Armed Conflicts in Online News: A Multilingual Study

Robert West*
School of Computer and Communication Sciences
Ecole polytechnique fédérale de Lausanne
robert.west@epfl.ch

Jürgen Pfeffer
Bavarian School of Public Policy
Technische Universität München
juergen.pfeffer@tum.de

Abstract

Wars and conflicts have constituted major events throughout history. Despite their importance, the general public typically learns about such events only indirectly, through the lens of news media, which necessarily select and distort events before relaying them to readers. Quantifying these processes is important, as they are fundamental to how we see the world, but the task is difficult, as it requires working with large and representative datasets of unstructured news text in many languages. To address these issues, we propose a set of unsupervised methods for compiling and analyzing a multilingual corpus of millions of online news documents about armed conflicts. We then apply our methods to answer a number of research questions: First, how widely are armed conflicts covered by online news media in various languages, and how does this change as conflicts progress? Second, what role does the level of violence of a conflict play? And third, how well informed is a reader when following a limited number of online news sources? We find that coverage levels are different across conflicts, but similar across languages for a given conflict; that Middle Eastern conflicts receive more attention than African conflicts, even when controlling for the level of violence; and that for most languages and conflicts, following very few sources is enough to stay continuously informed. Finally, given the prominence of conflicts in the Middle East, we further analyze them in a detailed case study.

1 Introduction

For millenia, wars and conflicts have been among the most influential events of human history. Wars have shaped our societies not only through the very facts they create, such as destruction, borders, and regimes, but also in less direct ways, through people's perception of the respective wars. For instance, the very tenets of modern German society are deeply rooted in how Germans see their country's role in two world wars; Turkey and Armenia have no diplomatic relations because the countries cannot agree on a common view of a conflict that happened 100 years ago; and millions of people have positive feelings toward Mexico's 1862 victory against France because it lets them eat nachos every May 5.

Although our perception of conflicts is so crucial, we rarely learn about them directly, but must rely on the news media instead. Journalists, however, create selection bias with their decisions about newsworthiness (Althaus et al. 2011), and their reporting is subject to description bias (Barranco and Wisler 1999). It has even been argued that the media have the power to provoke and escalate conflicts by steering public opinion, a process termed "agenda setting" or, more catchily, the "CNN effect" (Robinson 1999).

Since the media shape our perception of armed conflicts, and since our perception of conflicts shapes how we see the world, it is crucial to understand how armed conflicts are covered by the news media. Elucidating the underlying patterns and processes is difficult, however, because news comes as unstructured text, which is not straightforward to analyze, even in a given language. Moreover, news coverage might vary fundamentally across languages, so getting the full picture requires working in a multilingual setting. This is particularly hard if the researchers do not understand the languages being studied, and the reason why most previous research has focused on single languages.

To make statements about "the news media", one further needs to analyze *all* news outlets (or at the very least a representative sample). Obtaining such a dataset, and processing it at scale, poses challenges that have caused most prior work to focus on single or few news sources, thereby introducing sampling bias (Weaver and Bimber 2008).

Present work. Here we address these challenges by developing a set of unsupervised methods for compiling and analyzing a multilingual corpus of millions of online news documents about armed conflicts, and by applying our methods in a study of nine armed conflicts as covered by media outlets in 13 languages. Unsupervised methods are called for in this regime because the authors have no command of most of the 13 languages studied, such that hand-labeling examples for supervised machine learning would require hiring paid annotators, which is slow and expensive.

As put by Evans (2010), "Framing is manifested in, among other things, the amount of media coverage of a particular conflict and the language used to describe the actors and events in that conflict." Hence, we investigate both the amount of media coverage and the language used. We cluster our analyses around three core research questions:

RQ1 How widely are conflicts covered by online news in various languages, and how does this change with time?

*Research done mostly while at Stanford and TUM.

RQ2 What is the role of the level of violence of a conflict?

RQ3 How well informed is a reader when following a limited number of online news sources?

Summary of results. In summary, we observe that the amount of media coverage is rather different across conflicts, but similar across languages for a given conflict; that the Middle East receives significantly more attention than Africa, even for fixed levels of violence; and that for most languages and conflicts, following a handful of news sources suffices for reading *something* about the conflict on every relevant day, whereas staying informed about all, or particular, aspects of the conflict is much harder.

Overall, our main contributions are as follows:

- We develop a set of unsupervised methods for compiling and analyzing a multilingual corpus of millions of online news documents about armed conflicts (Sec. 3).
- We apply these methods to further our understanding of the nature of the media coverage of nine conflicts in 13 languages (Sec. 4).
- We conduct a detailed case study of four Middle Eastern conflicts to which the media pay particularly much attention (Sec. 5).

We start the paper by introducing our data sources (Sec. 2) and conclude by pointing out limitations and discussing our results in the context of related work (Sec. 6).

2 Datasets

2.1 Armed conflicts

Various endeavors in the field of peace and conflict studies aim to document armed conflicts in structured event databases. In these projects, human experts manually extract and code detailed event information (*e.g.*, type of event, location, date, involved parties, number of casualties) from secondary sources such as news reports. For instance, the Uppsala Conflict Data Program (UCDP) strives to cover all conflicts worldwide (Sundberg and Melander 2013), whereas the Armed Conflict Location and Event Data Project (ACLED) focuses on Africa and South/Southeast Asia (Raleigh et al. 2010), and the Iraq Body Count (IBC) on Iraq (Hsiao-Rei Hicks et al. 2009).

In this research, we are specifically interested in conflicts that started within the time range of our document corpus (*i.e.*, after June 2010; Sec. 2.2), as this allows us to track the news coverage they have received from the very beginning. We further concentrate on conflicts that are ongoing at the time of writing. About 15 conflicts—all of them in Africa, the Middle East,[1] or Ukraine—meet these criteria (Uppsala Conflict Data Program 2017), and we focus on nine of them that represent all three above geographical regions and span a wide range of intensity (Table 1). Note that we also include Iraq, although this conflict started as early as 2003, the reason being that the high resolution and quality of the Iraq Body Count (*cf.* above) lets us study the relationship between media coverage and casualty counts particularly well.

[1] We include Arab North Africa in the term "Middle East".

	Conflict	Region	Onset	Casualties
IQ	Iraq	Middle East	2003-03-20	28K
EG	Egypt	Middle East	2011-01-25	1K
YE	Yemen	Middle East	2011-01-27	10K
LY	Libya	Middle East	2011-02-15	5K
SY	Syria	Middle East	2011-03-15	121K
ML	Mali	Africa	2012-01-16	2K
CF	Centr. Afr. Rep.	Africa	2012-12-10	6K
MZ	Mozambique	Africa	2013-04-03	$<$100
UA	Ukraine	Europe	2013-11-21	6K

Table 1: Armed conflicts considered in this research. Onset dates are approximate and compiled from Wikipedia (2017). Casualty numbers are estimates from UCDP.

	Language	News sites	News docs	Conflict docs
en	English	2,671	339.9M	17,906K
de	German	336	61.0M	3,399K
es	Spanish	404	53.4M	1,882K
fr	French	156	22.5M	1,539K
pt	Portuguese	88	14.9M	982K
nl	Dutch	90	13.5M	729K
tr	Turkish	37	7.6M	625K
it	Italian	133	12.4M	436K
ro	Romanian	38	4.8M	509K
hu	Hungarian	38	6.9M	369K
pl	Polish	25	5.7M	273K
id	Indonesian	55	8.6M	259K
sv	Swedish	41	3.6M	255K

Table 2: Basic statistics of our news corpus. *Conflict docs* refers to documents published on news sites and mentioning at least one of the armed conflicts listed in Table 1.

2.2 Document corpus

As a data source of online news, we leverage the online media aggregation service *Spinn3r* (2017), from which we have collected all documents since August 2008 (about 2.6 billion), published in 53 languages[2] on about 9.4 million Web domains. Besides the main text content, documents consist of a title, a URL, and an (approximate) publication date.

Here, we focus on documents published in 13 Latin-character languages (Table 2) during the five years from July 2010 to June 2015. We clean the data by deduplicating documents with identical content (even when published under different URLs) and discarding URLs that appear multiple times with different contents, as such URLs tend to be landing pages (*e.g.*, `www.cnn.com`) rather than news articles.

To assess the completeness of this dataset with respect to the entirety of online news, we compare it with a comprehensive list of 151K online news articles about Osama bin Laden's death indexed by Google News (Bharat 2011). This list contains URLs from 7,765 Web domains. The Spinn3r dataset contains documents from all of these domains, so we conclude that Spinn3r gives us broad coverage with respect to the entirety of online news and is thus well suited for studying the research questions considered in this paper.

[2] Languages were detected from character *n*-grams by a naïve Bayes classifier (Nakatani 2010).

3 Methodology

Next, we describe our method for extracting a multilingual corpus of news reports about armed conflicts from the Spinn3r dataset introduced above. We proceed in two steps: first we identify documents about conflicts (Sec. 3.1), then we identify documents from news sites (Sec. 3.2). While this order of steps might seem counterintuitive at first, it will become obvious later (footnote 4) why we proceed this way.

3.1 Identifying documents about armed conflicts

To identify documents about armed conflicts, we follow a two-step approach: (1) we detect country mentions and then (2) decide whether the country is mentioned in the context of the armed conflict happening there.

For step 1, we manually construct a regular expression (regex) for each of the nine countries of Table 1 in each of the 13 languages of Table 2; *e.g.*, Ukraine is matched by `ukrain.*` in English, and by `oekraï.*` in Dutch. This effort took one researcher a few hours and was feasible even without command of the respective languages.

Since not all mentions of a country are about the conflict there, step 2 strives to automatically identify those that are. The most obvious approach would be to train a classifier using supervised machine learning. While straightforward and likely to succeed, this approach is impractical in our setting because it would require manually labeling hundreds of training documents per language. This is time-consuming and expensive, especially since the authors do not understand most of the 13 languages and would therefore need to enlist paid annotators.

We therefore develop a cheaper, unsupervised approach, inspired by prior work on unsupervised sentiment analysis (Turney 2002). We start with a high-precision classifier that labels a document as certainly conflict-related if its title matches a hand-crafted regex.[3] We created one regex per language, which took about one day all in all and was feasible even for languages not spoken by the authors, with the help of online dictionaries and Wikipedia.

Next, we compute, for each corpus word, the *normalized pointwise mutual information (NPMI)* (Bouma 2009) between the word appearing in a document and the same document's title matching the hand-crafted regex. NPMI measures how much more likely these two events are to co-occur, compared to a baseline that assumes the events to be independent, which provides us with a score for each word that captures how much the word is associated with conflict.

Equipped with this measure, we classify a document as referring to the conflict in country *C* if and only if it mentions *C* (step 1) and contains a word with NPMI above a threshold α in a 20-word window centered around a mention of *C*.

The reason for working with short windows rather than entire documents, is that, depending on the website, a single document may contain several news stories, and we want to label the document as related to a given conflict only if one of these stories refers to the conflict.

Using this approach, our classification problem reduces to choosing a single parameter α, which is easily done by inspection. We find that $\alpha = 0.1$ yields good results across three languages (achieving precision and recall of about 80% on each of English, German, and French) and therefore use this value for all 13 languages.

3.2 Identifying news websites

Our Spinn3r dataset contains many kinds of document, such as "social media, weblogs, news, video, and live web content" (Spinn3r 2017), whereas our focus is on online news, so we need a way to automatically identify news documents.

We proceed at the domain level, classifying either everything or nothing from the same domain as news. The most obvious approach would be to manually assemble a list of news websites; a slightly more complex approach would be to classify news *vs.* non-news domains using supervised machine learning. Unfortunately, both approaches are infeasible in our multilingual setting: while finding a precompiled list for a particular language might be possible, finding comparably comprehensive lists for all other 12 languages is difficult. Similarly, hand-labeling domains for training a supervised classifier is cumbersome even for languages understood by the researcher, let alone for other languages.

For these reasons, we again develop an unsupervised method. It works in two steps, (1) starting with a high-recall heuristic, then (2) increasing precision.

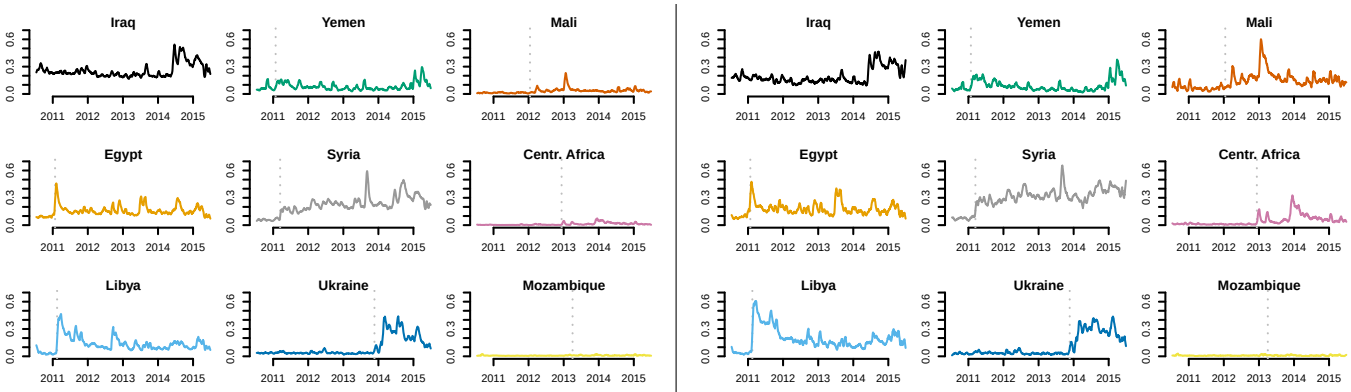
In step 1, assuming that every news outlet reported on Osama bin Laden's death (May 2, 2011), we find all domains in the Spinn3r dataset that mentioned his name (accounting for variations in spelling) on at least two days between May 1 and 5, 2011. As news websites tend to publish content frequently, we further include only domains from which we have documents on at least half of the days during our five-year period and from which we have at least five documents per active day in the median. Finally, we union the resulting set with the list of domains indexed by Google News mentioned in Sec. 2.2.

While step 1 can be expected to have high recall, it also yields many non-news domains that happened to mention bin Laden around the time he was killed. Hence, the objective of step 2 is to increase precision. It builds on the fact that different news websites follow similar temporal trends with respect to the content they publish—dictated by the events happening in the world—, whereas other types of website (such as blogs) are much less synchronized. Therefore, we aim to recognize news websites by the temporal patterns in which they mention external events such as armed conflicts.

We operationalize this intuition as follows. First, we represent each website by a binary vector with one bit per day and conflict that encodes if the site mentions the conflict that day.[4] We then stack all vectors as rows in a matrix, mean-center it, and run principal component analysis. Inspecting the results reveals that, indeed, the first principal component mirrors the overall daily frequencies with which the conflicts

[3] For instance, the English regex is the disjunction of `dead`, `kill.*`, `bomb.*`, `mortar.*`, `weapon.*`, `missile.*`, `war`, `wars`, `peace`, `ceasefire`, `truce`, `terror.*`, `fight.*`, `attack.*`, `combat.*`, `battle.*`, `soldier.*`, `milit.*`, `hostage.*`, `troop.*`, `rebel.*`, `army`, and `armies`.

[4] This is why we detect conflict mentions before news domains.



(a) English

(b) French

Figure 1: Coverage time series of all conflicts in two languages. The x-axes represent days, the y-axes the fractions of all news sources mentioning the conflict on day x (out of all sources active on day x). Dashed vertical lines mark the onset of the conflict.

are mentioned. Consequently, a website's score for the first principal component captures how much it is aligned to the overall "pulse of events", and thereby how "newslike" it is. So retaining only sites with a score above a threshold β lets us arbitrarily increase the precision of our initial high-recall filter; *e.g.*, the English websites with the highest score are `live5news.com`, `miamiherald.com`, and `abc40.com`; and those with the lowest score, `sampleaday.com`, `masnsports.com`, and `ninersnation.com`.

Automatically producing a list with 100% precision (*i.e.*, with no non-news sites) seems elusive, but in order to make cross-language comparisons, we certainly want to keep precision constant across languages. We achieve this calibration by choosing a separate threshold β per language, such that a fixed fraction γ of above-threshold sites mentions a given conflict during a given time range. (We set $\gamma = 0.6$, use Libya as the conflict, and the three days following the killing of the U.S. ambassador to Libya as the time range.)

Basic statistics of the resulting conflict news corpus are listed in Table 2. The number of news sites varies widely between languages, but we still have several millions of news documents even for the sparsest languages. About one in 20 news documents is about one of the nine conflicts we study.

3.3 Coverage time series

One of our fundamental objects of study will be what we refer to as *coverage time series*. We construct one coverage time series per language L and conflict C, as follows. For each day of our five-year period, we compute the fraction of all news sites of language L that published at least one document that day mentioning conflict C (where the fraction is taken with respect to the number of news sites of language L that published at least one document that day). Plotting this fraction as a function of time yields the respective coverage time series. For the sake of visual clarity, we process time series with Friedman's "super smoother" (Friedman 1984). Fig. 1, to be discussed later, contains examples.

Coverage time series are useful because they let us study how widely a conflict is covered by the media, and how this changes with time.

We emphasize that our notion of coverage is invariant to the number of articles published about a conflict by a given news source on a given day; all it cares about it is how many news sources cover the conflict in at least one article.

3.4 Topic modeling

At times, we will want to go beyond the mere binary notion of whether a conflict is covered or not, as captured by coverage time series, and analyze how the conflict is covered in terms of content. To facilitate meaningful insights, we represent documents by a small number K of topics, computed via latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) using Spark MLlib (Apache 2015). (We use $K = 5$ and $K = 20$; the parameters `docConcentration` and `topicConcentration` are each set to 1.1.)

As mentioned in Sec. 3.1, some documents encompass several news stories, and the standard representation of documents as bags of all their words would artificially mix the topics associated with different stories. To circumvent this problem, we work with short windows of text again (*cf.* Sec. 3.1): when considering a document in the context of conflict C, we consider the 20-word windows centered around all mentions of C and represent the document as the bag of words of the union of all windows. (We restrict ourselves to the 20K most frequent words in the corpus and discard stop words as well as names of conflict countries.)

In terms of output, LDA represents documents as distributions over the K topics, which are in turn represented as distributions over words. To have an intuitive handle of topics, we manually label them with succinct names by inspecting their corresponding word distributions.

LDA, being a fully unsupervised method, is well suited for this research. The only step to require an understanding of the respective language is the labeling of topics with names, which is easy and cheap, even if external annotators were to be paid. Therefore, this step, too, can readily scale to all languages (although, as a matter of focus, we here perform it in English only).

3.5 Coverage maximization framework

One of our research questions (RQ3) asks how many news sources a reader must follow in order to be well informed about a conflict. To formalize this question, we build a bipartite graph for each language and conflict, whose vertices represent news sources and the 365 days following the onset of conflict (one vertex for each day on which at least one source covers the conflict). We add an edge between a source and a day if the source mentions the conflict that day.

Given this graph representation, RQ3 corresponds to well-known combinatorial optimization problems: Determining the maximum fraction c of days that can be covered by choosing a fixed number k of sources is known as the MAXIMUM COVERAGE problem. Conversely, determining the minimum number k of sources required to cover at least a fixed fraction c of days is known as PARTIAL SET COVER.

While both problems are NP-hard, they are well approximated by a greedy algorithm, which yields a 63%-approximation for MAXIMUM COVERAGE, and a $\log n$-approximation for PARTIAL SET COVER, where n is the number of vertices representing days (Elomaa and Kujala 2010).

4 Results

4.1 RQ1: Media coverage over time

To answer RQ1—*How widely are conflicts covered by online news, and how does this change with time?*—, we begin with a discussion of coverage time series (Sec. 3.3).

Coverage time series. As we consider nine conflicts and 13 languages, there are 117 coverage time series. For space reasons, we depict only those for two exemplary languages, English and French (Fig. 1). Each of the remaining time series is summarized by computing its average over the 365 days following the onset of the respective conflict (Fig. 2).

The shape of time series is rather different across conflicts. Several conflicts see a spike of attention at their onset, followed by a decrease (Egypt, Libya, Ukraine), whereas other curves stay flat around the onset (Yemen; African conflicts in Mali, Central African Republic, Mozambique). Interestingly, media interest in some conflicts (*e.g.*, Mali; Central African Republic in French) spikes only later on.

A sudden increase in media interest tends to be followed by a gradual decrease. This may be explained by several reasons: the media might be losing interest over time, but the effect could also be explained by media-external factors such as a decrease in the inherent violence of the conflict. We aim to disentangle these two factors in RQ2 (Sec. 4.2).

A notable exception from the above observation can be found in Syria, whose time series spikes much less than those of other conflicts at the onset, but which then keeps growing for over four years, throughout our data period. This is especially interesting because the initial conditions of the Syrian conflict are rather similar to those of the other conflicts that started with the Arab Spring in 2011 (Egypt, Libya, Yemen), whose time series look very different. This observation led us to investigate the Middle Eastern conflicts in a more detailed case study, to be presented in Sec. 5.

Not only the shape, but also the amplitude, of coverage time series varies widely between conflicts. In particular, the Middle East and Ukraine attract much more attention than Africa. As Fig. 2 shows, this is true across languages.

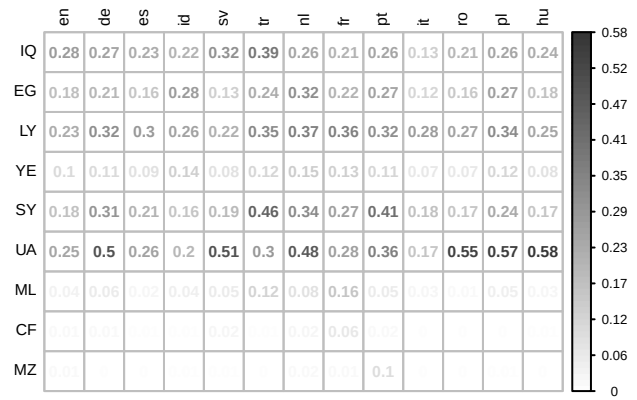


Figure 2: Summary of coverage time series (Fig. 1). Each value represents the average of the respective time series over the year following the onset of conflict (we use October 2013 for Iraq, when the conflict escalated into civil war).

Cross-language commonalities. We also measure the similarity of languages in a more fine-grained fashion in terms of the Pearson correlation of their respective time series. Concretely, we first compute a separate language-to-language correlation matrix for the time series of each single conflict, and then compute the pointwise average of all matrices to obtain an overall correlation matrix. Average correlations are very high, with a minimum of 0.73 and quartiles of 0.82, 0.87, and 0.91, so we conclude that, overall, different languages cover armed conflicts in similar ways.

Cross-language differences. Despite these similarities, there are several language-specific peculiarities. For instance, the francophone media care much more about Mali and the Central African Republic than other media do, mediated by France's colonial history and her military involvement in those conflicts (the spikes of French media interest in Mali and the Central African Republic coincide with the times France deployed troops there). Geographical proximity—and likely the associated fear of a potential spillover of conflict—plays a role as well, as evident in the elevated interest of Eastern European (Romanian, Polish, Hungarian) media in Ukraine, and of Turkish media in Iraq and Syria.

4.2 RQ2: Media coverage and violence of conflict

Above, we saw that some conflicts attract much wider media attention than others. One factor we expect to play an important role in this regard is the inherent violence of a conflict.

To quantify the *relationship between the violence of a conflict and the coverage it receives,* consider Fig. 3, which contains one data point for each conflict and year in English. Violence of conflict is shown on the x-axis in terms of the cumulative number of casualties caused by the conflict in the respective year, while the y-axis shows the fraction of mentioning sources (*i.e.*, the y-axes of Fig. 1) averaged over all days in the respective year.

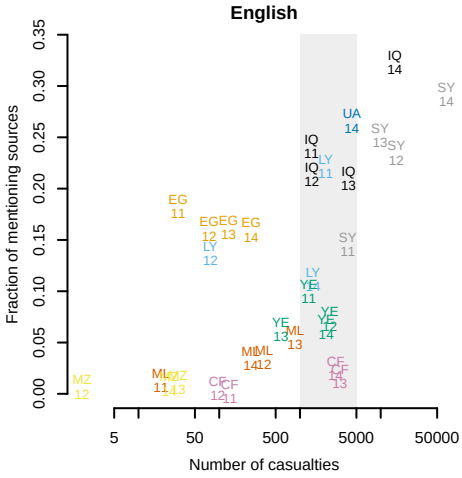


Figure 3: Casualty numbers (log axis) *vs.* English news coverage. Data points are conflict years (*cf.* Table 1 for abbreviations); news coverage computed as average of the respective coverage time series (Fig. 1) over the respective year.

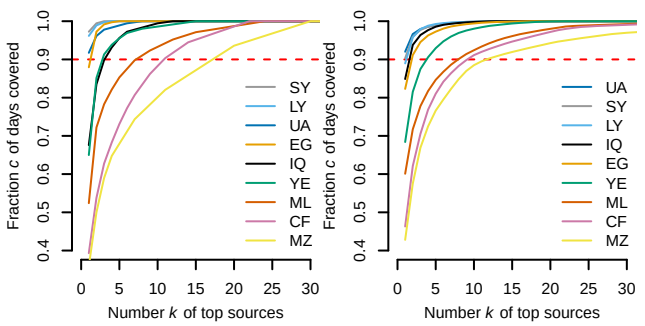
Inspecting the figure, we observe an overall positive correlation between inherent violence and media coverage. Moreover, this tends to also be the case when controlling for the conflict (consider, *e.g.*, the points for Syria or Iraq). Nonetheless, coverage levels can be quite different between conflicts even when their levels of violence are nearly identical. In particular, given a range of casualty numbers, Africa sees the lowest, and Ukraine and the Middle East the highest, levels of media coverage; *e.g.*, all of Ukraine, Iraq, Libya, Syria, Yemen, and the Central African Republic have incurred between 1,000 and 5,000 casualties in certain years (gray area in Fig. 3), but coverage ranges from 3% for the Central African Republic to 27% for Ukraine. Fig. 3 shows results for English only, but the same trends hold across languages, with the minor exception of French, which pays more attention to Mali and the Central African Republic, the former even being at par with the Middle Eastern conflicts.

4.3 RQ3: Coverage maximization

Above, we have considered all news websites, with the goal of understanding how widely conflicts are covered by the media, and how this varies with time and language. Next, we take the perspective of readers, who cannot possibly keep track of hundreds or thousands of news sources, but must instead focus their attention on just a few of them.

We ask: *How many news sources must one follow to be well informed about a conflict?* More technically, if a reader is able to follow a certain number k of news sources, what fraction c of all days of a conflict will they be informed about? Conversely, if a reader wants to be informed about a fraction c of all days of a conflict, how many news sources k must they follow? (We restrict ourselves to the first 365 days of each conflict.)

The optimization approach we adopt to answer these questions has been introduced in Sec. 3.5. Results are presented in Fig. 4, which contains one curve per conflict and



(a) Dutch (b) All languages

Figure 4: Fraction c of conflict days covered by the (approximately) optimal k news sources. One curve per conflict (*cf.* Table 1). *Left:* Dutch. *Right:* average curves across all languages. Legends are ordered according to the value at $k = 1$.

plots the number k of top sources on the x-axis, *vs.* the fraction c of days covered on the y-axis. We display the results for a single language—Dutch—in Fig. 4a, and averages over all languages in Fig. 4b.

We can see that, for most conflicts, extremely few sources suffice to cover a large fraction of all conflict days; *e.g.*, for five of the nine conflicts, a single source covers well over 80% of days, on average across languages (Fig. 4b).

The reason is that all languages contain a small fraction of sources that achieve large coverage on their own (whereas most sources have rather small coverage); *e.g.*, in English, these are big news outlets such as the BBC or the Washington Post, news agencies such as Reuters, as well as their "resellers" such as Yahoo! News.[5]

For space reasons, we cannot repeat Fig. 4a for all languages. To nonetheless grasp the full picture, we summarize each curve in one scalar value: we fix a coverage level c and compute the number k of sources necessary to achieve coverage c. We choose $c = 0.9$ here, corresponding to the dashed horizontal line in Fig. 4.[6] As summarized in Table 3, the trends identified in Fig. 4 hold for each language: very few sources tend to suffice for covering 90% of conflict days.

Cross-conflict source ranking. These numbers were obtained by optimizing for each language/conflict pair individually, thus obtaining one source ranking per pair. We now slightly modify this setup by pooling all conflicts for a given language. In the bipartite-graph formulation of Sec. 3.5, we replace days with day/conflict pairs and connect a source to a day/conflict pair if the source covers the conflict on the day. This lets us find a global, cross-conflict source ranking for each language. This global ranking will require at least as many sources for a given coverage level as the rankings optimized for single conflicts, and the respective increments are shown in parentheses in Table 3. Note that, in most

[5] As a robustness check, we manually constructed lists for the latter two kinds of domain for all languages, and even when discarding those domains, the results remain qualitatively unchanged.

[6] For robustness, we tried a wide range of values for c; the results are all highly correlated with those for $c = 0.9$ (Pearson/Spearman correlation 0.91/0.78 for values as low as $c = 0.6$).

	en	de	es	id	sv	tr	nl	fr	pt	it	ro	pl	hu	avg.
IQ	1 (+0)	2 (+0)	1 (+0)	2 (+0)	3 (+0)	1 (+0)	4 (+4)	1 (+0)	1 (+1)	3 (+1)	3 (+0)	2 (+1)	1 (+0)	1.9 (+0.5)
EG	1 (+0)	1 (+0)	1 (+0)	2 (+1)	6 (+1)	3 (+0)	2 (+0)	1 (+0)	1 (+0)	4 (+2)	6 (+1)	2 (+0)	1 (+0)	2.4 (+0.4)
LY	1 (+0)	1 (+0)	1 (+0)	2 (+1)	2 (+1)	2 (+0)	1 (+0)	1 (+0)	1 (+0)	2 (+0)	3 (+0)	2 (+0)	1 (+0)	1.5 (+0.2)
YE	1 (+0)	5 (+2)	3 (+2)	3 (+2)	7 (+1)	7 (+2)	3 (+4)	2 (+0)	2 (+0)	6 (+4)	9 (+1)	4 (+1)	2 (+1)	4.2 (+1.5)
SY	1 (+0)	1 (+0)	1 (+0)	3 (+2)	2 (+1)	1 (+1)	1 (+0)	1 (+0)	1 (+0)	2 (+1)	5 (+0)	2 (+0)	1 (+0)	1.7 (+0.4)
UA	1 (+0)	1 (+1)	1 (+0)	3 (+3)	1 (+0)	2 (+0)	1 (+1)	1 (+1)	1 (+0)	3 (+2)	1 (+0)	1 (+0)	1 (+0)	1.4 (+0.6)
ML	1 (+0)	17 (+14)	17 (+7)	9 (+1)	3 (+3)	7 (+0)	8 (+2)	1 (+0)	5 (+1)	13 (+8)	15 (+5)	8 (+0)	4 (+2)	8.3 (+3.3)
CF	4 (+1)	27 (+20)	20 (+6)	13 (+1)	6 (+2)	7 (+2)	11 (+6)	2 (+1)	13 (+1)	18 (+8)	4 (+3)	3 (+1)	5 (+0)	10.2 (+4.0)
MZ	2 (+0)	40 (+18)	46 (+21)	8 (+4)	6 (+1)	8 (+3)	18 (+4)	22 (+1)	1 (+1)	14 (+13)	8 (+6)	7 (+2)	5 (+2)	14.2 (+5.8)
avg.	1.4 (+0.1)	10.6 (+6.1)	10.1 (+4.0)	5.0 (+1.7)	4.0 (+1.1)	4.2 (+0.9)	5.4 (+2.3)	3.6 (+0.3)	2.9 (+0.4)	7.2 (+4.3)	6.0 (+1.8)	3.4 (+0.6)	2.3 (+0.6)	

Table 3: News sources required for covering 90% of all conflict days; "x ($+y$)" means we need x sources when using a ranking optimized for the respective conflict, and $x+y$ sources when using a ranking optimized for all conflicts simultaneously.

cases, using the global ranking does not require many more sources than the specialized ones; *i.e.*, a reader will generally be rather well informed about all conflicts even when using the same set of sources in all settings.

It is, once again, the African conflicts that constitute the main exception to the above findings: they (1) require more sources using specialized rankings, and (2) are significantly better covered by specialized rankings, compared to the global ones.

5 Case study: Middle East

As our above analysis of nine conflicts in 13 languages led us to conclude that the conflicts of the Middle East (Iraq, Egypt, Yemen, Libya, Syria) attract particularly much attention, we now revisit the three research questions of Sec. 4 with a focus on these conflicts. To gain deeper insights, we model the textual content of news reports using LDA (Sec. 3.4). While our techniques apply to all languages, we concentrate on English here for reasons of space.

5.1 RQ1: Media coverage over time

When studying temporal patterns of media coverage in Sec. 4.1, we made an interesting observation about the conflicts in Egypt, Yemen, Libya, and Syria: although all four conflicts were sparked in early 2011 under very similar conditions by the events of the so-called Arab Spring, their time series look rather different (Fig. 1). While Egypt and Libya's curves spike at the onset and then decrease gradually, Syria's does not spike but keeps increasing over many years, and Yemen's stays flat throughout. To better understand these differences is the goal of this subsection.

Source overlap. Given the different shapes of coverage time series, we are first interested in understanding if the four conflicts are covered by similar or different sets of news sources (called *source sets* here). We define the overlap of two sets as the fraction of elements of the smaller set that are also members of the larger set (*i.e.*, the maximum value of 1 is achieved if one set is a subset of the other), and measure source-set overlap for all pairs of the four conflicts for each day. We find the overlap during the first months of the Arab Spring to be very high, at about 80%, and therefore conclude that the difference in time series cannot be explained by the conflicts being of interest to different news outlets.

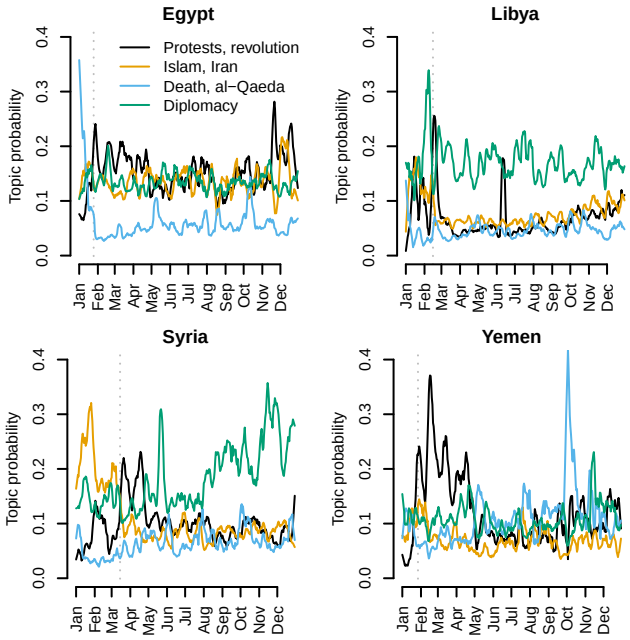


Figure 5: Probabilities of top 4 topics for Middle East in 2011. Sum is less than 1 because only 4 of 14 topics shown.

Topic analysis. Thus, we now focus on the textual content of news reports, modeling the topics they discuss via the LDA-based framework introduced in Sec. 3.4.

Since we are mostly interested in the differences in media coverage following the very onset of events, we focus on documents published during 2011 only. We sample 150K documents for each of the four conflicts, pool all 600K documents in one set used as the input to LDA, and extract 20 topics. Inspection of the word distributions associated with topics reveals that some topics capture similar concepts. Manually combining these yields 14 topics: *protests and revolution; Islam and Iran; death and al-Qaeda; diplomacy; information, media, and reporting; governments* (especially U.S. and Egypt); *Israel and Palestine; war; refugees;* as well as several conflict-specific topics.

Each document is represented as a distribution over these 14 topics. We visualize the prevalence of topics as time series in Fig. 5, where the probability of a topic on a day is

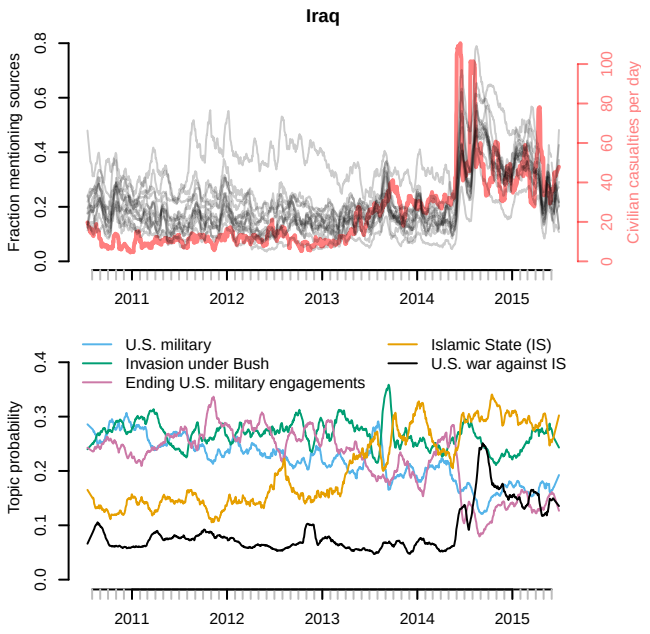


Figure 6: Iraq time series. *Top:* civilian casualties per day (red) and coverage time series (as Fig. 1; one gray line per language; outlier: Turkish). *Bottom:* topic probabilities.

computed by averaging over all documents, giving equal weight to all news sources. Across conflicts, four topics emerge as prevalent: *protests and revolution; Islam and Iran; death and al-Qaeda;* and *diplomacy.* For the sake of visual clarity, we plot only these four topics.[7]

Inspecting Fig. 5, we notice that *protests and revolution* is the most important topic everywhere at the very onset and decreases thereafter. The focus on this topic is especially strong for Yemen, whereas topics are more balanced for the other conflicts. The *diplomacy* topic is particularly important for Libya and Syria. For Libya, this is the case from the very onset, likely due to the international community's early involvement, with NATO airstrikes and Gaddafi's indictment by the International Criminal Court. For Syria, the *diplomacy* topic is less dominant during the first few months, but as the war drags on and escalates, and the international community becomes increasingly involved, the topic begins to dominate. This is accompanied by a growing interest in Syria on behalf of the media (Fig. 1).

5.2 RQ2: Media coverage and violence of conflict

Next, we revisit the question of the *relationship between media interest and the level of violence of a conflict,* as measured by casualty numbers. Fig. 3 answers this question for all nine conflicts, but at a rough, yearly granularity. We now focus on Iraq, which we can analyze at a much finer granularity because the Iraq Body Count project (Sec. 2.1) provides detailed records of civilian casualties at a daily level.

[7]Conflict-specific topics are also important, but they are not useful for comparing the four conflicts, so we ignore them in Fig. 5.

Fig. 6 (top) plots the Iraq Body Count time series of daily civilian casualty numbers in bold red, and that of media interest in thin gray (the gray curves are identical to the Iraq curves of Fig. 1, one gray curve per language). The red curve demonstrates that the Iraq conflict became much more violent starting in early 2013, primarily due to the involvement of the Islamic State (IS). This, however, did not immediately increase the number of news sources covering the conflict. It was only one year later, in mid-2014, when the conflict escalated into a full-fledged civil war and the U.S. launched airstrikes against the IS, that media interest picked up and started to be aligned with casualty counts.

The above observations are only with regard to the *amount* of media coverage, as captured by the fraction of all media outlets reporting on the conflict. For additional insights, we now investigate the *content* of media coverage during the same time. We do so using the same LDA framework as in Sec. 5.1. We process all English documents published about the Iraq conflict to extract five topics, which we hand-label as *U.S. military; invasion under Bush; ending U.S. military engagements* (in Iraq as well as Afghanistan); *Islamic State (IS);* and *U.S. war against IS.*

We turn topic probabilities into time series again as in Sec. 5.1 and visualize them in Fig. 6 (bottom). We observe that the *IS* topic started to rise in early 2013, synchronously with casualty counts (Fig. 6, top). In mid-2014, however, when the conflict escalated, and especially when the U.S. got involved, the *U.S. war against IS* topic made a leap, overtaking the topic of *ending U.S. military engagements*.

Interestingly, whereas the surge of the *IS* topic is not linked to a higher fraction of news outlets reporting on the conflict in Iraq, the surge of the *U.S. war against IS* topic is. This means that the core English media outlets swiftly paid attention to the increasing IS-inflicted violence (2013), whilst the overall number of English media outlets reporting on the conflict picked up only when the conflict exploded into war and the U.S. became involved militarily (2014).

5.3 RQ3: Coverage maximization

In Sec. 4.3, we saw that following just a few news sources is enough to be informed about all conflicts on all relevant days. Our notion of coverage, however, was binary and thus very simplistic there: as soon as a conflict was mentioned by a news outlet, we considered the conflict covered.

We now revisit the question of how many news sources one must follow to be well informed about a conflict by also taking into account what—not just whether—a news source writes about a conflict. In particular, we ask *how many news sources one must follow to be informed about a conflict with respect to all relevant topics* that are at play. Here, we define that a source covers a topic on a day if, in any of the documents the source published that day, the topic's probability lies above a threshold η. We use $\eta = 0.1$ here, which on average identifies the top two topics of a document.[8]

Given this definition, we adapt the bipartite-graph representation of Sec. 3.5 by replacing days with day/topic pairs, such that 100% coverage for a given conflict now requires

[8]Our analyses give similar results with $\eta = 0.05$ and $\eta = 0.2$.

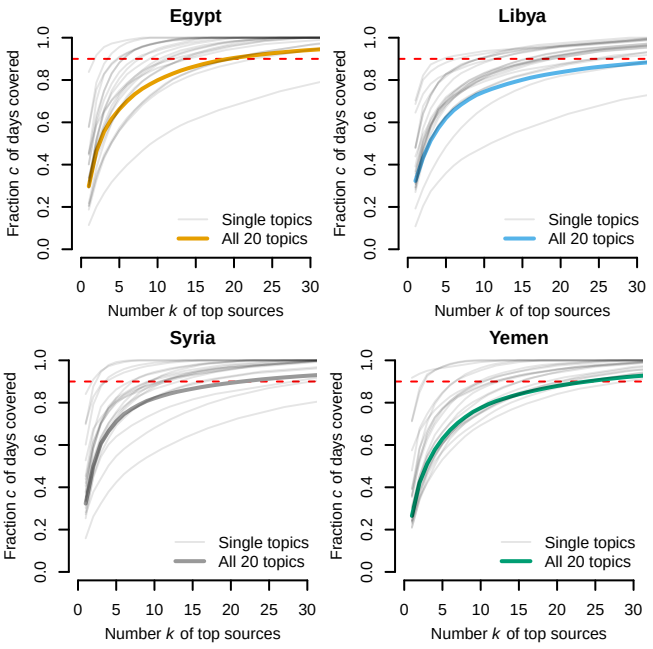


Figure 7: Fraction c of conflict days covered by the (approximately) optimal k news sources. *Bold:* coverage of all 20 topics. *Thin gray:* coverage of a single topic.

choosing a set of sources that cover all topics on all days on which the topic was mentioned in the context of the conflict.

Fig. 7 (same axes as Fig. 4) presents the results for the four conflicts for which we have already extracted topics in Sec. 5.1. Note that at least 20 sources are necessary to cover 90% of all day/topic pairs, many more than are required when ignoring topics: as shown by Table 3, a single source suffices to cover 90% of days in that scenario.

Moreover, even when considering single topics (the thin gray lines in Fig. 7), rather than all topics, we typically need many more than a single source to achieve good coverage.

In a nutshell, it is easy to read *something* about every conflict on every relevant day by following single news sources, but obtaining a faceted view of all—or even just specific—topics pertaining to the conflict requires following a much larger set of sources. That said, even this larger number—about 20 sources for 90% coverage—seems manageable for someone who is really interested.

6 Discussion and related work

We conclude by summarizing our findings, discussing them in the context of related work, and pointing out limitations as well as future work.

Related work. The media are of paramount importance for how conflicts are perceived by the general public. Thus, there is ample research about the media coverage of armed conflicts in the field of journalism studies (for an overview, *cf.* Allan and Zelizer 2004). One much-discussed topic is what is called "agenda setting" or the "CNN effect", *i.e.*, the media's power to provoke government interventions in conflicts (Robinson 1999) or influence the allocation of emergency relief funds to conflict countries (Jakobsen 2000).

Most research has focused on single (or very few) languages, conflicts, and news sources. Since a comprehensive list is not feasible here, we restrict ourselves to some select examples: Evans (2010) focuses on the New York Times' coverage of Palestine; Jasperson and Kikhia (2003) on CNN and Al-Jazeera's coverage of Afghanistan; and Robinson (2005) on two TV stations and two newspapers.

We, on the contrary, operate at a scale unprecedented in this context, considering media coverage of nine conflicts in a news corpus comprising 13 languages and representative of the entire media landscape in these languages. To be able to work in this regime, we employ unsupervised text analysis techniques that allow us to draw conclusions that go beyond single languages and conflicts, summarized below.

Summary of findings. By studying coverage time series **(RQ1),** we find that the patterns of media attention are rather different for different conflicts, but are similar for the same conflict in different languages. Notable exceptions may be explained by historical ties between the countries of the conflict and of the reporting media (*e.g.*, Mozambique is especially prominent in the Portuguese media), by current military involvement (*e.g.*, Mali and French media), or by geographical proximity (*e.g.*, Syria and Turkish media).

We further find that, while media attention is overall correlated with the violence of a conflict **(RQ2),** some conflicts receive much more attention than others even when fixing a level of violence, with Africa ranking much lower than Ukraine and the Middle East. Our fine-grained analysis of Iraq yields that the core English media shift attention to the Islamic State as soon as the latter emerges, but the overall number of media outlets reporting on the Iraq conflict does not increase until much later, when the country slides into an all-out war that eventually prompts a U.S.-led military intervention. It is an interesting direction of future research to investigate the role of such catalytic events in sparking media interest. In other words, is a certain "activation enthalpy" necessary to kickstart sustained media interest?

Finally, we discover that a handful of news source tend to suffice to inform a reader about all conflicts on most relevant days **(RQ3).** Our textual analysis reveals that it is, however, much more challenging to cover all, or specific, topical aspects of a conflict by following only a few sources.

The case of Africa. All our analyses point to the conclusion that the African conflicts receive overall considerably less media attention than the Middle East and Ukraine. As this remains true even when controlling for the number of casualties incurred, we may say that Africa receives less media attention per death. Consequently, the number of news sources a reader must follow to be well informed about African conflicts is much higher than for the other conflicts. Moreover, following specific news sources dedicated to the region of interest is more useful for Africa than elsewhere; *e.g.*, the sites `afrik.com` and `camer.be` provide better coverage of Africa in French than mainstream outlets.

The lack of interest in Africa is mitigated in languages spoken in countries with historical, cultural, or political ties

to the country of conflict (*e.g.*, French for Mali and the Central African Republic, and Portuguese for Mozambique), especially once the country of the media outlet gets itself involved in the conflict militarily (*e.g.*, France in Mali and the Central African Republic).

Future work should further investigate the case of Africa, in particular by analyzing the text of news articles to understand which particular aspects the media fail to cover.

Limitations and further future work. The breadth of our analysis—nine conflicts and 13 languages—necessarily comes with certain limitations. For instance, the basic notion of coverage we adopt here is binary and thus rather coarse: either a news source mentions a conflict on a given day, or it does not. We do increase the granularity to the topical level (Sec. 5), but the topics emerging from our LDA framework are broad and therefore not suitable for tracking mentions of specific persons, places, *etc*. Furthermore, our framework currently extracts topics for separate languages. In the future, it would be interesting to use cross-lingual topic models (Jagarlamudi and Daumé III 2010) to extract equivalent topics across languages, which would allow us to analyze to what extent different languages focus on different aspects of a given conflict.

Another limitation stems from our partitioning of news outlets by language, rather than by country. We do so because labeling websites with countries of origin turned out to be considerably harder than labeling documents with languages. Future work should, however, address this challenge, since it would enable more fine-grained insights.

Our research raises numerous further questions. For instance, the amount of content each media outlet can publish is limited. How does this affect the coverage of simultaneous conflicts? What is the relationship between mainstream news, as analyzed here, and social media, which have been used extensively during the Arab Spring (Khondker 2011) and by the Islamic State (Friis 2015)? How do the news production patterns studied here relate to news consumption?

Conclusion. To conclude, this paper contributes to our understanding of media coverage of armed conflicts, by developing a set of methods for multilingual document analysis and applying it to a large corpus of dozens of millions of news reports in 13 languages about nine conflicts. We hope our work will inspire other scholars to follow up on the new research questions raised by our results and to apply our framework to other kinds of event beyond armed conflicts.

Acknowledgments. We thank Patrick Mello of TUM for helpful discussions; Andrej Krevl, Peter Kacin, and Jure Leskovec of Stanford for technical support; and Spinn3r for data.

References

Allan, S., and Zelizer, B. 2004. *Reporting war: Journalism in wartime*. Routledge.

Althaus, S. L.; Swigger, N.; Chernykh, S.; Hendry, D. J.; Wals, S. C.; and Tiwald, C. 2011. Assumed transmission in political science: A call for bringing description back in. *Journal of Politics* 73(04):1065–1080.

Apache. 2015. Spark MLlib clustering. `http://bit.ly/2iXuEcq`.

Barranco, J., and Wisler, D. 1999. Validity and systematicity of newspaper data in event analysis. *European Sociological Review* 15(3):301–322.

Bharat, K. 2011. Google News and the coverage of bin Laden. `http://bit.ly/2iXFfEo`.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proc. International Conf. of the German Society for Computational Linguistics and Language Technology*.

Elomaa, T., and Kujala, J. 2010. Covering analysis of the greedy algorithm for partial cover. In *Algorithms and Applications*. Springer.

Evans, M. 2010. Framing international conflicts: Media coverage of fighting in the Middle East. *International Journal of Media & Cultural Politics* 6(2):209–233.

Friedman, J. H. 1984. A variable span smoother. Technical Report 5, Stanford Laboratory for Computational Statistics.

Friis, S. M. 2015. 'Beyond anything we have ever seen': Beheading videos and the visibility of violence in the war against ISIS. *International Affairs* 91(4):725–746.

Hsiao-Rei Hicks, M.; Dardagan, H.; Guerrero Serdán, G.; Bagnall, P. M.; Sloboda, J. A.; and Spagat, M. 2009. The weapons that kill civilians: Deaths of children and noncombatants in Iraq, 2003–2008. *New England Journal of Medicine* 360(16):1585–1588.

Jagarlamudi, J., and Daumé III, H. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Proc. European Conference on Information Retrieval*.

Jakobsen, P. V. 2000. Focus on the CNN effect misses the point: The real media impact on conflict management is invisible and indirect. *Journal of Peace Research* 37(2):131–143.

Jasperson, A., and Kikhia, M. 2003. CNN and Al-Jazeera's media coverage of America's war in Afghanistan. In *Framing terrorism: The news media, the government, and the public*. Routledge.

Khondker, H. H. 2011. Role of the new media in the Arab Spring. *Globalizations* 8(5):675–679.

Nakatani, S. 2010. Language detection library for Java. `http://bit.ly/2idWJxm`.

Raleigh, C.; Linke, A.; Hegre, H.; and Karlsen, J. 2010. Introducing ACLED: An armed conflict location and event dataset. *Journal of Peace Research* 47(5):651–660.

Robinson, P. 1999. The CNN effect: Can the news media drive foreign policy? *Review of International Studies* 25(02):301–309.

Robinson, P. 2005. *The CNN effect: The myth of news, foreign policy and intervention*. Routledge.

Spinn3r. 2017. Website. `http://bit.ly/2j34zeu`.

Sundberg, R., and Melander, E. 2013. Introducing the UCDP Georeferenced Event Dataset. *Journal of Peace Research* 50(4):523–532.

Turney, P. D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proc. Annual Meeting of the Association for Computational Linguistics*.

Uppsala Conflict Data Program. 2017. Website. `http://ucdp.uu.se`.

Weaver, D., and Bimber, B. 2008. Finding news stories: A comparison of searches using LexisNexis and Google News. *Journalism and Mass Communication Quarterly* 83:515–530.

Wikipedia. 2017. List of ongoing armed conflicts — Wikipedia, the free encyclopedia. (Online; accessed 2017-01-05).