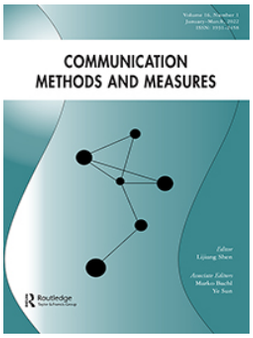

Routledge
Taylor & Francis Group

Communication Methods and Measures

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hcms20

Linking event archives to news: a computational method for analyzing the gatekeeping process

Kasper Welbers, Wouter Van Atteveldt, Joe Bajjalieh, Dan Shalmon, Pradnyesh Vineet Joshi, Scott Althaus, Chung-Hong Chan, Hartmut Wessler & Marc Jungblut

To cite this article: Kasper Welbers, Wouter Van Atteveldt, Joe Bajjalieh, Dan Shalmon, Pradnyesh Vineet Joshi, Scott Althaus, Chung-Hong Chan, Hartmut Wessler & Marc Jungblut (2022) Linking event archives to news: a computational method for analyzing the gatekeeping process, Communication Methods and Measures, 16:1, 59-78, DOI: 10.1080/19312458.2021.1953455

To link to this article: https://doi.org/10.1080/19312458.2021.1953455

Published online: 22 Aug 2021.

COMMUNICATION METHODS AND MEASURES
2022, VOL. 16, NO. 1, 59–78
https://doi.org/10.1080/19312458.2021.1953455

Routledge
Taylor & Francis Group

Linking event archives to news: a computational method for analyzing the gatekeeping process

Kasper Welbers[a], Wouter Van Atteveldt[a], Joe Bajjalieh[b], Dan Shalmon[b], Pradnyesh Vineet Joshi[b], Scott Althaus[b], Chung-Hong Chan[c], Hartmut Wessler[c], and Marc Jungblut[d]

[a]Department of Communication Science, VU University Amsterdam, Amsterdam, Netherlands; [b]Cline Center for Advanced Social Research, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; [c]Institute for Media and Communication Studies, University of Mannheim, Mannheim, Germany; [d]Department of Media and Communication, Ludwig Maximilian University of München, Munich, Germany

ABSTRACT

Digital archives that document real-world events provide new opportunities for large-scale analyses of how news coverage represents reality. We present a method and open-source tool for linking event data to news articles, and demonstrate its application with an analysis of event and country level predictors of terrorism coverage in *The Guardian* from 2006 to 2018, using event data from the Global Terrorism Database (GTD). Our method builds on established techniques for calculating document similarity, and we propose a novel strategy for fine-tuning parameters of the event matching algorithm that requires no manual coding. An online appendix is provided that documents all code to replicate our analysis and reuse our tools.

Introduction

Of all real-world events, the news only covers those that are believed to be the most important and interesting for the audience (Galtung & Ruge, 1965; Harcup & O'neill, 2017; Shoemaker & Vos, 2009). This selection process serves a critical social purpose, because it reduces a vast and complex reality into a small number of bite-sized stories that people are capable and willing to consume. However, the distorted picture of reality in the news can also distort citizens' worldviews, with potentially harmful consequences (Gadarian, 2010; Iyengar, 1990; Lewis, 2012). In this paper, we propose and validate a computational method and tool that enables a large-scale analysis of this distortion at the level of specific news events.

In order to accurately understand the factors that cause certain events to be covered and others to be ignored, we need to relate news coverage to information regarding the real-world occurrence of events (Rosengren, 1970). Today, there are various data sources that document individual events on topics such as disasters, crime and terrorism. These databases typically report various features that would be reported in news coverage, most importantly the type of event, location, and the people and organizations that are involved. Using computational text analysis, we can use these textual descriptions and the time of the event to link individual events to individual news articles. We propose a method and tool for performing this task, which we demonstrate by linking 112,091 events from the Global Terrorism Database (GTD) to 446,612 articles from *The Guardian* between 2006 and 2018. We then show how these linked event-article data can be used to analyze news selection in a way that aligns closely with the conceptual model underlying gatekeeping theory, building on Soroka's (2012) conceptualization of gatekeeping as a function.

CONTACT Kasper Welbers k.welbers@vu.nl VU University Amsterdam, Amsterdam, Netherlands

Our method builds on established techniques for measuring the similarity of texts, but is specialized for the application of matching event data to news content. We also propose a novel technique for estimating parameters for the matching algorithm. Specifically, we demonstrate that this can be used to determine a good threshold for how similar an event and news article need to be in order to consider them a match. Normally, determining this threshold requires hard to obtain gold standard data (see, e.g., Burggraaff & Trilling, 2020; Linder et al., 2018; Welbers et al., 2018), but our fully automated technique shows promising results as an alternative or complementary approach. To validate our method and analysis, we manually coded a gold standard of 500 news articles for references to GTD events, and performed an additional manual validation of the precision of 100 matches. Our tools are published in an R package,[1] and an online appendix is provided that contains code for replicating our analyses.[2]

Analyzing the Gatekeeping Function with Event Data

Even the thickest newspaper can contain only a very limited perspective on a very specific selection of real-world events. The process through which reality is transformed into this set of messages has been conceptualized as *gatekeeping* (Lewin, 1947; Shoemaker et al., 2001). In its simplest form, this process involves a single gatekeeper that, given a certain information input, produces a certain output. In the seminal gatekeeping study by David Manning White (1950), the gatekeeping choices of a single editor are studied by analyzing which of the news stories are selected to be "in" or "out." However, this is still only a single gate. The full gatekeeping process involves many interdependent gatekeepers, whose selection choices are affected by a diversity of values, routines, and organizational and institutional constraints (Shoemaker & Vos, 2009).

Empirically analyzing every gate and how the gatekeepers act in concert is practically infeasible. Yet, we can study the outcome of this complicated process by comparing the picture of reality in the media to real-world data (Rosengren, 1970). This type of analysis cannot directly tell us *why* specific events are reported, but it can tell us what types of events are more likely to become news. In other words, we can see which events are "in" or "out," but we don't know the selection criteria of each individual gatekeeper involved in this selection. One of the first explorations of how certain types of "world events" are more likely to pass through a chain of gatekeepers and become part of the "media image" was the paper by Galtung and Ruge (1965) on the structure of foreign news coverage. As scholars pursued this line of inquiry, they identified various elements that determine the newsworthiness of events, often referred to as *news values* (Harcup & O'neill, 2017; Schultz, 2007).

A novel conceptual framework for comparing news content to real-world data was proposed by Soroka (2012). Soroka conceptualized gatekeeping as a function, that takes an information distribution in reality as input, and outputs the information distribution of news reports. As an example, Soroka compares the distribution of real-world monthly unemployment rate to the distribution of monthly sentiment (i.e., a scale of negative to positive) scores in news about employment. By obtaining data about these two distributions, we can understand the gatekeeping process by estimating the function that transforms the former into the latter. Soroka coined this as a "distributional perspective" (p. 516) to gatekeeping.

The greatest strength of this distributional perspective is that it makes a clear and elegant connection between gatekeeping theory and analysis. It sticks close to the definition and metaphor of the gatekeeper, but "provides a model that can in principle be applied to real data, in a roughly comparable way across media outlets and countries and policy domains" (Soroka, 2012, p. 518). The idea of conceptualizing the gatekeeping process as a function that transforms input into output is not new, but by measuring reality and news coverage as distributions it becomes possible to actually calculate the gatekeeping function. Hypotheses regarding the gatekeeping process can then be expressed unambiguously in terms of the parameters of this function.

In this paper, we propose an extension of this approach that identifies the gatekeeping function based on an *event perspective*. Instead of distributions as input and output, the input is a set of real-

world events, and the output is a subset of these events reported in the news. This perspective aligns closely with the classic definition and metaphor of gatekeeping as a process that determines what information passes through a gate (Shoemaker & Vos, 2009; White, 1950). Our contribution lies in showing how we can obtain the data about the input and output events that are required to approximate this gatekeeping function.

Linking Event Databases to News Articles

Our method leverages the fact that today various types of events are documented in large digital archives. There are multiple databases for different types of events, and these are often maintained with active assistance of scholars and organizations that go through great lengths to standardize compilation and validation. These databases are generally not designed with news research in mind, but many do strive to provide a complete overview of all events of a particular type with consistent and well-documented criteria for inclusion.

For example, the International Disaster Database (EM-DAT) reports over 21.000 international disasters since 1900, and is actively maintained and documented by the Center for Research on the Epidemiology of Disasters (Guha-Sapir et al., 2014). For terrorism research, many studies rely on the RAND database of Worldwide Terrorism Incidents, the International Terrorism: Attributes of Terrorist Events (ITERATE) and the Global Terrorism Database (GTD). The GTD, which is used for the example analysis in this paper, currently documents over 200.000 terrorist events since 1970 (LaFree, 2019). Furthermore, communicative actions such as social media messages, speeches and press releases can also be considered events, and many of these can now be obtained from online platforms. Although these databases will only ever provide an approximate account of all events of a particular type, it is often our best approximation of the real-world set of events. This can be of tremendous value for gatekeeping research, where the lack of knowledge about the real-world set of events has long been a blind spot.

The main challenge is to link these event data to news articles. Depending on the number of unique events and news articles, the number of potential matches can be huge, and only a very small percentage of these will be an actual match.[3] Performing this task manually is extremely time-consuming while requiring high cognitive load, which makes it prone to human error. For these types of coding tasks, computational methods can often achieve or at least approach human level accuracy (King & Lowe, 2003).

Several studies have indeed used computational techniques to (partially) automate this coding task. Gruenewald et al. (2009) linked homicide events from the Newark Police Department investigation files to news articles by using information, such as names of victims and offenders, as keywords that were entered into a searchable database. This can certainly be a viable strategy, but it can only be used under certain conditions. The event information should be cleanly organized, or the number of events should be small enough for humans to extract the keywords. Moreover, it relies on the news database not only for what articles it contains but also for its search functionalities. Sui et al. (2017) conducted an analysis of news coverage in the United States about global terrorist incidents, for which they linked the GTD to news articles. To perform this analysis, the authors wrote an algorithm to link a GTD event to a news article if (1) the article was published within 7 days after the event, and (2) the city and attack type documented in the GTD event description were mentioned in the article. While this strategy might give results that can be used for certain types of analysis, it only makes little use of the textual information from the event description and news text. Although city and attack type are important information for distinguishing events, the specific words used in the GTD will often not match the words used in the news article.

The fields of computer science and computational linguistics provide refined techniques for matching event descriptions to news articles based on (latent) textual similarities (Mozer et al., 2020). Although these fields seem to have focused little on the specific task of analyzing event selection bias (Hamborg et al., 2019) the task of linking events to news is strongly related to techniques for

information retrieval and document clustering. In the field of communication science, scholars have also started adopting these techniques for linking news articles based on similarities in what events are covered (Trilling & Van Hoof, 2020; Welbers et al., 2018). The current study builds on and contributes to these approaches, but with a specialized application for the task of matching event databases to news. We also propose a novel method for using strong assumptions that can be made about the temporal order and distance of events and news coverage.

It should be noted that, for certain event databases and news sources, readily available linkage data can also be obtained from news databases, such as the Global Database of Events, Language and Tone (GDELT). GDELT documents events with aggregated data from news articles, and since the 2.0 update in 2015 also records individual *mentions* of events across news articles. The database is built by experts and uses powerful resources (Leetaru & Schrodt, 2013), so if it provides the data that a researcher needs, the accuracy could be higher than one might achieve by performing the linking oneself. The main limitations of using this type of off-the-shelf data are that it limits control over the data, and the algorithms used are often black boxes (Van Atteveldt & Peng, 2018). As we discuss in our analysis, one of the benefits of the approach used in this paper is that it allows the researcher to adjust parameters that affect precision and recall, to test for robustness and to choose which is more important for a given use case. Still, using GDELT is a viable option if the data are available, and researcher can validate the linkage data to determine whether they are appropriate for their study. For an excellent application of using GDELT for using events in news analysis, we refer to Hopp et al. (2020). The authors of this study also published an article that presents an interface for working with GDELT for communication research, and that discusses the promises and pitfalls of using GDELT (Hopp et al., 2019).

The analysis presented in this paper demonstrates how linking an event archive to news articles can be used to analyze the journalistic gatekeeping process We show that the gatekeeping function can be approximated with a (multilevel) logistic regression model. Hypotheses regarding the event characteristics that increase the probability of the event being covered in the news can then be formulated unambiguously in terms of odds ratios. Thus, we can perform a fine-grained event-level analysis of the gatekeeping process using an established statistical model while maintaining the elegant connection between theory and empirical analysis in Soroka's (2012) gatekeeping function approach.

Predictors of Terrorism Coverage

Our analysis of terrorism coverage serves as a demonstration of how our method can be applied to test relevant hypotheses pertaining to news selection. To contextualize this example, we provide a brief introduction of research on news selection of terrorist events.

The investigation of news values in the context of terrorism coverage reveals a complicated symbiotic relation between journalism and terrorism (Farnen, 1990; Weimann & Winn, 1994). Terrorists have a communicative goal (Hoffman, 2017), and to achieve this goal they parasitically exploit journalistic selection by constructing a package so spectacular and violent that journalists can hardly ignore it (Bell, 1978). For journalists, this results in an unfortunate intersection of news selection criteria and terrorist motives. Their role as gatekeepers compels them to cover these attacks on society but doing so may cause them to inadvertently amplify terrorists' messaging. Abubakar (2020) interviewed journalists to investigate this dilemma, and concluded that while journalist's ethics do play a role, newsworthiness is indeed "the driving force in news media's extensive coverage of violent extremism" (p. 293).

Given the methodological goal of this paper, we narrow our substantive focus to a selection of important news values for terrorism coverage. Firstly, we will focus on news values that intersect with the communication strategy of terrorists, which Bell (1978) describes as constructing spectacular and violent packages that media would find hard to ignore. We focus on two event characteristics documented in the GTD that reflect these values: the number of fatal victims (bad news, magnitude), and whether or not it was a suicide attack (drama). A positive effect of the number of fatal victims on the likelihood of coverage has been found in U.S. coverage of international terrorism (Sui et al., 2017)

and terrorist attack on U.S. soil (Kearns et al., 2019). Interestingly, Zhang et al. (2013) also observed a positive effect of the number of victims in U.S. newspapers on the prominence of terrorism coverage, but found no such effect in Chinese newspapers. In contrast, the effect of a suicide attack on likelihood of coverage is not strongly corroborated. In the study by Sui et al. (2017) the effect of suicide attacks was also hypothesized but not supported (though some indication of an effect was found in the robustness check).

H1a: Terrorist attacks with more fatal victims are more likely to be covered.

H1b: Terrorist attacks where a perpetrator committed suicide are more likely to be covered.

Secondly, we will investigate how the country in which the event occurred affects coverage. Research into international news flows has shown that, controlled by event level characteristics, the location of an event often greatly affects the likelihood that it is covered (Van Belle, 2000; Wu, 2000). As a general rule of thumb, distance has a negative relation with coverage. For terrorist attacks, Sui et al. (2017) found support that geographical distance between the U.S. and other countries indeed has a negative effect on coverage. However, distance does not have to be geographical, but can also be cultural (Chang et al., 1987; Galtung & Ruge, 1965; Sheafer et al., 2014). For people in the United Kingdom – *The Guardian's* home market – we expect that attacks in countries that are more culturally distant are less likely to be covered. Prior support for this hypothesis was also found in the U.S. by Sui et al. (2017), and reflected in the finding of Elmasry and El Nawawy (2020) that American papers covered terrorist attacks in non-Muslim countries much more prominently compared to attacks in Muslim countries.

H2a: Terrorist attacks in countries that are more geographically distant from the United Kingdom are less likely to be covered.

H2b: Terrorist attacks in countries that are more culturally distant from the United Kingdom are less likely to be covered.

Method

We build on recent work in the field of communication that uses document similarity calculations to study news flows (Boumans et al., 2018; Mozer et al., 2020; Nicholls, 2019; Trilling & Van Hoof, 2020; Welbers et al., 2018). Events and news articles are transformed to sparse weighted vectors, and whether an event is likely to be covered in a news articles is then measured based on the similarity of these vectors and the time between the event and the news article publication. In this section, we first describe the data and methods for calculating document similarity. Then we address the problem of determining a suitable similarity threshold, for which we propose a novel approach. Finally, we use a manually coded gold standard to validate our measurements for the example analysis, and to evaluate our method for estimating the similarity threshold.

Data

For our analysis, we use all GTD events (N = 112,091) and all coverage of violent events in *The Guardian* (N = 446,612) over a period of 13 years, from 2006 to 2018. We chose the time period from 2006 to 2018 because before 2006 the number of newspaper articles available via the API from *The Guardian* was very inconsistent, and the GTD data were at the time of data collection available until December 2018.

The GTD is one of the largest databases of terrorist events that has commonly been used for research (LaFree, 2019). Each event is documented according to a range of standardized fields, including location, type of attack, weapons used, targets, victims, terrorist group and a summary. Our method is designed to enable the use of all relevant textual features of the event description as information for matching news articles. We therefore combined all textual features from the GTD data with information about the event (e.g., summary, attack type, city) into a single textual representation.[4] For non-textual features, such as the number of victims, textual descriptions can be added for additional information. In our case, the GTD includes the number of killed and wounded victims, so we included common words to describe this (e.g., killed, injured, casualties) if these numbers are non-zero. Examples of GTD events can immediately be viewed on the GTD website.[5] The whole GTD database is available on request.[6]

The Guardian is a highly influential elite news outlet, with many readers both within the UK and across the globe. An important advantage of focusing on *The Guardian* for the methodological contribution of this paper is that all data can be searched and obtained for free via the *Open Platform* API.[7] Our online appendix provides scripts for downloading the data used in our analysis, which we will maintain as long as these conditions do not change. We used a broad query[8] to collect all articles that potentially cover a terrorist attack. We used all newspaper articles provided by the API, which includes the UK and International sections of the website and the Sunday newspaper *The Observer*. For all articles, we only used the first 200 words. News articles typically have an inversed-pyramid structure, where the who, what, when, where and why questions are presented first (Christian et al., 2014; Hamborg et al., 2018). Focusing on the first 200 words is a simple but effective way to focus the textual similarity measurement on the most relevant content.

Calculating Document Similarity

Calculating the document similarities covers two general steps. First, the documents are *preprocessed* to transform them into vector representations. Second, the similarity of documents is measured by calculating the similarity between document vectors. There are various choices to make in each of these steps, and which choices work best for what tasks are still an active area of research (Mozer et al., 2020). Here we describe some key considerations for calculating document similarities for the purpose of linking events to news articles.

From Text to Vector

A broad distinction can be made between two common types of vector representations of documents. One is the bag-of-words (BOW), in which the vector elements are words or short sequences of words (Boumans et al., 2018; Nicholls, 2019; Welbers et al., 2018). This works well for finding document matches on specific content elements, because information about specific words (e.g., people, organization, locations) is maintained. An implication is that words in two texts need to be identical in order to contribute to the similarity score. For preprocessing, it is therefore recommended to apply *lowercasing* and *stemming* (or *lemmatization*) to remove insignificant differences between words due to capitalization, singular-plural and verb conjugations (e.g., President versus president, attack versus attacked).

The alternative is to use lower-dimensional vectors that represent more latent topical information about documents, that can for instance, be obtained via topic modeling, word embeddings or factor analysis (Blei et al., 2003; Mikolov et al., 2013). The strength of this approach is that it can find semantic similarities even if the specific words are different. For instance, that "London bombings" is similar to "Paris shootings." Conversely, by focusing more on topical similarity between documents, the discriminative value of specific keywords is diminished. For matching event data, we do not actually want "London" to be considered similar to "Paris."

As a general guideline, a lower-dimensional representation works well for matching documents on topical content, but for more specific matches where specific keywords matter, a BOW approach can

be more appropriate (Mozer et al., 2020; Trilling & Van Hoof, 2020). For matching many events to many news articles, keywords pertaining to who, what and where are particularly important, so we used a BOW approach. For our example analysis, we used standard preprocessing techniques for English language data (lowercasing and stemming). Although more advanced techniques, such as lemmatization might give better results, for English language data the difference would be minor, and the simple preprocessing techniques make our analysis easier to replicate.[9] The preprocessed texts are then represented as a Document Term Matrix (DTM).

We also applied two additional preprocessing steps, for which we provide code in our online appendix. Firstly, terms with similar spellings (e.g., Bogota, Bogotã) are merged together as single columns in the DTM. To determine whether terms had similar spelling we used overlapping character tri-grams.[10] Fixing minor spelling differences is particularly important for the data in our example analysis, because the proper nouns that are key for identifying terrorist events often have different spelling variants. Also, the GTD contains fairly many spelling mistakes. It does sometimes occur that words with similar spelling but different meanings are incidentally merged, but this is much less harmful for accuracy than failing to merge slightly different words with the same meaning.

Secondly, we combine terms to add more informative features. Some terms, such as "London" and "bomb," are not very informative on their own in our data, but very informative combined. This step basically adds more and more specific content elements to our document vectors. It is comparable to adding bi-grams (i.e., sequences of two terms) to a DTM to add more information, but without the strict condition of term order. It matters for the calculation of document similarity because measures based on the dot product only sum the products of individual terms and do not consider their interactions. Applying this technique does come with two caveats. Firstly, it requires some criteria for deciding which combinations are included in the vector, because including all combinations (a quadratic increase of vector length) would impose memory issues. In short, we only keep combinations that occur at least once in the GTD data, that occur in less than 1% of all articles, and of which the observed frequency is less than 1.2 times the expected frequency. The function to compute these combinations, and more information on this technique, is provided in the online appendix. Secondly, this step greatly increases the vector magnitude, which needs to be taken into account in deciding the measure for vector similarity, as discussed in the next section.

A final step for preparing the document vectors is to weight the terms. We use the tf-idf weighting scheme (Sparck Jones, 1972), that weights down terms (and term combinations) that occur in many documents by multiplying the term frequency (tf) by the inverse document frequency (idf). We only look at whether or not a query occurs in a news article (i.e., tf is binary), so the term score is effectively the idf. We calculate the idf scores based only on term frequencies in the news articles, because including the GTD scores would weight down the importance of terms that occur in many terrorist events.[11]

Examples of elements of our vector after merging terms with similar spelling (the OR operators) and combining terms (the AND operator) are given in Table 1. Here we for instance, see that "fire AND al-sattar" is a highly informative-term combination, and that some spelling variations of the term "pro-government" have been merged together.

Table 1. DTM columns after lowercasing, stemming, merging words with similar spelling and adding term combinations.

Weight	Column
1	fire AND al-sattar
0.959	iraq AND (al-sukkariyah OR al-sukkariya)
0.724	iraq AND (archbishop OR archbish OR arch-bishop OR)
0.588	pro-govern OR pro-government OR pro-governemnt

Calculating the Similarity of Vectors

A common recommendation for calculating document similarity is to calculate the cosine similarity of the document vectors (Trilling & Van Hoof, 2020). However, for matching short event descriptions to news articles the regular dot product can be a good alternative, and is used in our example analysis. Cosine similarity is equivalent to the dot product after normalizing the vector magnitudes. This can be nice because it bounds the dot product between 0 and 1, which can intuitively be interpreted as the proportion of similarity. However, for matching event data to news, we are actually not interested in the proportion of similarity. We only want to know whether there is sufficient reference to the event in the news article, regardless of the other information in the news article. For example, say we have the event description "Bomb in London," and there are two news articles: "Bomb in London underground" and "The impact of 9–11 and the London Bombings." The dot products of the event description (after prepossessing) with the two news articles would be identical, but the cosine similarity for the second news article would be lower because of the larger vector magnitude. To determine whether a news article mentions an event, we use the dot product because we are interested in the scores of terms they have in common, regardless of any additional terms. Note that this is particularly important if preprocessing techniques are used that inflate the amount of information in document vectors, such as adding bi-grams or tri-grams, or our technique for adding term combinations. Our R package provides a specialized implementation of matrix multiplication for efficiently computing the dot products (and other similarity measures) of all events and news articles within a given time window with the option to only store results above a given similarity threshold, making this analysis feasible even on current-generation personal computers.

Determining the Similarity Threshold

The document comparison procedure gives us similarity scores, so to obtain discrete matches of events to news articles we need to use a threshold (Mozer et al., 2020). The higher the similarity score of an event-article pair, the more likely it is that the article actually covers the event. Accordingly, a higher threshold gives better precision (fewer false positives) but if the threshold is too high it harms the recall (more false negatives). A common approach to determine a good threshold is to develop a gold standard based on manual coding (see, e.g., Burggraaff & Trilling, 2020; Linder et al., 2018; Welbers et al., 2018). The precision and recall can then be calculated for the discrete matches obtained with different thresholds to see which threshold gives the best results. The drawback of this approach is that it requires a difficult manual coding task to get even a small gold standard. For the specific task of matching event data to news, we propose a novel approach that can be used to estimate a good threshold if gold standard data are not or not sufficiently available. In the next section, we explain and apply this approach, and then we use a manually coded gold standard to show that our estimation gives a similar suggested threshold.

Let us first be very specific about how we measure the validity of discrete matches. We define a *match* as a document pair, in our case, a GTD event paired with a news article, for which the similarity score is above a certain threshold. A match is a *true positive* (TP) if the news article actually covers to the GTD event, and a *false positives* (FP) if it does not. A *false negative* (FN) means that we did not find a match between an event and a news article even though the news article actually does cover the event. Given these numbers, we can calculate validity in terms of the precision (P) and recall (R), and their harmonic mean (F1).

Estimating Precision and Recall

The gist of our technique is that we estimate the precision, recall and F1 score by making the assumption that the false-positive rate (FPR) is constant over time. To show the intuition behind this approach, Figure 1 presents three bar charts that show how many matches were found over time with different similarity thresholds (2, 7 and 14), where the x-axis represents how many days a news article was published after the event. For the analysis, we are only interested in the matches with

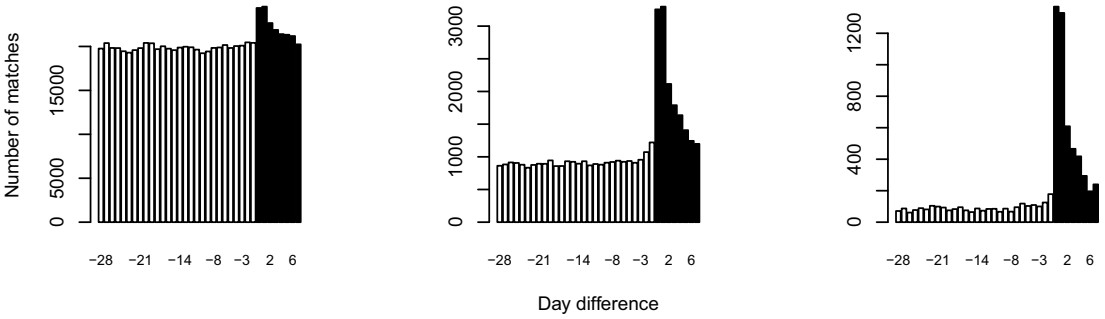

Figure 1. Frequency of matches with news articles published before (white) and after (black) events for similarity thresholds 2, 7 and 14.

a positive time difference, for which the bars are colored black. After all, we can be certain that matches with a negative time difference are false positives, because there cannot be news coverage of an event that has not yet happened. Yet, this does give us an estimate of the rate of finding false positives. Assuming this rate is constant, we can estimate how many of the matches with a positive time difference are true positives.

This appears to be a fairly strong assumption. In case of a false positive, the article does not cover the event, so whether the article was published before or after the event is irrelevant. But the assumption does rely on the independence of events, and this can be violated in the case of chains of terrorist attacks. In these cases, new events are more likely to be similar to news articles covering the previous attacks. This can be seen in the small rises in Figure 1 just before the zero-day difference point in the second and third histogram. More sophisticated variations of this estimation might correct for this, but for the current application, the impact of violating the assumption appears to be minor.

Roughly speaking, the average height of the white bars indicates the false-positive rate.[12] If the algorithm works, the black bars should be higher, because these can contain true positives on top of the false positives. This is especially the case for 1 or 2 days after the event, because this is when most news coverage should occur. As we increase the threshold, the *proportion* of the black bars that is higher than the false-positive rate increases, indicating higher precision. However, the *absolute number* by which the black bars are higher decreases, indicating that we might now have more false negatives. By more formally putting numbers to these estimates, we can calculate an F1 score that balances precision and recall. Details on how to perform the calculation are provided in the online appendix[13]

Figure 2 shows the estimated precision, recall, and the corresponding estimated F1 score, for a range of similarity thresholds. The vertical line, which indicates where the estimated F1 score is highest, shows that the suggested weight threshold is 6.54.

Validation 1

We first created a gold standard by manually coding which GTD events are referred to in 500 random news articles. A single expert coder followed a standardized procedure. In the first stage, the news article is scanned for whether it covers a violent attack (or multiple attacks) that occurred in the past 7 days. Any type of attack is used at this point, so the coder does not have to interpret whether the attack should be defined as terrorism. For each attack, the coder used a script to filter the GTD data by date and country for a list of potential matches. In the second stage, the coder reads the event information for all of the potential matches to determine whether it is mentioned in the article. We found 113 event-article pairs, for 27 articles that matched one or more GTD events.

To verify that the judgment of the expert coder in the second stage is reproducible, we conducted an inter-rater reliability (IRR) test with two additional coders. The IRR test set was sampled from the gold

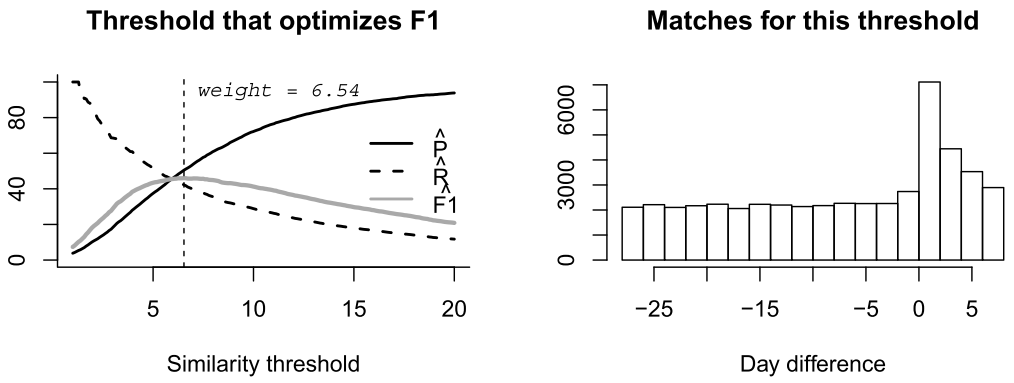

Figure 2. The threshold that optimizes the estimated F1 (left, vertical line) and the matches for this threshold (6.54) over time (right).

standard so that approximately one out of four potential event-article pairs was coded as true in the gold standard, and so that for all included articles all possible matches need to be coded (identical to the original task). This yielded 26 articles with 142 possible matches that all three coders rated, on which they achieved a Krippendorf's alpha of 0.775. This is an acceptable reliability given that references to events in news are often ambiguous.[14]

Figure 3 reports the precision, recall and F1 scores for different weight thresholds, given separately for the *article* and *match* level. The match level represents the accuracy of the specific event-article pairs, as also discussed above. By article level, we refer to the precision and recall of identifying which news articles cover *at least one* GTD event. We included this as additional proof of the validity of the matching algorithm and to show how aggregating matches to the article (or event) level generally improves accuracy.

The highest F1 score (65%) is obtained with a threshold of 5.96, at which point the precision is 54.5%, and the recall is 80.53%. For the article level, the highest F1 score is much better, mainly because recall decays less steeply, allowing a higher threshold to be used to increase precision. This makes sense, because aggregating event-article pairs to binary measures of whether an article covers *at*

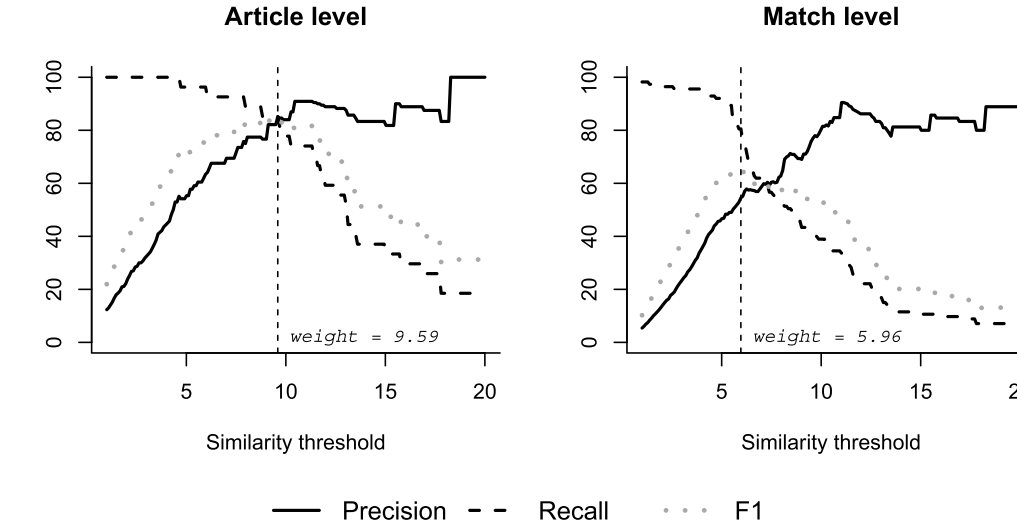

Figure 3. Precision, recall and F1 scores for different threshold values based on gold standard.

least one terrorist attack makes it easier to obtain high-recall. The same should apply to measuring whether or not an event is covered in at least one article, meaning that the match level validity is likely to underestimate the validity of our data as used in the analysis. However, we cannot verify whether this is truly the case with these validation data.

Validation 2

In addition to the gold standard, we also performed a manual precision measurement for 100 event-article pairs that according to the algorithm are matched if we use a threshold of 6.54 (as used in our analysis). Where the gold standard was carefully designed to enable a proper measurement of recall by coding all possible event-article pairs for 500 articles, the rareness of an actual match resulted in only 113 matches for 27 articles. This additional validation provides a complementary measurement of the precision score if a threshold of 6.54 is used, based on 100 unique articles and events. The validation set is also included in the online appendix, and contains comments from the coder for all cases where the reference to the GTD event was ambiguous, and urls for viewing *The Guardian* article and GTD item. This makes the strengths and weaknesses of the algorithm more transparent.

The most important finding is that this validation indicates that the gold standard validation does not overestimate the precision, and even slightly underestimates it. At a threshold of 6.54 the precision in the gold standard is 57.4%, and 64% in this additional precision measurement.

Comparison of Estimation Method and Gold Standard

Our threshold estimation technique provides a cheap and fast measure that can be used to see if changes in the parameters of the matching algorithm, such as the similarity threshold, are likely to improve the validity. For our method to be useful, the parameters of the matching algorithm that optimize the estimated F1 score should be close to the parameters that optimize the real F1 score.

In Figure 4 we compare our estimated F1 score with the gold standard F1 score for different similarity thresholds. Here we see that the scores are clearly correlated, and that the thresholds that optimize the scores are indeed close. For the match level, the highest F1 score is found at a weight threshold of 5.96, which is quite close to the threshold of 6.54 for the estimated F1 score. Also, note

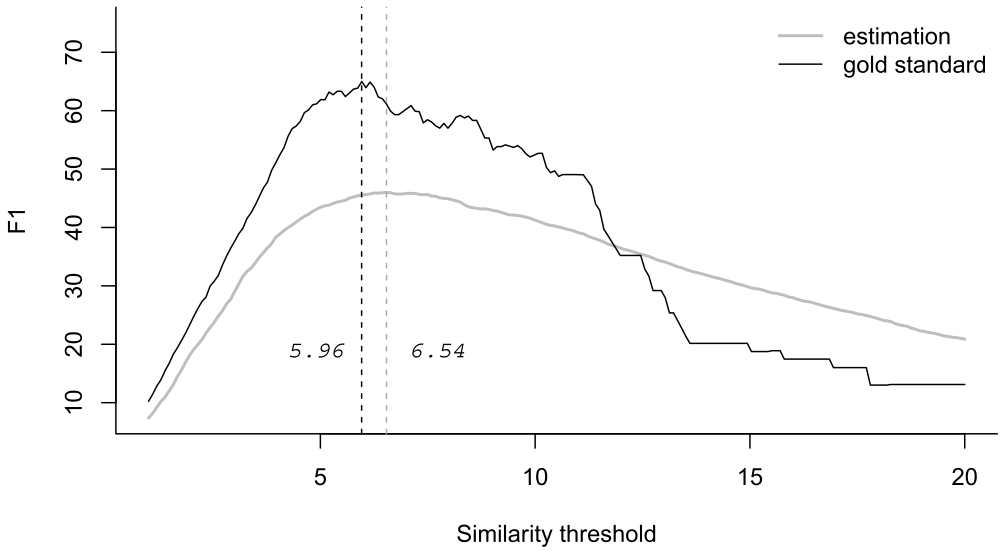

Figure 4. Relation between the similarity threshold and F1 score for both the automated estimation and gold standard validation.

that for both validation approaches the F1 score does not vary much within the range from 5.96 to 6.54. In the analysis, we show that choosing one or the other makes no difference to the conclusions from our analysis.

We also used both the gold standard and the estimation method to compare the results with and without the two additional preprocessing techniques that we proposed. The results are presented in Table 2. The two preprocessing techniques, summarized as *merged* and *combined*, refer to the step of merging terms with similar spellings to a single column, and adding columns with combinations of terms. The similarity threshold column reports the recommended threshold for both the gold standard and our estimation method. The highlighted row is the combination with the best performance, where we used both preprocessing techniques. This table mainly shows two things. Firstly, it provides evidence that the combination of these choices is important. This was our expectation, because merging terms with similar spelling mainly serves to improve recall by avoiding 'missing' terms that match, and combining terms serves to improve precision by adding more weight to rare combinations. The second observation is that the estimated similarity threshold stays close to the gold standard threshold even as the threshold varies wildly. This provides additional support for the usefulness of the estimation procedure.

Analysis

For our analysis, we use the results of the matching algorithm with a time window of 7 days after the event, using a similarity threshold of 6.54. For this analysis, we only use GTD events in which at least one fatal casualty was reported (N = 56,359). This narrows our focus to the subset of *fatal* terrorist attacks, which has the benefit of reducing ambiguity in the GTD data regarding at what point of impact an attack warrants documentation and can be defined as a terrorist attack. Also, this makes the test of H1a more appropriate, because we focus on the effect of *more* fatal victims.[15]

To better understand the multilevel structure of our data, we first visualize country-level differences in media coverage. In absolute numbers, by far most event-article pairs are found for countries with a high frequency of terrorist attacks, such as Iraq and Afghanistan. But when we look at the relative news coverage, the picture emerges that countries that are geographically and culturally closest to the United Kingdom are over-represented. Figure 5 shows the average number of news articles per event per country, which higher scoring countries colored black. This number is much higher in countries, such as France, Norway, the USA and Australia (where *The Guardian* has sister outlets) and the UK itself. While this aligns with the news values underlying H2a and H2b, the total number of attacks in these countries is also a factor. In Norway, in particular, the number of fatal attacks is very low, but the massacre of 77 people by Anders Breivic in 2011 dominated the news.

We use a multilevel logistic regression model (Bates et al., 2015) to investigate the predictors of which events are (1) or are not (0) covered by *The Guardian*, with events nested in countries. This model serves as an approximation of the gatekeeping function from an event perspective.[16] The effects of event characteristics on the probability that an event is covered are then expressed clearly in terms of the odds ratios.

Table 2. Effect of preprocessing steps on optimal thresholds.

Preprocessing		**Similarity threshold**		**Performance***		
Merged**	Combined***	Gold standard	Estimated	F1	P	R
No	No	17.55	20.39	55.94	49.32	64.60
Yes	No	17.71	19.35	57.49	52.99	62.83
No	Yes	12.82	12.62	55.28	63.95	48.67
Yes	Yes	5.96	6.54	65.00	54.49	80.53

* Based on gold standard using the threshold that optimizes F1.
** Have terms with similar spelling been merged?
*** Have term combinations been added?

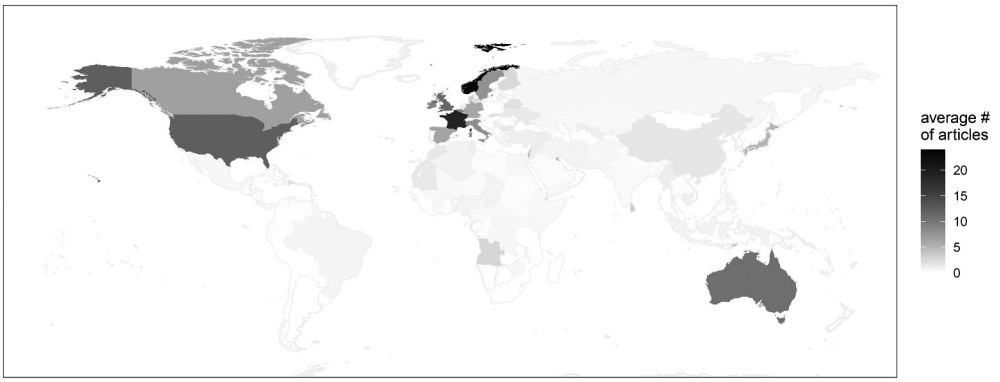

Figure 5. Average number of news articles for every fatal GTD event per country.

We include the following independent variables, corresponding to our hypotheses:

H1a: fatal victims (log). The number of people killed in the attack, using a natural log transformation to account for heavy skew.

H1b: suicide attack. Binary value for whether or not a perpetrator committed suicide.

H2a: geographical distance. Country-level variable for distance of the UK (London) to other countries. For the specific coordinates the center point of terrorist attacks in a country is used. Distance is transformed to z-scores for more interpretable odds ratios and model convergence.

H2b: cultural distance. Country-level variables for cultural distance, measured as the difference in cultural values and political decisions (see, e.g., Sheafer et al., 2014). We do not make a single score out of these variables because they represent different cultural dimensions that are only moderately correlated ($r(38) = 0.358$, $p < .05$).

- **Cultural values distance** Euclidean distance of countries on the Traditional versus Secular-Rational and Survival-Self Expression scales (Inglehart & Welzel, 2005). The scales are calculated using data from Wave 5 (2005–2009) of the World Values Survey (Inglehart et al., 2005).[17]
- **UN voting difference** Percentage of votes in the United Nations general assembly where a country voted different from the UK. Calculated over all votes with yes or no answers from both countries since 2000.

For the full analysis, we can only include countries for which we could calculate the cultural distance values (i.e., UN countries that participated in the WVS wave 5). The results presented in Table 3 are based on this selection (N_{events}= 7,606; $N_{countries}$= 38). For comparison, we also include the results for a model with all countries (N_{events}= 56,171; $N_{countries}$= 131). The coefficients are reported as odds ratios, with standard errors for the log odds ratios, and the level 1 error variance is fixed to $\pi^2/3$. Starting from a base model with only random intercepts (1), we add the event level predictors (2), geographical distance (3) and cultural distance (4). We evaluate our hypotheses based on the full model (4), and in addition report the findings for the data including all countries to show how our findings are robust in this larger and more diverse set of events.

The first two hypotheses are clearly supported. For every unit increase on the natural log scale of the number of fatal victims, the odds increase by a factor of 1.77 (95% CI = 1.64–1.92, $p < .001$), meaning that the events with more fatal victims are more likely to be covered (H1a). Suicide attacks (H1b) have 6.59 times the odds of non-suicide attacks (95% CI = 4.76–9.14, $p < .001$). The *all countries* model

Table 3. Multilevel logistic regression analysis, predicting news coverage of fatal GTD events nested in countries.

	Fatal terrorist attack (GTD) covered in The Guardian				
	UN countries that participated in WVS				*All countries*
	1	2	3	4	
Fixed effects Odds Ratios ($SE_{log\ OR}$[1])					
Constant	0.50 (0.28)*	0.29 (0.28)***	0.23 (0.27)***	5.05 (0.51)**	0.19 (0.12)***
H1a: fatal victims (log)		1.76 (0.04)***	1.76 (0.04)***	1.77 (0.04)***	1.46 (0.01)***
H1b: suicide attack		6.57 (0.17)***	6.54 (0.17)***	6.59 (0.17)***	8.03 (0.03)***
H2a: geographical distance (z)			0.67 (0.17)*	0.98 (0.14)	0.77 (0.08)**
H2b: cultural values distance				0.06 (0.70)***	
H2b: UN voting difference				0.01 (1.43)**	
Random Effects					
σ^2_{event}	3.29	3.29	3.29	3.29	3.29
$\tau_{country}$	2.16	1.99	1.63	0.47	1.16
Deviance	5972.1	5590.9	5585.8	5556.8	51,657
χ^2		381.2***	5.1*	29.0***	
AIC	5,976.1	5,598.9	5,595.8	5,570.8	51,667
BIC	5,990.0	5,626.7	5,630.5	5,619.4	51,711
N_{events}	7,608	7,608	7,608	7,608	56,175
$N_{countries}$	38	38	38	38	131

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.
[a]Coefficients are transformed to odds ratios, but SE reported for log odds ratios.

shows similar results that also support H1a (odds ratio = 1.46, 95% CI = 1.43–1.49, $p < .001$) and H1b (odds ratio = 8.03, 95% CI = 7.53–8.57, $p < .001$).

Our findings for the hypotheses regarding the news values of geographical proximity (H2a) and cultural proximity (H2b) have an interesting implication for news values theory. If we do not take the cultural distance into account, as in model 3, we do find support for H2a, confirming that events physically close to home are generally more newsworthy. For every standard deviation that a country is further away, the odds decrease by a factor of 0.67 (95% CI = 0.48–0.94, $p < .05$). This hypothesis is also supported by the *all countries* model (odds ratio = 0.77, 95% CI = 0.66–0.91, $p < .001$). However, when we include cultural distance in the model, the effect of geographical distance disappears (odds ratio = 0.98, 95% CI = 0.74–1.30, $p = .895$). For reference, geographical distance is moderately correlated with UN voting difference ($r(38) = 0.496$, $p < .001$) and is not correlated with cultural values distance ($r(38) = -0.110$, $p = .471$). Instead, we now see negative effects of both cultural values distance (odds ratio = 0.06, 95% CI = 0.02–0.23, $p < .001$) and UN voting difference (odds ratio = 0.01, 95% CI = 0.00–0.12, $p < .01$). Thus, based on the full model, we reject H2a and accept H2b.

Effect of Threshold on Results

Given the difficulty of determining the best possible threshold to use, it is important to confirm whether using a slightly different threshold would change the conclusions. In Figure 6 we present the results of the full model (model 4 in Table 3) for different similarity thresholds. The odds ratios (y-axis) for each independent variable are plotted for similarity thresholds ranging from 1 to 20 (x-axis). The whiskers indicate the 95% confidence interval, and dotted horizontal lines are drawn at $y = 1$ (equal odds) to more clearly show the direction and significance of effects.

The results show that the odds ratios are remarkably robust for changes in the similarity threshold. Within a wide range of thresholds surrounding the threshold that optimizes the F1 score (6.54) the odds ratios are very similar, and conclusions based on the direction and statistical reliability of effects would have been identical. It is mainly for very low thresholds, where the precision gets very weak that we get different results.[18] For higher thresholds, the recall decreases and only events with at least one very high similarity score remain. This mainly widens the confidence intervals, but also slightly affects the odds ratios.

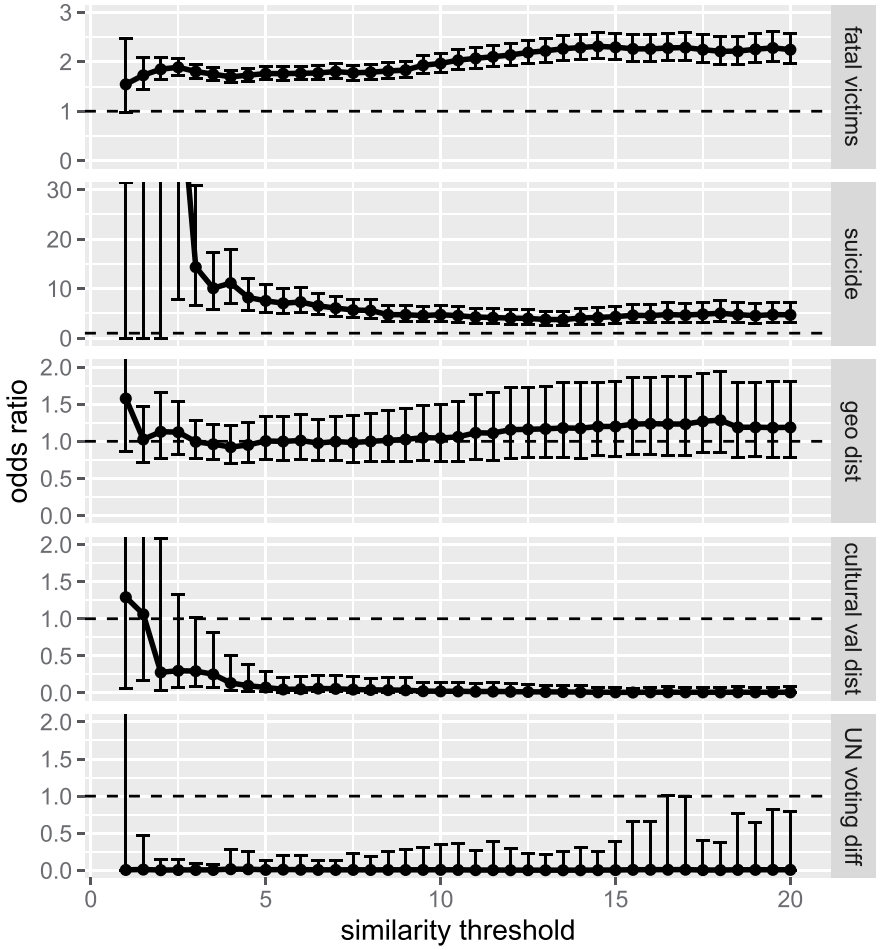

Figure 6. Odds ratios of full model with 95% CI for different similarity thresholds.

For our best approximation of the true gatekeeping function, it makes sense to use the threshold that optimizes the F1 score of the event matches. However, it is also interesting to see that the results are mostly the same if we would have used a much higher threshold, where we sacrifice recall for better precision. Overall, this indicates that the results of the matching algorithm, even with imperfect accuracy and using our weakly supervised method for determining the threshold, can reliably be used to estimate the gatekeeping function. Note that for other data and predictors of coverage, changing the threshold might have a stronger impact on odds ratios, so for other studies we recommend performing the same comparison of statistical models for different thresholds as shown in Figure 6.

Discussion

We presented a computational method for linking event data to news articles, and demonstrated how this method can be used to investigate news selection. Our validity tests suggest that the method performs good enough to use the linked data directly in an analysis, and our example analysis demonstrates how these data can be used to analyze the journalistic gatekeeping process on a large scale. We conclude that this method enables the conceptual gatekeeping model to be formalized as a statistical model, similar to the method proposed by (Soroka, 2012) but from an event perspective.

Hypotheses concerning multiple event level predictors of coverage can then be tested in a multivariate analysis.

The example analysis presented in this paper demonstrated a simple case where we only investigated event-level predictors of *whether* an event is covered. Beyond this classic gatekeeping question, the linked event-article data can be used to combine event-level data with a content analysis of the news message to also investigate *how* an event is covered. For example, a sentiment analysis could reveal whether certain types of terrorist attacks are covered with different emotions, and we could analyze which attacks are actually labeled as "terrorism" by journalists. Furthermore, data about news outlet and time could be included, which could be used to investigate organizational level factors and institutional level shifts. By leveraging event databases and computational techniques for linking them to news articles, large-scale event-level analysis can allow us to ask new questions, and approach old questions from a different perspective.

A novel methodological contribution of this study is the method for estimating a good similarity threshold without relying on a gold standard. Based on the fact that news coverage of an event is most likely to follow shortly after it occurs, and can never precede it, we can estimate a precision and recall score, and their harmonic mean (F1). We found that the similarity threshold in the algorithm that optimizes this estimated F1 score lies close to the threshold that optimize the 'real' F1 score. This should be tested on more cases, because if it holds true within a passable margin of error, it could be used to determine good parameters for event-to-news matching tasks in cases where extensive gold standards are not easily available. The parameter estimation technique might also be generalized to other tasks where document similarity is used to study news coverage. In this paper, we proposed a simple implementation of the idea that we can use time to estimate a false-positive rate based on matches that are *impossible*, but this idea can be generalized to matches that are *less plausible*. For example, for the task of identifying whether different news articles cover the same event, it is safe to assume that these news articles should be published around the same time. By fitting the observed matches to an expected probability distribution of the time differences of news article pairs, we can calculate which settings of the matching algorithm optimize this fit. Thus, this idea might more broadly have applications in the study of news flows and text reuse.

An important practical observation is that the results of the statistical analysis are remarkably robust for small to moderate changes in the threshold. This suggests that for statistically investigating the effects of event characteristics on the odds of news coverage, when using this event matching approach on a sufficiently large scale, it is good enough to use an approximately good threshold. As a pragmatic first step for this type of analysis, one could use our parameter estimation technique for determining a good threshold value, and then perform the analysis for a range of thresholds around this value, as we did in Figure 6.

Regarding the limitations of the event matching approach, it should be assumed that large event archives are never completely accurate, and that the criteria for including certain events and excluding others can be inconsistent (for a discussion of terrorism archives, see LaFree, 2019). Since news data are often an important information source for the creation of event databases, distortions in the news coverage can also spill over, thus confounding our *real-world* input. This can also lead to differences between countries due to different media systems and press freedom, which affects the supply of information based on which event databases are often maintained. There can also be structural biases in the matching data, because different types of events and countries tend to be covered in different ways. This can affect the accuracy of a matching algorithm based on text similarity, because certain ways of covering events might be less similar to how events are described in the database.

A key avenue for progress is to pursue higher precision and recall. Using the threshold with the highest F1 score (F1 = 65%), the recall was good (80.53%), but the precision was on the low side (54.49%). In other words, the algorithm managed to find about four out of five of the actual event-news pairs, but almost one out of two matches was a false positive. This is not bad for this type of extreme rare event problem, but it does mean that there is still much noise in our analysis, and likely also structural bias. We identify two main ways in which our current approach can contribute toward

further development. Firstly, it can be used as a preparatory step to cull the vast majority of irrelevant event-article pairs. With a recall of 80.53%, over 99.943% of all possible matches (i.e., news articles within 7 days after the event) could be removed. Another method, possibly involving more computationally demanding techniques, could then be used to improve precision with less harm to recall. Secondly, this preparatory step also makes manual content analysis of the remaining article pairs feasible. With minor concessions to recall, this could be used to overcome human limitations in developing gold standards and training data for machine learning approaches.

We have argued that linking event archives to news articles can be a powerful method for analyzing the gatekeeping process. The tool provided with this paper can be used to perform this type of analysis or as a starting point for improving the method and creating better gold standards. To accommodate a closer investigation of improvements or flaws in our code and method, the online appendix makes our analysis fully replicable and has detailed documentation. The open-source software will actively be maintained, and relevant new developments will be integrated.

Notes

1. The RNewsflow package, available on CRAN: https://cran.r-project.org/web/packages/RNewsflow/index.html
2. https://github.com/kasperwelbers/gtdnews
3. In our data, the 446,612 news articles and 112,019 events yielded 86,274,229 possible matches within a 7 day time window, and only 49,079 of these were actual matches (according to our results using a threshold of 6.54).
4. Specifically, we used the fields: *summary, city, country_txt, attacktype_txt, targtype_txt* and *targsubtype_txt* (target type), *natlty_txt* (target nationality), *gname* (terrorist group name), *gsubname, weapptype_txt* and *weapdetail* (weapon type and detail), and *victims*. For numbered fields (*weaptype1_txt, weaptype2_txt*, etc.) all fields were included.
5. Example of a GTD event: https://www.start.umd.edu/gtd/search/IncidentSummary.aspx?gtdid=201911290008
6. https://www.start.umd.edu/gtd/
7. https://open-platform.theguardian.com/
8. Our population query contained 165 terms such as "dead", "attack", "bomb" and "kidnapped". The full query is available in the online appendix
9. Note that certain languages it is generally better to use lemmatization instead of stemming. In R this is supported for some languages in the udpipe (Wijffels, 2019) and spacyr (Benoit & Matsuo, 2020) packages.
10. Specifically, we merged terms that (1) start with the same character, (2) at least 2/3 character tri-grams were identical, and (3) no more than 4 tri-grams were different
11. We also divided the idf scores by the highest possible idf score so that values are bounded between 0 and 1. This division by a constant does not at all affect the analysis, but makes interpretation easier.
12. Our actual calculation is different, because we calculate the rate of matches over the total number of event-article comparisons, instead of the rate over time
13. https://github.com/kasperwelbers/gtdnews/blob/master/online_appendix/threshold_estimation.pdf
14. To illustrate this, the second validation set includes comments about ambiguous cases
15. The analysis using all data is available in the online appendix. The only notable difference in findings is that the effect of cultural values distance – one of the two cultural distance variables – disappeared.
16. We also fit a multilevel negative binomial model using the number of articles per event as the response, to test whether analyzing the amount of articles rather than whether or not an event is covered would have changed the conclusions. The incidence rate ratios were very similar to the odds ratios of the logistic regression, and the substantive conclusions would have been the same. However, the assumption that incidents (i.e., the articles to which an event is linked) are independent is clearly violated, and the model also failed to converge. Since most variance is in whether or not an event is covered, we reported the simpler and more reliable logistic model.
17. Wave 6 (2010-2014) is available, but without the UK. We assume that these values changed little within the period of our study, which is consistent with comparisons based on previous waves.
18. This lowest threshold is already fairly high. Of all comparisons between events and news articles published within 7 days after the event, only 0.71% pass this threshold.

ACKNOWLEDGEMENT

We would like to thank the anonymous reviewers at Communication Methods and Measures for their detailed and insightful comments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft [WE 2888/7-1]; National Endowment for the Humanities [HJ-253500-17]; Dutch: Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) [463-17-004].

ORCID

Kasper Welbers http://orcid.org/0000-0003-2929-3815
Chung-Hong Chan http://orcid.org/0000-0002-6232-7530
Hartmut Wessler http://orcid.org/0000-0003-4216-5471
Marc Jungblut http://orcid.org/0000-0002-2677-0738

References

Abubakar, A. T. (2020). News values and the ethical dilemmas of covering violent extremism. *Journalism & Mass Communication Quarterly*, *97*(1), 278–298. https://doi.org/10.1177/1077699019847258

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bell, J. B. (1978). terrorist scripts and live-action spectaculars. *Columbia Journalism Review*, *17*(1), 47.

Benoit, K., & Matsuo, A. (2020). *spacyr: Wrapper to the 'spacy' 'NLP' library*. R package version 1.2.1. https://CRAN.R-project.org/package=spacyr

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022. https://www.jmlr.org/

Boumans, J., Trilling, D., Vliegenthart, R., & Boomgaarden, H. (2018). The agency makes the (Online) news world go round: The impact of news agency content on print and online news. *International Journal of Communication*, *12*, 1768–1789. https://ijoc.org

Burggraaff, C., & Trilling, D. (2020). Through a different gate: An automated content analysis of how online news and print news differ. *Journalism*, *21*(1), 112–129. https://doi.org/10.1177/1464884917716699

Chang, T.-K., Shoemaker, P. J., & Brendlinger, N. (1987). Determinants of international news coverage in the US media. *Communication Research*, *14*(4), 396–414. https://doi.org/10.1177/009365087014004002

Christian, D., Froke, P., Jacobsen, S., & Minthorn, D. (2014). *The associated press stylebook and briefing on media law*. The Associated Press.

Elmasry, M. H., & El Nawawy, M. (2020). The value of Muslim and non-Muslim life: A comparative content analysis of elite American newspaper coverage of terrorism victims. *Journalism*, https://doi.org/10.1177/1464884920922388 .

Farnen, R. F. (1990). Terrorism and the mass media: A systemic analysis of a symbiotic process. *Studies in Conflict & Terrorism*, *13*(2), 99–143. https://doi.org/10.1080/10576109008435820

Gadarian, S. K. (2010). The politics of threat: How terrorism news shapes foreign policy attitudes. *The Journal of Politics*, *72*(2), 469–483. https://doi.org/10.1017/S0022381609990910

Galtung, J., & Ruge, M. H. (1965). The structure of foreign news. The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research*, *2*(1), 64–90. https://doi.org/10.1177/002234336500200104

Gruenewald, J., Pizarro, J., & Chermak, S. M. (2009). Race, gender, and the newsworthiness of homicide incidents. *Journal of Criminal Justice*, *37*(3), 262–272. https://doi.org/10.1016/j.jcrimjus.2009.04.006

Guha-Sapir, D., Below, R., & Hoyois, P. (2014). EM-DAT: International disaster database. *Centre for Research on the Epidemiology of Disasters*. http://www.cred.be/sites/default/files/ADSR_2014.pdf

Hamborg, F., Donnay, K., & Gipp, B. (2019). Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries*, *20*(4), 391–415. https://doi.org/10.1007/s00799-018-0261-y

Hamborg, F., Lachnit, S., Schubotz, M., Hepp, T., & Gipp, B. (2018). Giveme5W: Main event retrieval from news articles by extraction of the five journalistic W questions. In: G. Chowdhury, J. McLeod, V. Gillet, & P. Willett (eds), *Transforming digital worlds (356-366). iConference 2018*. Springer. https://doi.org/10.1007/978-3-319-78105-1_39

Harcup, T., & O'neill, D. (2017). What is news? News values revisited (again). *Journalism Studies*, *18*(12), 1470–1488. https://doi.org/10.1080/1461670X.2016.1150193

Hoffman, B. (2017). *Inside Terrorism*. Columbia University Press. https://doi.org/10.7312/hoff17476

Hopp, F. R., Fisher, J. T., & Weber, R. (2020). Dynamic transactions between news frames and sociopolitical events: An integrative, hidden Markov model approach. *Journal of Communication*, *70*(3), 335–355. https://doi.org/10.1093/joc/jqaa015

Hopp, F. R., Schaffer, J. A., Fisher, J. T., & Weber, R. (2019). CoRe: The GDELT interface for the advancement of communication research. *Computational Communication Research*, *1*(1), 13–44. https://doi.org/10.31235/osf.io/smjwb

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, D. (2005). *World values survey: Round five - country-pooled datafile version*. JD Systems Institute. https://www.worldvaluessurvey.org

Inglehart, R., & Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press. https://doi.org/10.1017/cbo9780511790881

Iyengar, S. (1990). The accessibility bias in politics: Television news and public opinion. *International Journal of Public Opinion Research*, *2*(1), 1–15. https://doi.org/10.1093/ijpor/2.1.1

Kearns, E. M., Betus, A. E., & Lemieux, A. F. (2019). Why do some terrorist attacks receive more media attention than others? *Justice Quarterly*, *36*(6), 985–1022. https://doi.org/10.1080/07418825.2018.1524507

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*(3), 617–642. https://doi.org/10.1017/S0020818303573064

LaFree, G. (2019). *The evolution of terrorism event databases. In the oxford handbook of terrorism*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198732914.013.3

Leetaru, K., & Schrodt, P. A. (2013). GDELT: Global data on events, location, and tone, 1979–2012. In *ISA annual convention* (1–49). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.686.6605

Lewin, K. (1947). Frontiers in group dynamics II. Channels of group life; social planning and action research. *Human Relations*, *1*(2), 143–153. https://doi.org/10.1177/001872674700100201

Lewis, J. (2012). Terrorism and news narratives. In D. Freedman, & D. K. Thussu (Eds.), *Media & Terrorism: Global Perspectives*, 257–270. SAGE Publications Ltd, https://doi.org/10.4135/9781446288429.n15

Linder, F., Desmarais, B., Burgess, M., & Giraudy, E. (2018). Text as policy: Measuring policy similarity through bill text reuse. *Policy Studies Journal*, *48*(2), 546–574. https://doi.org/10.2139/ssrn.2812607

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. Proceedings of the International Conference on Learning Representations (ICLR 2013). https://arxiv.org/abs/1301.3781

Mozer, R., Miratrix, L., Kaufman, A. R., & Jason Anastasopoulos, L. (2020). Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Political Analysis*, *28* (4), 445–468. https://doi.org/10.1017/pan.2020.1

Nicholls, T. (2019). Detecting textual reuse in news stories, at scale. *International Journal of Communication*, *13*, 4173–4197. http://ijoc.org

Rosengren, K. E. (1970). International news: Intra and extra media data. *Acta Sociologica*, *13*(2), 96–109. https://doi.org/10.1177/000169937001300202

Schultz, I. (2007). The journalistic gut feeling: Journalistic doxa, news habitus and orthodox news values. *Journalism Practice*, *1*(2), 190–207. https://doi.org/10.1080/17512780701275507

Sheafer, T., Shenhav, S. R., Takens, J., & Van Atteveldt, W. (2014). Relative political and value proximity in mediated public diplomacy: The effect of state-level homophily on international frame building. *Political Communication*, *31* (1), 149–167. https://doi.org/10.1080/10584609.2013.799107

Shoemaker, P. J., Eichholz, M., Kim, E., & Wrigley, B. (2001). Individual and routine forces in gatekeeping. *Journalism & Mass Communication Quarterly*, *78*(2), 233–246. https://doi.org/10.1177/107769900107800202

Shoemaker, P. J., & Vos, T. (2009). *Gatekeeping theory*. Routledge. https://doi.org/10.4324/9780203931653

Soroka, S. N. (2012). The gatekeeping function: Distributions of information in media and the real world. *The Journal of Politics*, *74*(2), 514–528. https://doi.org/10.1017/S002238161100171X

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Sui, M., Dunaway, J., Sobek, D., Abad, A., Goodman, L., & Saha, P. (2017). US news coverage of global terrorist incidents. *Mass Communication and Society*, *20*(6), 895–908. https://doi.org/10.1080/15205436.2017.1350716

Trilling, D., & van Hoof, M. (2020). Between article and topic: News events as level of analysis and their computational identification. *Digital Journalism*, *8* (10), 1317–1337. https://doi.org/10.1080/21670811.2020.1839352

Van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Van Belle, D. A. (2000). New York times and network TV news coverage of foreign disasters: The significance of the insignificant variables. *Journalism & Mass Communication Quarterly*, *77*(1), 50–70. https://doi.org/10.1177/107769900007700105

Weimann, G., & Winn, C. (1994). *The theater of terror: Mass media and international terrorism*. Longman.

Welbers, K., Van Atteveldt, W., Kleinnijenhuis, J., & Ruigrok, N. (2018). A gatekeeper among gatekeepers: News agency influence in print and online newspapers in the netherlands. *Journalism Studies*, *19*(3), 315–333. https://doi.org/10.1080/1461670X.2016.1190663

White, D. M. (1950). The Gatekeeper: A case study in the selection of news. *Journalism Quarterly*, *27*(4), 383–390. https://doi.org/10.1177/107769905002700403

Wijffels, J. (2019). *Udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the 'UDPipe' 'NLP' toolkit*. R package version 0.8.3. https://CRAN.R-project.org/package=udpipe

Wu, H. D. (2000). Systemic determinants of international news coverage: A comparison of 38 countries. *Journal of Communication*, *50*(2), 110–130. https://doi.org/10.1111/j.1460-2466.2000.tb02844.x

Zhang, D., Shoemaker, P. J., & Wang, X. (2013). Reality and newsworthiness: Press coverage of international terrorism by China and the United States. *Asian Journal of Communication*, *23*(5), 449–471. https://doi.org/10.1080/01292986.2013.764904