a data set is requested, a query for a heat map cannot be made at the same time)
- Handling/GUI: Graphical User Interface (GUI) available; search queries can be made on the website; easy to use
- Type of data processed: News media
- Supported languages: 65 languages; non-English news are translated into English in real-time
- Type of events/metrics: Claims analysis. The CAMEO ontology
- Input text: Not controllable, no own input possible

Testing of the Tool

In our test we considered the last week of April 2018 as time frame. We initially planned to analyze the media content in Germany, to evaluate the possibility to obtain country specific events or claims, which is required in the context of the ESS. Therefore, we specified Germany as a selection keyword. The output, shown in Table 1, contains information such as the date, the code and name of the first actor, his/her type/function (like GOV - Government, or LEG – Legislature), the recipient, his/her type of function, an event code, and numerous other geographical information. All codes are documented in the CAMEO manual.

Table 1: Events/Claims reported by GDELT (Extract)

…	Date	Actor1	Actor1 Name	Actor1 Country	…	Actor1 Type1	…	Actor2	Actor2 Name	Actor2 Country	…	Event	…
	201804	DEU	GERMANY	DEU								120	
	201804	DEU	GERMAN	DEU				GBRGOVHRI	THERESA MAY	GBR		40	
	201804	DEU	GERMAN	DEU				CAN	CANADA	CAN		46	
	201804	DEUGOV	GERMANY	DEU		GOV		IGOEUREEC	THE EU	EUR		20	
	201804	DEU	GERMAN	DEU				CAN	CANADA	CAN		46	
	201804	DEU	GERMAN	DEU				USA	WASHINGTON	USA		42	
	201804	DEUGOV	GERMANY	DEU		GOV		USA	THE US	USA		20	
	201804	DEU	GERMAN	DEU				DEUGOV	ANGELA MERKEL	DEU		20	
	201804	DEUGOV	ANGELA MERKEL	DEU		GOV		FRAGOV	FRENCH	FRA		46	
	201804	DEUGOV	GERMANY	DEU		GOV		USA	WASHINGTON	USA		20	
	201804	DEUGOV	ANGELA MERKEL	DEU		GOV		USA	WASHINGTON	USA		164	
	201804	DEU	GERMAN	DEU				DEUGOV	ANGELA MERKEL	DEU		20	

Advantages and Limitations

The GDELT is very easy to handle due to the clear user interface. Another positive aspect is that the database is steadily maintained and includes a large amount of texts and of different text types.

Aside from this positive functionality, there are many limitations of the tool for use in the ESS. The main problem is the language or country selection. GDELT does not support selecting and analyzing texts reported in a specific country. When a country is specified as a key word for search, it is maintained as a part of the claim structure (e.g. actor or object). The results obtained are then events or texts, when a country was handled as an actor and not the texts published in the media of a specific country. To give an example, if one uses "Germany" as country, he/she obtains all events containing Germany in the texts, regardless of where a text was published. Thus, if for example Polish media discuss something about "bad Germans" this would be relevant for the Polish respondents of the ESS, but nearly none of the German respondents will be aware of it. This is a serious limitation with respect to the use in the ESS.

The output of the Event Record Exporter is similar in the structure to that of the claim coding, formerly used in the ESS. Hence, extracted information is of lower relevance for the ESS, since it contains predominantly geographical information. To identify which events are reported, the CAMEO coding manual should be consulted, which is hard due to very detailed event structure.

The tools in GDELT are designed to export small subsets of the data based on search criteria, like the type or country of the actor or of the recipient, or a specific event code. Searching for events/claims not known in advance is not possible. Therefore, when using GDELT the relevant issues, which are to be searched, should be identified beforehand.

The GDELT Record Exporter is the only tool in GDELT which could support the ESS in looking for events/claims. All other options of GDELT are geographical or network outputs.

Some minor criticisms are that the underlying software codes are not open source and therefore no modifications by the users are possible. The composition of the text corpus is not clearly documented and not extensible.

In sum, with GDELT, neither a control of the media included in the analysis is possible, nor data analysis in a specific country can be conducted. The main limitation is that a systematic data collection on significant topics is not supported as well, because the keywords for the search need to be identified before the analysis.

Topic Models

General Description

Topic Modeling is a widely used text mining approach which allows identifying topics and themes in a large corpus of texts. Given the weakness of the claim analysis tools to structure the media content without a predefined topic, we decided to test this approach although it does not present a fully automatic application. The main principle is that the frequently used words which co-occur in a document are used to classify the documents into themes or topics. The

technique used for the classification is LDA (Latent Dirichlet Allocation, see Blei, Ng & Jordan, 2003).

Some Technical Specifications

- Costs: The tools are available in R or Python
- Platform: Windows
- Handling/GUI: R syntax file is available
- Type of data processed: user collected text corpus
- Supported languages: language independent; Stop word lists and lemmatization/ stemming are available
- Type of events/metrics: topics/themes
- Input text: Raw text collections

Testing of the Tool

For our test purposes we decided to use an already available database of articles published in the Guardian in July 2006 (about 1200 documents). For this text data base, an analysis of the topics was conducted by an alternative approach (Landmann & Zuell, 2008). We adapted the R software source published by Blei et al. (2003). When applying this procedure, the user has to provide the number of topics to be selected and the number of words, which are requested in the output to specify a topic. How these parameters should be specified, depends on the size of the text corpus. The number of events to be identified in a month should be higher than in a week. We decided to limit the number of topics to three and the number of words describing a topic to eight. With this parametrization, we expected to identify three significant topics in the text corpus and to obtain interpretable results.

The result is a list of three topics described by eight words each (Table 2). The task now is to define the three topics with the help of the words listed. For each word, the probability that the word belongs to this topic is also provided by the analysis (not shown in the table). The words are stemmed. Thus we added for better readability the endings as italic characters.

Table 2: Result of the topic modeling

Topic 1	**Topic 2**	**Topic 3**
minist*ry*	offic*e*	govern*ment*
peopl*e*	british	polic*e*
hizbullah	peopl*e*	isra*eli*
kill	london	israel
war	offici*al*	forc*es*
hous*es*	nation	attack
bomb*s*	blair	famili*es*
secur*ity*	militari*es*	lebanon

Topic 1 could be labeled as the "war of Hizbullah throwing bombs and killing people". The second topic explains the national British position on the (same) conflict. Finally, topic 3 describes the attacks of Israel's military forces in Lebanon. All three topics from July 2006 describe the same event under different perspectives. The results obtained by the topic modeling are in line with former analysis of these texts conducted by Landmann and Zuell (2008).

Advantages and Limitations

One advantage of the procedure is the flexibility in the textual material selected. It is applicable with any language. The main advantage is that the process of identification of the topics is systematic and controllable. The duration of the topic modeling for the given large text corpus was of five minutes.

Nevertheless, such a procedure requires preparation work as the text selection from the media under investigation is not supported by the tool. Hence, the text corpuses in each country should be prepared prior to the analysis. The next disadvantage is that the interpretation of the results is hardly possible without having an idea of relevant news in a country.

As mentioned above, topic modeling can be conducted with every media text, independent of the language. The decision on which media should be used can be done by the ESS with the support of NCs. Next, a schedule should be set up concerning the time interval, for which the texts should be collected (e.g. on a weekly basis). There are tools which allow systematic downloads depending on the type of texts required (e.g. http://boilerpipe-web.appspot.com/ for online news).

In the second step, stop words have to be defined for each language. R offers stop word lists for different languages. Stop words are words excluded from the automatic analysis which are frequently used but are meaningless for the analysis, for example, prepositions, articles, conjunctions, auxiliary verbs or others. Thus, the lists of the stop words should be extended in order to exclude them from the analysis to obtain meaningful list of words describing a topic. An example can be found in Table 3. To define the individual stop words one may use the word frequency list sorted descending by word frequencies, which is also an output, generated from the topic modeling. The most frequently used words, not useful for the analysis, may be selected from this list. This stop word list has to be created once and can be re-used in all rounds.

Table 3: Examples of additional "user-defined" stop words

can	say	one	way	use	also	however
tell	will	much	need	take	tend	even
like	particular	rather	said	get	well	make
ask	come	end	first	two	help	often
may	might	see	thing	point	look	now

In the third step – and depending on the time interval as well as the material – both, the number of topics and the number of words describing a topic should be defined. In our example, we used a text base of one month and found that not more than one event described by three topics

was identified. We expect that setting the parameter to 3 topics should work for monthly reports. However, the method is very flexible, so that for a given context different specifications can be easily tested to identify those that provide the best results.

Finally, the output word lists have to be interpreted and described as significant topics, preferable by a country expert. The identified topics could be stored in a database similar to the "old" event data base from round 1 to 5.

Wikipedia Live Monitor

We also tested Wikipedia Live Monitor. The idea is that events influencing respondents should be salient and effects could be measured by active participation of people in the discussion processes. Thus, tools like Wikipedia or Twitter (retweet) could be of interest.

The Wikipedia Live Monitor identifies events depending on changes in Wikipedia in real time. The changed or added topics/events in Wikipedia could be anything such as political claims or topics on sports or biological issues, which are discussed in specific communities. Our test shows that the tool does not support selection of countries; activities are to be analyzed for. With the tool one can track changes in Wikipedia in every country, e.g. Asia or African countries but not a single country. An example of the output of Wikipedia Live Monitor can be seen in Figure 1. In the first part of the output one can see three entries just added/changed by different people (e.g. from the U.S., and the Ukraine). In the second part of the output under the headline "Article clusters …" most active changed/reported entries are combined to "clusters" or events. In the example the main event is the French Open tennis championship which was running when the example was created.

Our test shows that the tool can help to identify relevant discussions in Wikipedia. The disadvantages are that a country specific data collection is not supported, also a focused data collection for the events or claims is not possible, which are relevant in the context of the ESS. Next disadvantage is that the users are often part of a very specific type of the population and these people could not be seen as representative of the general population of a country.

Figure 1: Output of Wikipedia Live Monitor

6. General Discussion and Recommendations for the ESS

The project aimed to develop a methodology for the media analysis in the ESS, which regards an easy to handle identification of relevant media content. The procedure should rather involve usage of easy to handle tools to ensure a (fully)-automatic data analysis and provide reliable and valid results with a relevance for the ESS.

During the project, we defined criteria for the selection of the tools, generated an overview of the available tools and selected tools and methods for a further test, when using the pre-defined criteria. The first tool we selected was GDELT, which is a claims analysis tool. The second main approach we tested was topic modeling.

The results show that the GDELT would be only of a reduced use to the ESS, as the primary goal, to identify relevant media context, could not be fulfilled. By the way of contrast, the tool is powerful in searching for claims for a known keyword. Next, the tool neither allows the user to control the countries the search is conducted for, nor to control of the media considered during the search. Alternative tools for the claim analysis are even less suitable for use by the ESS or could not be used due their specificity with respect to the topic of the analysis (political protest) or their availability for internal users only, i.e. those who belong to a specific organization.

The test results with respect to the topic modeling were more promising, because with the topic modeling a systematic search for relevant media content was possible. An application of topic modeling on a large text corpus lead to interpretable results, which also were in line with the results of former analyses of the same text corpus. This analysis was conducted by Landman and Zuell (2008), who identified media events by means of factor analysis. However, as topic modeling does not provide results in the form of sentences but just in the form of word stems, interpretation of the results requires involvement of country experts, who have an overview of the relevant media discussions in the country. Next, applying topic modeling requires a systematic selection of media texts. There are tools available, which can be used for automatic collection of the content of predefined online media (e.g. newspapers).

For the Workprogramme in 2019-21 the ESS can decide between different scenarios, which are described as follows.

Application of Topic Modeling to Measure Media Content

When applying topic modeling for the data collection on media context, the following procedure can be implemented:

1) Identify relevant media for each of the participating countries.
There should be online media, which address a broad country audience. In addition, media which is used especially by certain groups, e.g. younger, older, lower educated people can be considered. Since relevant country or worldwide events or claims are reported by every newspaper, the selection of media is a less crucial aspect. Therefore, considering a rather small number of relevant mass media (online versions) would provide satisfactory results. Nevertheless,

effort should be made to select media, which cover as much as possible the targeted population of the ESS in a country as audience. Social media can be included as well. For Germany, examples of selected media would be “FAZ”, “Sueddeutsche”, “Bild” and “Zeit”. The NCs of the in the ESS participating countries can provide suggestions for the relevant mass media, for which online applications are available.

2) Identify the period of time and frequency of media observation.
The relevant time would be shortly before (e.g. three months) and during the data collection in a country. The media can be collected on a weekly basis, whereby the topic modeling can be conducted once a month on the entire material.

3) Conduct text selection and store the texts in a text base.
For this step, automatic procedures can be used (e.g. boilerpipe: http://boilerpipe-web.appspot.com/) or adapted (e.g. NewsAPI: https://newsapi.org/). The texts should be stored for each country separately to obtain specific results for a country.

4) Conduct topic modeling of the selected texts to identify relevant topics.

5) Interpret, label and summarize the topics to obtain an exhaustive and disjunct selection.
At this step, country experts should be involved to arrive at the sensible and valid interpretations of the topics.

6) Develop a variable or a number of variables to provide data on media context in each country.

To be able to implement this method, a further pilot-project is needed, which tests the application of the proposed method for different ESS countries. We also would suggest to establish a central coordination of media analysis in the ESS, with the following tasks:

1) negotiate the online media to be selected in a country with a country expert (would be NCs)

2) agree on the time of media observation

3) select texts from the media

4) conduct topic modeling

5) agree with NCs on the interpretation of the topics

6) provide data

The central coordination would be more efficient, because one or two experts would quickly develop routines in conducting topic modeling, which would eliminate the need for other qualified country experts. When NCs had to conduct topic modeling, a central coordination would be necessary as well to define procedures, provide training, coordinate and control the results. The manual work needed to prepare topic modeling does not require a high scientific qualification and can be centrally done for all countries in a consistent manner. Based on our experiences from the testing of topic modeling during this project, we would suggest to establish a central coordinator (0.5 person month), who is assisted by one technical expert (up to 0.5 person month) and student assistants. The tasks of the latter would be selection of the texts and their storage, running of topic modeling and providing the data.

Application of Event Coding Methods

Alternatively or in parallel to the topic modeling, which is applicable at this point of time at relatively low costs, the ESS can consider to further develop and apply methods based on a computational social science approach to extract events relevant to the content of the ESS. This will entail a specific process (see Figure 2); the creation of a common codebook for all participating countries, a centrally coordinated collection of data (online media in each country and/or twitter media); data exploration through NLP tools; data analysis and visualization of results. Event coding has been an innovative methodological approach adopted by the ESS since its launch in 2002. It was the first time that a large scale comparative survey implemented a tool to measure the social and political climate in each country through media analysis. Until then event coding, known as PEA (Protest Event analysis) was used in the field of international relations and social movement research to measure cooperation and conflict or protest. The ESS can take upon an equally pioneering role in using state-of-the art computational social science techniques to develop event coding.

The following steps should be considered when implementing this approach:

1. Provide central effort to create a common codebook, with the aid of some countries that would be willing to undertake the task. Apply for funding for a core NLP team that will be able to develop and maintain a centrally designed and approved by all participating countries coding scheme of media measurement based on the topics that ESS covers (immigration, trust in political institutions, electoral behavior, religion, etc.)

Figure 2: Steps for applying a computational method

2. Link with CLARIN RI, the infrastructure for digital language resources data and tools for knowledge transfer and collaboration. In the framework of CLARIN, several efforts by substantive research teams participating in the infrastructure are being developed with the aim to analyze online language corpora like newspapers and social media texts.

Data Collection on the ESS Respondents

In addition, the ESS can consider to incorporate a question in the ESS questionnaire asking directly to the respondents, which events have affected them during a certain period of time. The question can be open or semi-open. Example can be given using questions on events from the "Comparative Panel Survey on the Dynamics of Change: Belief Formation and Political Engagement in Egypt, Tunisia, and Turkey" (MEVS, 2016).

Incorporation of such a question would be an easy to implement method which at least allows a control of socio-political media context from the point of view of the respondents. Adding the systematically collected data via topic modeling or event analysis would foster research on the subjectively perceived and non-reactively collected media context.

7. Outlook

The rapid expansion of Machine Learning (ML) Techniques, Information extraction and retrieval tools as well as the advancement of Natural Language Processing Tools has opened up new prospects for automated extraction of the desired content. As mentioned above, the establishment of Research Infrastructures like CLARIN or DARIAH has further facilitated the development of advanced tools for automated information extraction from texts. The field of computational linguistics provides a series of innovative techniques to discover, explore, exploit, annotate, and analyze large corpora like the ones used in the media (printed and online).

As Evans and Aceves (2016, p.18.2) note "supervised ML prediction tools can "learn" and reliably extend many sociologically interesting textual classifications to massive text samples far beyond human capacity to read, curate, and code. Second, unsupervised ML approaches can "discover" unnoticed, surprising regularities in these massive samples of text that may merit sociological consideration and theorization". Attitudinal large-scale surveys like the ESS have to respond to the challenges that the era of big data brings for survey research (Callegaro, 2016, Couper, 2013, Japec et.al., 2015).

However the wealth of technologies at hand still has limitations; Tools have to be maintained and updated regularly to sustain their cutting-edge usability, an effort that requires significant investment in human and material resources. Many of the methods and tools are subject to only temporary research, and once it is completed, there is no maintenance or further development of these methods and tools. Another problem is that most tools are not cross-lingual. To this end, the creation of a common codebook for all ESS countries for media measurement requires a considerable effort. Until then, the evaluated tools work best either on English text only or with the text of the country the tool was developed in. Last but not least, tools and methods analyzing social media are quite promising, but still in development. Most of them, like the Wikipedia Live Monitor, collect promising events based on retweets, likes or the number of edits of a Wikipedia page, but at the same time, they are in a prototype phase, which lacks applications that can be regularly used.

Admittedly, as of today, there is no tool for non-informatics which is easy to install and to use with satisfactory performance in detecting relevant events in news media and/or in user generated content, such as Twitter or Wikipedia. One interesting project is the PALOMAR framework that combines event detection in news and social media, but it is not production-ready yet.

To conclude, the ESS can consider to initiate and conduct pilot projects with application of topic modeling and develop improved procedures for the media context analysis in the social

science research. In addition, attention should be paid to the developments of tools in the area of event coding, claim analysis and machine learning, due to their rapid conceptual, methodological and technical development.

References

Allan, J., Carbonell, J., Doddington, G., Yamron, J. & Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription Workshop*.

Alvarez, M. (2016) (Ed.) *Computational social science. Discovery and prediction*. Cambridge: Cambridge University Press.

Bantucci, S. & Stevens, D. (2015). Surveys in context how timing in the electoral cycle influences response propensity and satisficing. *Public Opinion Quarterly*, 79, 214–243.

Becker, H., Naaman, M. & Gravano, L. (2011). Beyond trending topics: Real-world event identification on Twitter. In *Proceedings of the International Conference on Weblogs and Social Media*.

Blei, D., Ng., A. & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993-1022.

Callegaro, M. (2016) *Importance of Surveys in the Era of Big Data*. "IPSDS Talks". International Program in Survey and Data Science. University of Mannheim.

Couper, M. P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods* 7(3), 145–156.

Chinchor, N. & Marsh, E. (1998). MUC-7 Information extraction task definition. In *Proceedings of the Seventh Message Understanding Conference*, MUC-7.

Diani, M. & Kousis, M. (2014). The Duality of Claims and Events: The Greek campaign against the Troika's Memoranda and austerity, 2010–2012. *Mobilization: An International Quarterly*, 19(4), 387–404.

Evans J. & Aceves P. (2016). Machine translation: Mining text for social theory. *Annual Review of Sociology* 42, 18.1–18.30.

Fernee, H, Stoop, I. & Harisson, E (2012) *Coding media claims in the European Social Survey, Round 6. Background, guidelines and codebook*. Version 3.2

Fernee, H. & Stoop, I. (2013). Contextual data in the European Social Survey. In C. Aarts, and en M. Wittenberg (eds.). Stabiliteit en verandering in Europa, *Proceedings Vierde Nederlandse Workshop ESS.* (pp. 165-169).

Filatova, E. & Hatzivassiloglou, V. (2003). *Domain-independent detection, extraction, and labeling of atomic events*. *Proceedings of RANLP* (Vol. 3, pp. 145–152). Borovetz, Bulgaria.

Filatova, E. & Hovy, E. (2001). Assigning time-stamps to event-clauses. In *Proceedings of the 2001 ACL Workshop on Temporal and Spatial Information Processing*. Toulouse, France. Retrieved from: http://www1.cs.columbia.edu/~filatova/filatovaACLtemporal01.pdf

Gerner, D., Schrodt, P., Yilmaz, O. & Abu-Jabr, R. (2003). Conflict and mediation event observations (CAMEO): A New event data framework for the analysis of foreign policy interactions. Retrieved from: http://eventdata.parusanalytics.com/papers.dir/gerner02.pdf

Hogenboom, F., Frasincar, F., Kaymak, U. & de Jong, F. (2011). An overview of event extraction from text. In M. van Erp, W. R. van Hage, L. Hollink, A. Jameson, and R. Troncy,

(eds.), *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web at Tenth International Semantic Web Conference*. 779. (pp-48-57), Aachen: CEUR-WS.org..

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C. & Usher, A. (2015). AAPOR report on big data. *American Association for Public Opinion Research*. Retrieved from: https://www.aapor.org/Education-Resources/Reports/Big-Data.aspx

Keegan, B., Gergle, D. & Contractor, N. (2013). Hot off the Wiki: Structure and dynamics of Wikipedia's Coverage of Breaking News Events. *American Behavioral Scientist, 57*(5), 595-622.

King, G. & Lowe, W. (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(03), 617–642.

Kish, L. (1994). Multipopulation survey designs: Five types with seven shared aspects. *International Statistical Review*, 62 (2), 167-186.

Koopmans, R. & Statham, P. (1999). Political claims analysis: integrating protest event and political discourse approaches. *Mobilization: An International Quarterly*, 4(2), 203–221.

Landmann, J. & Zuell, C. (2008). Identifying events using computer-assisted text analysis. *Social Science Computer Review*, 26(4), 483-497.

Leetaru, K. & Schrodt, P. (2013). *GDELT: Global Data on Events, Language and Tone, 1979-2012.*

Li, C., Sun, A. & Datta, A. (2012). Twevent: segment-based event detection from tweets. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, ACM, (pp. 155–164). Retrieved from: http://www.ntu.edu.sg/home/axsun/paper/sun_cikm12li.pdf

Li, Z., Wang, B., Li, M. & Ma, W. (2005). A probabilistic model for retrospective news event detection. *SIGIR* 05. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.9651&rep=rep1&type=pdf

Lin Y.-R., Keegan B., Margolin D. & Lazer D, (2014). Rising tides or rising stars?: Dynamics of shared attention on Twitter during media events. *PLoS ONE 9*(5): e94093. https://doi.org/10.1371/journal.pone.0094093

MEVS (2016). *Middle Eastern Values Study. Michigan Population Studies Center. Maryland University*. Retrieved from: https://mevs.org/research-programs.

O'Brien, S. (2012). A multi-method approach for near real-time conflict and crisis early warning. In V. Subrahmanian (ed.) *Handbook on Computational Approaches to Counterterrorism,* (pp.401-419). New York: Springer..

Olteanu, A., Castillo, C., Diakopoulos, N. & Aberer, K. (2015). Comparing events coverage in online news and social media: The case of climate change. In *International AAAI Conference on Web and Social Media, ICWSM,* (pp. 288–297). Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10583.

Papageorgiou, H. & Papanikolaou, K. (2018). Data analytics meet social sciences: the PROMAP project. In T. Stathopoulou (Ed.), *Transformations of protest in Greece.* Athens: Papazisis Publishers and National Centre for Social Research. In press.

Papanikolaou, K., Papageorgiou, H., Papasarantopoulos, N., Stathopoulou, T. & Papastefanatos, G. (2016). "Just the Facts" with PALOMAR: Detecting protest events in Media Outlets and Twitter. Proceedings of the *tenth International AAAI Conference on Web and Social Media*. Social media in the newsroom (pp. 136-142). Cologne: Germany.

Papastefanatos, G. & Maroulis, S. (2018). Socioscope. A visual analytics platform for open social data. In T. Stathopoulou (Ed.), *Transformations of protest in Greece.* Athens: Papazisis Publishers and National Centre for Social Research. In press

Parikh, R. & Karlapalem, K. (2013). *Et: events from tweets. Proceedings of the 22nd International Conference on World Wide Web*, (pp. 613–620) New York: ACM.

Popescu, A. M., Pennacchiotti, M. & Paranjpe, D. A. (2011). Extracting events and event descriptions from Twitter. In *Proceedings of International Conference on World Wide Web*. Retrieved from: http://www.ambuehler.ethz.ch/CDstore/www2011/companion/p105.pdf

Pustejovsky, J. (2000). Events and the semantics of opposition. In C.Tenny and J. Pustejovsky (eds.), *Events as grammatical objects* (pp. 445–482). Stanford: CSLI Publications.

Qin, Y., Zhang, Y., Zhang, M. & Zheng, D. (2013). Feature-rich segment-based news event detection on Twitter. In *International Joint Conference on Natural Language Processing*, (pp. 302–310). Nagoya, Japan.

Ritter, A., Etzioni, M., Etzioni, O. & Clark, S. (2012). Open domain event extraction from Twitter. In *KDD '12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1104-1112). New York: ACM..

Schober, M., F., Pasek, J., Guggenheim, L., Lampe, C. & Conrad, F G. (2016). Research synthesis. Social media analyses for social measurement. *Public Opinion Quarterly, 80*, 1, 180–211.

Schrodt, P., Shannon, D. & Weddle, J. (1994). Political Science: KEDS - A Program for the machine coding of event data. *Social Science Computer Review*, *12* (14), 561-517. https://doi.org/10.1177%2F089443939401200408

Statham, P. & Howard, T. (2013). Relating news analysis and public opinion: Applying a communications method as a "tool" to aid interpretation of survey results. *Journalism* 14 (6), 737-753.

Stathopoulou, T. (2005). Using text mining tools for event data analysis. In S. Sirmakessis, (Ed.), *Knowledge Mining.* Series: Studies in Fuzziness and Soft Computing (pp. 239–255). Berlin: Springer.

Stathopoulou, T. (2007a). *The identification of immigration issues through the use of event data.* Paper presented at the 2nd EASR conference of the European Survey Research Association. Prague.

Stathopoulou T. (2007b). Event Data in the European Social Survey: Problems of recording and analysis. In P. Kafetzis, T. Maloutas, and J. Tsiganou, (Eds.), *Politics, Society, Citizens: The European Social Survey – ESS* (pp. 315–39). Athens: National Centre for Social Research. [In Greek]

Stathopoulou, T. (2008). *Why event reporting is important*. Presentation at the 4th ESS National Coordinators meeting. ESSi (European Social Survey Infrastructure - Improving Social Measurement in Europe) Program Warsaw. Poland.

Stathopoulou, T., Stasinopoulos, N., & Diakoumakos G. (2015). *Mapping "old" and "new" forms of protest in Greece.* Paper presented at the 12th European Sociological Association Conference. Prague, Czech Republic.

Stathopoulou, T. (2011a). *Greek event reports for Rounds 1-5 of ESS* -European Social Survey.

Stathopoulou, T. (2011b). *The climate of crisis. Monitoring events and attitudes in a cross-national perspective*. Paper presented at the 4th EASR (European Association for Survey Research) conference. Lausanne. Switzerland.

Stathopoulou, T. (2012). Historical conjuncture and social survey: the impact of events on attitude formation. In A. Maratou, M. Thanopoulou, A.Teperoglou, and E. Fronimou. (Eds.), *Sociology in Greece today*, Festschrift for I. Lambiri-Dimaki, (pp. 89–99). Athens: Sakkoulas. [In Greek]

Stathopoulou, T., Papageorgiou, H., Papanikolaou, K. & Kolovou, A. (2018). Exploring the dynamics of protest with automated computational tools. A Greek case study. In C. M. Stuetzer, M. Welker, & M. Egger, (Eds.), *Computational Social Science in the Age of Big Data. Concepts Methodologies, Tools, and Applications.* German Society for Online Research (DGOF), pp. 326-355, Köln: Herbert von Halem Verlag.

Stathopoulou, T. (2018). (Ed.). *Transformations of protest in Greece*. Athens: Papazisis Publishers and National Centre for Social Research. In press.

Stoop, I. (2004). *Event data collection Round 2 - Guidelines for ESS National Coordinators*. Retrieved from https://www.europeansocialsurvey.org/docs/round2/methods/ESS2_event_reporting_guidelines.pdf

Stoop, I. (2007). If it bleeds, it leads: the impact of media-reported events. In R. Jowell, C Roberts, R. Fitzgerald and E. Gillian. Measuring attitudes cross-nationally. Lessons from the European Social Survey (pp. 95-111). London: Sage.

Strohmaier M. & Wagner C. (2014) Computational Social Science for the World Wide Web, *IEEE Intelligent Systems 29*(5): 84-88.

Zachary, C. S., Mocanu, D., Vespignani, A. & Fowler, J. (2015). Online social networks and offline protest. *EPJ Data Science*, 4(19). Springer Open.

Walther M. & Kaisser M. (2013) Geo-spatial event detection in the Twitter Stream. In P. Serdyukov. et al. (eds): *Advances in Information Retrieval. ECIR 2013. Lecture Notes in Computer Science,* vol 7814. Berlin, Heidelberg: Springer.

Wang, Y., Fink, D. & Agichtein, E. (2015). SEEFT: Planned Social event discovery and attribute extraction by fusing Twitter and Web content. International AAAI Conference on Web and Social Media**,** North America. Retreived from: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10509/10531.

Wang W. & Zhao D. (2012) Ontology-based event modeling for semantic understanding of chinese news story. In M. Zhouet al. (eds.): *Natural Language Processing and Chinese Computing. Communications in Computer and Information Science*, vol 333. Berlin, Heidelberg: Springer.

Weng, J. & Lee, B. (2011). Event detection in Twitter. International AAAI Conference on Web and Social Media**,** North America. Retrieved from: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2767

Wuest, B., Rothenhäusler, K. & Hutter, S. (2013). Using computational linguistics to enhance protest event analysis. Paper presented *at the annual conference of the Swiss Political Science Association*. Zurich: University of Zurich.

Appendix

A.1 Guideline for the Interview to Evaluate Past Approaches in the ESS

We are working on a proposal for the evaluation of the tools for monitoring national (MNC) contexts (media claims/event data) developed so far in the ESS. As you were previously engaged in national context reporting we would be interested in your experience regarding assessment of the claim/event reporting in the ESS.

1. General experiences with reporting

2. Usability of the tools for monitoring national contexts (MNC)

   - Usefulness of MNC for the analysis of ESS data
   - Advantages

3. Coding procedure

   - General (evaluating the measurement of national contexts in the ESS so far)
   - Methods/tools used to code media claims/events
     - Experience with coding
     - Problems accessing media content (newspapers and/or online material)
     - Practices/methods: most important aspect of MNC that should be addressed from now on
     - Costs
     - Time
     - Homogenization of claims
     - Visualization of claims

   - Evaluation of media claims coding
     - Usage of the coding scheme provided by ESS
     - Usability: Positive experiences
     - Problems
     - Evaluation of the instructions (were the instructions clear enough/simple to follow)

4. Country/culture specificities

5. Criteria choosing the media used

   - Print and online media in your country (How would you describe the media (print and online) landscape in your country?)

6. Suggestions for future development of the MNC

   - Experiences with automated coding – machine learning

- Knowledge of any relevant automated approach that could be used by NCs
- Use of social media content to retrieve claims/data for MNC

A.2 Tools for (Semi)-Automatic Claim Detection

Tool	Description	Pros	Cons
TABARI	Uses *actors* (proper nouns that identify political actors), *verbs*, and *phrases* (used to distinguish different meanings of a verb) and employs several general English grammatical rules	Very fast processing due parsing rather sparsely than making a full syntactical analysis	Not cross-lingual and relies on text dictionaries that have to be updated regularly (not updated as PETRARCH and PETRARCH2 are successors)
PETRARCH	Uses CoreNLP to distinguish nouns and verbs in phrases by looking up terms in a set of verb, actor, agent, issue and discard dictionaries (available on GitHub).	More accurate than TABARI and modules like geotagging can be easily added	High computational power required due to the deep parse with CoreNLP. Also, not cross-lingual and dictionaries have to be updated regularly.
PETRARCH2	Stores Penn TreeBank data from CoreNLP as a tree of linked nodes, where each node is a phrase object. This allows 1) PETRARCH2 to be easily adapted to other languages, and 2) for identifying any noun phrase that could potentially be a political actor.	Has simpler verb dictionary PETRARCH which increases speed and is more robust. The TreeBank tree structure makes matching patterns more sufficient and adaptable to other languages	Dictionaries still have to be updated, last update was in 2016, and the tool is command-line only, thus difficult to install and run by non-informatics
PALOMAR	Uses the ELK stack for indexing, searching and analyzing data, which enables an efficient event extraction steps with the steps: 1) establishing a coding framework, 2) data collection (from news media and social media), 3) NLP based data pre-processing to produce a set of annotations for POS tags and Named Entities, and 4) classification into categories such as person, organization, location, facility, time, and issue.	Can be applied to every event coding scheme and also across different languages and data sources	The coding framework has to be developed first, which takes a lot of effort and must be repeated for every new language. Not production-ready yet.
BBN Accent	Based on BBN SERIF: an NLP analysis engine that extracts structured information (e.g. entities, relationships, and events) from text.	None found	Not transparent regarding the exact methodology and there is no publicly available detailed description of the method

Phoenix Pipeline (PHOX)	Links a web scraper, CoreNLP, and PETRARCH (version 1 or 2) together as a pipeline to make the entire process easier from beginning to end.	Once installed it is easy to handle and generates reliable results.	It cannot be handled by a single consumer computer, but rather needs a high performance computer to process real time data.
EL:DIABLO	A virtual machine (running in e.g. VirtualBox) including a web scraper and the Phoenix pipeline, the web scraper runs once an hour and the Phoenix pipeline runs once a day.	It is production-ready and quite easy to handle	Not easy at all to install without informatics expertise and it is not helpful if one wants to process an existing set of texts. Also, not cross-lingual
hypnos	Uses the two minimal components of EL:DIABLO: the event coder of PETRARCH2 and CoreNLP in a REST API, which allows users to make HTTP requests.	Same as EL:DIABLO	Same as EL:DIABLO
Jabari NLP	Java implementation of TABARI, but improved by NLP concepts with three dictionaries to aid in event coding: actor, agent, and verb dictionary.	Improved techniques with NLP concepts in comparison with TABARI: Shallow parsing allows to use just enough of the NLP capability to produce good results, while keeping the computational complexity low.	Same as TABARI and there is no detailed description of the methodology available
GDELT Analysis Service / Google BigQuery	Uses the CAMEO codebook and a global knowledge graph (GKG) to provide data about the connections of all persons, organizations, locations, themes, counts, events, and sources each day (updates live every 15 minutes)	Cloud-based analysis service with 14 tools for geographic, temporal, network, and contextual visualizations of the Event Database and the Global knowledge graph. No need for installation and a large amount of events captured (due to massive data collection)	Not open source, thus not transparent which methodology used. No information about how accurate GDELT's model performs and only base features are free to use
Virtual Research Associates (VRA) Reader	Extracts the first sentence from Reuter Business Briefing (RBB) articles and quantitatively summarizes all the events that are described in the lead	None found	No information available and not maintained anymore (i.e. website down). It also only works with Reuters Business Briefing News stories

Twied	Collects Tweets based on the Twitter API and infers locations based on gazetteers, clusters the tweets and then uses a relevance classifier to store relevant events in an event database	Can detect tweets that are relevant tot he public, most probably based on retweets and likes	Not much information available about how exactly it works, especially the relevance classifier and no scientific references found.
Lydia	Tracks the temporal and spatial distribution of entities in news media. It obtains newspaper text via crawling and parsing methods, performs a named entity recognition and identifies which other entities occur near it, and uses a temporal and spatial analysis to establish a relative frequency given entities that are mentioned in different news sources.	Does not rely on an ontology but works with spatial and temporal analysis	Depends on a database of entities (for named entity recognition) that needs to be maintained regularly, and for every language another database is needed.
Wikipedia Live Monitor	Monitors article edits on different language versions of Wikipedia. When concurrent edits peak to a certain point, it is an indication for potential breaking news. Plausibility of potential breaking news is then checked with full-text cross-language searches on multiple social networks (Twitter, Facebook , Instagram, YouTube, Flickr, MobyPicture, TwitPic, Wikimedia Commons, and Google+).	Unlike the reverse approach of monitoring social networks first, and potentially checking plausibility on Wikipedia second, this approach is less prone to false-positive alerts, while being equally sensitive to true-positive events. It also has relatively low processing costs and almost real-time coverage of national and especially international events	Plausibility check is not fully automated and needs manual input. It is also hard to detect the actual relevant events. Last but not least, it is still in development, such that there is no API yet, that can export the results for a given period of time