Semi-automated Content Analysis of Media Frames

This is a work in progress.

Johannes B. Gruber

University of Glasgow

j.gruber.1@research.gla.ac.uk

Abstract

Successful applications of framing analysis using automated content analysis methods have been notably rare. One of the reasons is that frames are an abstract variable, notoriously hard to identify and code in content analysis, which can make automatic classification difficult as well. To overcome this issue, I propose a supervised learning approach to content analysis of media frames. The procedure first codes frame elements, based on Entman's (1993) definition, in a training sample before applying a collection of supervised machine learning algorithms to reproduce the classification on a larger dataset. In contrast to previous approaches, actual frames are identified only after the coding, by clustering elements into groups. Coding frame elements instead of holistic categories improves reliability and validity of manual coding results and could also improve supervised learning classification of frame elements. This approach has several further advantages over other automated content analysis methods as it: gives the researcher control over what is important for category building – in contrast to e.g. topic models; allows a more direct comparison with results from the vast body of framing studies; and can employ already implemented techniques to validate results of supervised learning methods. To test the approach, I analyse newspaper articles about protest and demonstrations, published in eight major UK national newspapers from 1992-2017. Findings pave the way for large scale framing analysis and contribute to our understanding of reporting patterns on protests in mainstream news media. [*]

Word count: 8,989

I. Introduction

Framing is arguably one of the most often employed concepts in communication and media research (as highlighted by overviews such as Borah, 2011; D'Angelo and Kuypers, 2010; Matthes,

[*]**Disclaimer: This paper is a work in progress! Some results and tests are still pending. Updates of the paper can be found at JohannesBGruber.eu/papers/JGruber_Semi-automated-Content-Analysis-of-Media-Frames.pdf.**

2009; Entman et al., 2009; D'Angelo, 2002). Its underlying main theme is that in order to transform the plethora of information that is produced by any event into a consisted narrative by selecting and organising certain aspects of what is happening and omitting others (e.g. de Vreese et al., 2001; Goffman, 1974). In communication, framing, therefore, means pre-selecting and emphasising some information while disregarding other, following a central organising idea in order to tell a coherent story (Entman, 1993; Gamson and Modigliani, 1989). Doing so, communication actors are thought to affect whether their audience notices certain aspects of an issue or event, potentially influencing how people evaluate and choose to act upon them (Chong and Druckman, 2007a; Druckman, 2001; Entman, 1993). This is what makes framing inherently interesting for communication researchers—especially for those who study mainstream media—as scrutinising the patterns of these choices provides a way to systematically study discourse, even when the underlying language and arguments change over time or between specific topics. Framing can furthermore connect research on a levels of a communication process as it is thought to be used consciously by communication actors (*strategic framing*), by journalists to make sense of a story (*journalistic frames*), can be observed in media content (*media frames*) and finally have an impact on the audience (*framing effects*) (Entman et al., 2009; Matthes, 2013).

However, *how* frames can be extracted reliably from text in order to be analysed is still disputed. In fact, meta-studies such as Matthes (2009) or Entman et al. (2009) reveal a plethora of approaches on how frames should be operationalised and measured. Meanwhile, successful applications of framing analysis using automated text analysis have been relatively rare. [1]. A reason for both is that frames are an abstract variable, notoriously hard to identify and code in content analysis (Matthes and Kohring, 2008; van Gorp, 2005).

To overcome this issue, I propose a supervised learning approach to content analysis of media frames. The procedure follows three main steps: First, a coding scheme is created based on the notion of frame elements found in Entman's (1993) definition. Secondly, a proportion of documents (training set) is coded manually using this scheme. Thirdly, patterns of frame elements used together in multiple documents are identified and interpreted as frames. In contrast to previous approaches, this means that actual frames are identified only after the classification by clustering elements into groups. Finally, I use machine learning document classification to assess the remaining documents.

Coding frame elements instead of holistic categories improves reliability and validity of manual coding results and can also improve supervised learning classification of frames. The approach has several advantages over other automated content analysis methods as it: gives the researcher control over what is important for category building – in contrast to, for example, topic models; allows a more direct comparison with results from the vast body of framing studies; and can

[1] With only a few exceptions: Card et al., 2016; Klüver and Mahoney, 2015; Burscher et al., 2014; van Atteveldt et al., 2013 Several researchers have claimed that fully automated classification using Latent Dirichlet Allocation is able to discover frames. This idea is discussed further below.

employ already implemented techniques to validate results of supervised learning methods.

The aim of this article is to outline an empirical procedure to pave the way for reliable large-scale framing analyses. The rest of this paper is structured as follows: First, I review several existing approaches of manual framing analysis. In a second step, previous automated content analysis approaches to text classification are examined. Based on that, I present a measurement procedure that combines a method of frame extraction which is deemed reliable with automated content analysis methods. To demonstrate the usefulness of the approach, I analyse a corpus of newspaper articles about protest and demonstrations, published in eight major UK national newspapers from 1992-2017. Results show that, in contrast to previous findings, only a relatively small proportion of articles focuses on the violent aspects of protest or peculiarities and dubiety of protesters. Finally, I briefly discuss the benefits of the proposed method and outline implications for future research.

II. Literature Review

There are hundreds of studies that employ content analysis of frames as their approach.[2] In essence, these studies try to find patterns in how aspects of an issue or story are selected and made salient in media content and describe those in terms of different frames. However, how frames are extracted from the material differs considerably between studies. Studies have been using text-based and number based approaches, have studied frames inductively or deductively and have extracted generic or specific frames from the material. In their overview of the framing concept, Entman et al. (2009) distinguish between four basic approaches:[3]

The first ones are *Qualitative approaches*, which identify frames based on the interpretation of the text itself. Usually, the approaches are rooted in qualitative research traditions, proceed inductively, frames are described in-depth and little or no quantification of elements or the distribution of frames within a discourse is provided by the researcher. Since coding is complex, it is also more labour intensive than other approaches, which is likely why most of the studies in this category rely on small samples of text which renders generalisation difficult. Despite commonly thorough descriptions of the individual frames, it is often not quite clear how researchers arrive at their conclusions. Why did a specific number of frames emerge from the material and which are their distinctive features? If these questions can't be answered, a replication of a specific study can turn out to be very difficult. Additionally, a researcher employing an inductive qualitative approach runs risk of identifying frames "they are consciously or unconsciously looking for" (Matthes and Kohring, 2008, p.259). The procedure proposed by Pan and Kosicki (1993), which is often described as a linguistic approach, can also be subsumed under this category. They are based on the notion that frames can be identified by the use of the same vocabulary and semantic structures

[2] An overview is provided by Matthes (2009).

[3] Also see Matthes (2009) and Matthes and Kohring (2008) who arrive at similar categories.

between texts. Studies following this approach usually make coding more explicit as they describe the linguistic features that indicate a frame.

Secondly, *Manual-Holistic approaches* code frames as holistic variables usually in quantitative content analyses. The frames can be either derived from the literature or identified inductively in a pilot study of a small sample. As in qualitative approaches, validity and reliability of an approach depend on the transparency with which the study communicates the coding decisions. Otherwise, the assessment of frames remains a black box. Considering the immense complexity of a frame as a content analytical variable, results are believed to inevitably receive a subjective tint based on the researcher's perception of the material (van Gorp, 2007). A second problem could emerge from the relative inflexibility of the approach: Once a list of frames is defined, new frames can usually not emerge from the material. That means that if a relatively uncommon frame is found only later in the coding process, researchers will usually not change their initial assessment or will not notice a new pattern at all (Matthes, 2009).

In order to evade these problems, *manual-clustering approaches*, the third category of approaches, split up frames into sub-variables which are easier to code in content analysis. The results are then combined to frames by dimensionality reduction techniques, such as factor or cluster analysis—resembling the procedure of analysing the results of a regular questionnaire. Most often, frames are operationalised as a set of indicator questions that are relatively easy to answer by coders—usually coders are asked if a certain aspect is mentioned in the text or not (see e.g., Semetko and Valkenburg, 2000; de Vreese et al., 2001; Burscher et al., 2014; Card et al., 2015). This procedure increases the reliability and validity of framing analysis considerably as the decisions are far more transparent and subjective interpretation is reduced to a minimum. The task also becomes much easier for the coders as they receive clear instructions on what and how to code, which makes it easier to split up the workload and makes coding overall less labour intensive. However, a downside of this approach is that frames must be known ex-ante since they form the base of the indicator questions. A solution is to either draw frames from the literature or to conduct a pilot, similar to the ones conducted by manual-holistic approaches.

An alternative is to derive coding instructions directly from a definition of framing. The definition by Entman (1993), which is the most commonly used one (Matthes, 2009), seems the most suitable for this task:

> "To frame is to *select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation* for the item described" (Entman, 1993, original emphasis).[4]

One important advantage of this definition is that it is rather specific and can thus be used directly to operationalise frames. The *problem definition*, *causal interpretation*, *moral evaluation*, and *treatment*

[4]For a comparison of framing definitions see Matthes, 2009.

recommendation can be seen as the elements of a frame. Matthes and Kohring (2008) suggest that these elements can be further divided to arrive at content analytical variables which can be clearly defined and coded relatively easily. In this understanding of framing, a frame is a pattern of different frame elements used together in a text. The goal of this operationalisation is to reduce subjective decision making not only during the extraction of frames but also during the identification of frames to a minimum. This is because, in contrast to indicator questions, the procedure is flexible enough to identify new frames. Known frames from the literature can be split into elements, while, for example, a new *problem definition* can be added to an existing coding scheme during the coding. A downside of using this definition is that it promotes a specific understanding of framing. This renders it unsuitable to code other proposed frame types such as equivalency frames (Tversky and Kahneman, 1987), emphasis vs issue frames (Chong and Druckman, 2007b), episodic vs. thematic frames (Iyengar, 1991) and emotional vs. rational frames. Nevertheless, Matthes and Kohring's (2008) approach of dividing frames into frame elements seems promising to conduct valid and reliable framing analyses.

The fourth and final category of approaches was named *computer-assisted* by Entman et al. (2009) or *inductive quantitative computer-assisted studies* by Matthes (2009). Commonly these approaches employ automated content analysis methods to replace manual coding completely. Some examples will be discussed in the next section.

Early on, studies have tried to improve coding reliability and reduce manual labour by turning to computational methods of content analysis. Miller (1997), for example, proposes so-called "frame mapping", which employs a combination of cluster analysis and multidimensional scaling based on the co-occurrence of key terms in texts. The key terms are chosen based on frequency and researcher assessment of relevance to the topic. The results must be validated by a human. Otherwise, no manual coding needs to be done. This reduces subjectivity of the frame identification significantly, as the clusters are extracted by the computer via statistical methods instead of being found through interpretation by the researcher. However, some scholars believe that the frames extracted by these methods do not do the concept justice as they reduce frames to clusters of words which are used together while omitting other dimensions (Carragee and Roefs, 2004; Matthes and Kohring, 2008).

This point can also be made about common topic model approaches. While they do take word choices into account—which are an important part of framing (Entman, 1993)—other choices, regarding what information is presented in a text and what information is left out, do not play a role. Jacobi et al. (2015) have argued nevertheless that topic models, such as Latent Dirichlet Allocation, should theoretically wield results congruent with framing analysis. Yet, in practice, retrieved topics follow the analysed material closely and are prone to pick up words specific to story content while ignoring more subtle variables such as *moral evaluation* and *treatment recommendation* (Jacobi et al., 2015). As an example, Jacobi et al. (2015) tried to replicate the

pioneering framing analysis study by Gamson and Modigliani (1989). However, while the original study identified several "anti-nuclear" frames, no topic could be clearly linked to that frame—even though reports about nuclear accidents and the dangers of nuclear power mostly employed this frame. An exception to this problem might be homogeneous corpora in which all documents focus on the same or very similar stories and only differ from each other in the choice of vocabulary.

Ordering automated content analysis approaches on a continuum ranging from inductive to deductive, clustering of words and topic modelling are probably the most inductive ones (Boumans and Trilling, 2015). Since they are "bottom-up" with very little or completely without any assumptions of what is meaningful in a corpus of texts, one could say that the discovery of patterns—although not the interpretation—is left to the computer. On the other hand of this spectrum are deductive approaches—i. e. where a researcher knows beforehand what they are looking for. The most deductive approaches are counting- or dictionary-based techniques (Boumans and Trilling, 2015). Using this method the computer counts a series of pre-defined character sequences and sometimes syntactic rules of how they co-occur. Sometimes this is directly used to reveal frames (Roggeband and Vliegenthart, 2007), while other studies use keywords to measure concepts, which are then combined to frames (Shah et al., 2002; Ruigrok and van Atteveldt, 2007). However, constructing the initial list of words (i. e. dictionaries) to use for coding requires detailed knowledge of the material as well as the frames they intend to code. After that first step, the validity of the dictionary needs to be checked and often words need to be added or removed from the list. But even after multiple iterations, the method runs risk of omitting some important words which do not occur often, yet are central to a specific frame (Reese, 2010). Dictionaries also do not 'travel well' into different domains as words often have specific meanings in different contexts (see, e.g., Loughran and Mcdonald, 2011) Furthermore, the method re-introduces a great deal of subjectivity as researchers are actively looking for words they believe are suitable to distinguish between concepts or frames.

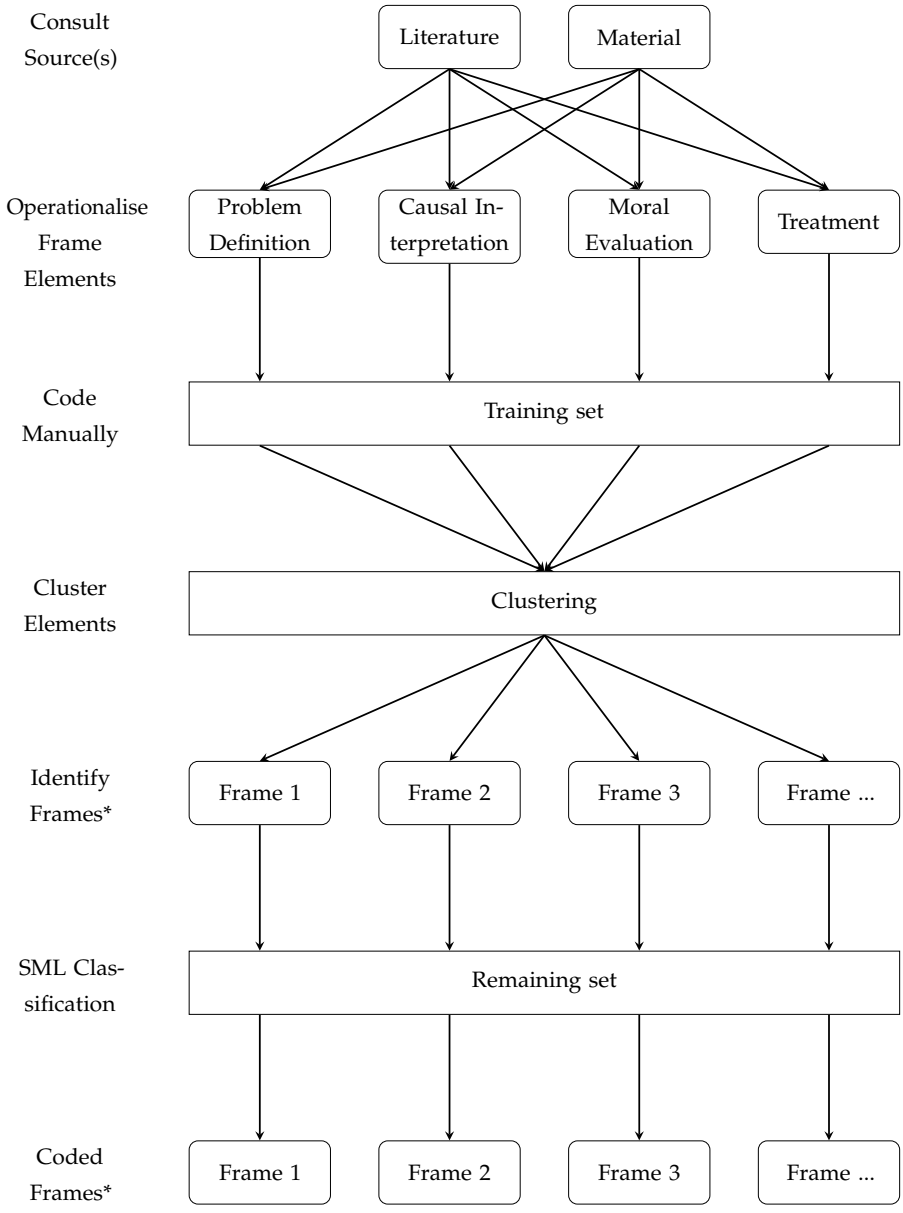
An intermediate between inductive and deductive automated content analysis methods are supervised machine learning (SML) approaches (Boumans and Trilling, 2015). SML has the advantage that methods aim to reproduce manual coding—in contrast to, for example, fully automated methods (Maier et al., 2018). This lays out a clear path for validation as traditional methods such as Krippendorff's alpha can be used to assess agreement between the results produced by a computer and the human gold standard—just as if the computer was another coder (Grimmer and Stewart, 2013). Alternatively, the machine learning literature also often measures precision and recall. Precision means, in this case, the percentage of how often a trained algorithm chooses the same category as human coders. Recall is the percentage of articles an algorithm classifies differently than a human coder in each category. This distinction is important since different usage scenarios might require a high precision (fewer false positives) or a high recall (fewer false negatives)—a high score in both is, of course, the best outcome (Eisenstein, 2019). If

the goal is to remove irrelevant documents from a set of articles, for example, false positives can be removed later in the process, while false negatives are permanently removed from the set, making a high recall for the relevant category more desirable.

In practice, the manually coded material is split into two groups: a larger sample called training set, which is used to "train" the supervised learning algorithms how to classify texts, and a smaller test set. Regarding the size of these samples, tests have shown that after increasing the training set to a size of roughly 500 documents, algorithms only incrementally get better by adding more hand-coded documents to the training set (Hopkins and King, 2010), even though rarely occurring categories might need more input to be validly coded (Burscher et al., 2014).

To code frames, SML has—to my knowledge—only been employed once before. Burscher et al. (2014) used indicator based questions to manually code a sample of more than 11,000 newspaper articles. The four generic frames they code were derived from the literature, as is necessary when employing indicator questions. As different SML algorithms have different strengths and weaknesses, Burscher et al. (2014) use a so-called *ensemble of classifiers*, consisting of Support Vector Machines (SVM), (Joachims, 1998), a Polynomial SVM classifier (Chang et al., 2010), and a Perception algorithm (Lippmann, 1987). Burscher et al. (2014) also tested if SML performed better on each indicator question separately, or if the algorithms performed better when texts were first classified using cluster analysis before the classified data was fed to the computer. In short, they coded the material twice, once using indicator-based and once using holistic coding. Interestingly, while the indicator-based approach is thought to improve the validity and reliability of manual coding, the algorithms performed better when coding holistic frames. Burscher et al. (2014) suggest that while indicator based approaches are better suited to "translate" theoretical concepts into variables which can easily be coded by humans, SML algorithms base their decisions on statistical analysis of vocabulary. While humans would perform better with a small set of questions to keep in mind, the computer usually works better when using a larger amount of indicators, in this case all words in a text (Burscher et al., 2014).

All of the above approaches can be used to extract either generic or specific frames. Specific frames can be very narrow and focus on a single event or on a number of similar events. A generic frame, on the other hand, can comprise a number of different topics. However, there is some debate on how generic a frame can be until it cannot be regarded as a frame any more. The distinction between episodic and thematic frames, made by Iyengar (1991), for example, is today often regarded as a choice between meta-categories of framing—as any frame can be described as either thematic or episodic (Callaghan and Schnell, 2002)—or a distinction between news formats (van Gorp, 2007).



**The actual number of frames varies between projects*

Figure 1: *Workflow: Code Frames*

III. Method

Figure 1 shows a flowchart of the proposed method. The proposed method consists of a combination of Matthes and Kohring's (2008) approach for a manual content analysis of media frames (step 1-5 in Figure 1) with supervised machine learning (step 6 and 7). In contrast to Burscher et al.'s (2014) procedure, frames are operationalised as a cluster of frame elements instead of indicator questions—which require frames to be known ex-ante—during the manual coding phase. First, the frame elements were coded manually based on Entman's (1993) definition in a training sample before applying a collection of machine learning algorithms to reproduce the classification on a larger dataset. In contrast to previous approaches, actual frames are identified only afterwards by clustering elements into groups using cluster analysis.

Each coded unit has a number of binary categories on which they can score either 1 if a certain variable is present or 0 if it is not. Frame elements are operationalised in content analytical variables, such that the *problem definition*, for example, can be coded by identifying the central issue or *topic* and the main *actor* in a text. [5] This results in a matrix with documents as rows and variables as columns (documents $\times$ variables), which can easily be used for cluster analysis. Frames then emerge by determining the optimal number of clusters and classifying units of text accordingly. Statistical methods to do so are widely available and implemented.[6] Burscher et al. (2014) suggest that machine learning algorithms perform better when reproducing the holistic frame categories instead of running classification of text for each variable. Therefore I chose to train the algorithms on the assigned cluster membership of the articles in the training set.

The procedure during manual coding follows closely the approach suggested by Matthes and Kohring (2008). In comparison to indicator based coding, this procedure has the advantage that frames do not need to be known ex-ante but are revealed by clustering the coded articles and determining which elements are most common in each cluster. Frames already identified in the literature can simply be divided into frame elements, while the method is flexible to incorporate newly identified elements during manual coding. A similar approach has—to my best knowledge—not been implemented so far.

In theory, the procedure allows a few different units of analysis: the whole article could be employed which maximises the chance that each unit will have information about each of the frame elements. However, as reporters will often try to at least make their stories appear balanced to some degree, different parts of a story might stress different views, which can make coding decisions harder for humans. Consequently, coding on the sentence level might appear to make decisions for easy decisions. In reality though, sentences often do not contain enough information to make any decision at all. The paragraph thus probably often is the best compromise. They are usually considered the smallest unit of meaning, while sentences rely on the context of the

[5] A more detailed explanation is given in the section Codebook and Manual Coding below.

[6] For example in the `R` package `NbClust` (Charrad et al., 2014).

paragraph and hence, would typically not convey a frame on their own (Jasperson et al., 1998).

The method has several advantages over other approaches: Compared to fully automated methods, such as topic modelling, it gives the researcher control over what is important for category building. In contrast to deductive methods, such as dictionary-based methods or supervised learning approaches based on indicator questions, it allows the researcher to add new variables during manual coding. Furthermore, since the frames are identified by clustering individual frame elements, rather than directly by the researcher, an additional layer of subjectivity is removed from the procedure (Matthes and Kohring, 2008).

IV. Application: Framing protest

To demonstrate the usefulness of the proposed procedure, I analysed the framing of protest events by mainstream newspapers in the United Kingdom from 1992-2017. A review of the literature on the topic suggests that there is a long-standing quasi-consensus that the *protest paradigm*, a default journalistic frame—i. e. a schemata journalists employ to select main issues and narratives for a story (Entman et al., 2009)—determines media reports of protest (Chan and Lee, 1984; McLeod and Hertog, 1992, 1999). When adhering to this paradigm, journalists would highlight the clash between protesters and police, the spectacle surrounding an event and unconventional outfits and hairstyles of protesters, while the message behind an event is generally neglected. However, so far two important gaps remain in the literature: most studies so far focused on the existence of the *protest paradigm* in their empirical studies, leaving the question open how prevalent it really is; and second, most studies only focus on a specific protest event or a limited time frame, while the landmark studies in the field are several decades old. A long-term empirical study which assesses the coverage on a broader scale is, as to my knowledge, missing so far. This might be an especially important gap, since the media and society went through important changes, many driven by new technologies, in the last decades, which could have wide-ranging influences on how protest is reported (Cottle, 2008).

The application of the method follows the research questions:

1. *Which frames are used by mainstream news media to portray protest?*
2. *How, if at all, does the mix of frames change over time?*

i. Data

Data for the study consists of newspaper articles from eight major UK national newspapers published from 1992-2017. The articles were retrieved from the commercial newspaper archive *LexisNexis*. After choosing the newspaper outlets which best represent the population of interest, a pilot study was conducted, which entailed downloading a sample of possibly relevant articles, coding a random subset manually and summarising the results to identify relevant keywords. The

initial list of keywords was identified using literature which describes different protest form (Tilly, 2012; Della Porta, 2013), an online dictionary with synonyms of "protest" and "demonstrations" (Thesaurus.com, 2018) and the Wikipedia entry on "protest" (Wikipedia, 2018). This was done as King et al. (2017) suggest that automated content analysis methods can often be biased as they fail to include all keywords that are relevant while building a database. From the pilot it was concluded, however, that all relevant texts can be retrieved using the keywords "protest", "demonstrations" and several variations of the two.[7]

Using the key phrase determined by the pilot, newspaper articles from eight UK outlets were then retrieved: *The Mirror, The Sun, Daily Mail, The Guardian, Observer, The Independent, Financial Times, Daily Telegraph and The Times*. The downloaded raw articles were read into `R` (R Core Team, 2019) using the package `LexisNexisTools` (Gruber, 2018) and extensive cleaning was performed to eliminate duplicates, irrelevant articles (i. e. false positives) and articles about foreign protests. Duplicates in the *LexisNexis* database are one of several limitations for studies relying on the digital newspaper archive and need to be removed before analysis (Deacon, 2007). Two or more articles were regarded duplicates if they were published on the same day, had a cosine word similarity of 0.95 or above and a relative Levenshtein distance of 0.2 or less to another article in the database (Gruber, 2018). Including highly similar articles in this definition is necessary as *LexisNexis* collects several version of the same article, often with minimal changes made for different (local) editions. Another problem was that some of the articles were from sources other than the requested ones, did not contain any of the keywords or contained only a description of a graphic—whereas the image itself was removed by LexisNexis—and were otherwise empty. This behaviour of LexisNexis was remarked by (Deacon, 2007) already and represents problems by the database. The respective articles were therefore removed.

Articles were deemed relevant if they concerned a domestic public protest of more than one actor. If, for example, a single member of parliament protested against a government decision, this was excluded from the data. One difficulty of data preparation was thus to eliminate false positives from the dataset. However, this step was deemed crucial, since "protest" and "demonstrations" are relatively common words and a large number of off-topic texts would significantly spoil SML performance.

In a first cleaning step, a list of keywords and phrases indicating irrelevant content, such as "FOOTBALL:", "TENNIS:" or "Obituaries:", was used to remove articles. The list of words was created and constantly extended during manual coding. Since the database still contained a

[7]The full key phrase was "protest OR protests OR protesters OR demonstration OR demonstrations OR demonstrators OR demonstrating OR protesting OR protester OR protestations OR protestors OR demonstrator OR counter-protesters OR undemonstrative OR protestor OR enltrprotesters OR counter-demonstration OR counter-protest OR counter-demonstrators OR protester's OR demonstraters OR protestsreflect OR protest-related OR counterprotests OR counter-demonstrations OR counterprotest OR demonstration's OR minidemonstration OR eco-protests OR protest-marches OR eco-protest OR protest's OR eco-protestors OR eco-protesters OR protest-as-theatre OR counterdemonstrators OR counter-protests".

large number of irrelevant articles, Latent Dirichlet Allocation was used to divide the dataset into "topics" (Blei et al., 2003; Grün and Hornik, 2011). After several runs using different numbers of topics, it was concluded that 75 topics were fine-grained enough to distinguish articles thematically, yet the number of topics was still small enough to manually assess the outcome. For each topic, the words with the highest word probabilities were assessed and a sample of 20 articles per topic (1,500 in total) were manually coded for relevance. LDA was effective in clustering false positives of several kinds, such as obituaries, sport events or celebrity gossip. From each topic, 25 articles were randomly chosen and coded. If not at least one of these contained description of a protest, the topic was removed.

As a second step, articles referring to protest events outside the UK were removed. Reports about foreign protests are generally thought to be significantly different from reporting about domestic protest events (Boyle et al., 2012; Mueller, 1997). To remove these articles, the names of locations were identified through named entity recognition. The identified locations were then matched with the GeoNames dataset (GeoNames, 2018) to divide locations into two categories: inside and outside the UK. Where an article did not mention at least one location within the UK but at least one foreign location, it was removed from the dataset.

Another set of articles which was falsely selected contained so-called homonyms. In this case this meant forms of "protest" or "demonstration" which had nothing to do with public protest in the sense relevant here. Examples include "cookery demonstrations", "demonstration effect" and "protest vote". To make sure no relevant articles were removed just because they contain these words, the words were removed from the articles before checking again if they still contained any form of "protest" or "demonstration". If they did not, they were removed.

In a final cleaning step, 1,000 paragraphs were manually coded for relevance. Using Naïve Bayes classification, this training sample was used to identify irrelevant articles. To perform classification, the text was turned into a document feature matrix, where columns of the matrix represent the words in the data and rows represent the paragraphs. This step effectively turns words into numbers since the contained information is the frequency in which a word occurs in a text. Not all features of the text were preserved, however, as a list of stopwords suggested by `quanteda`, as well as any punctuation, numbers, symbols and URLs were omitted. The implementation of multinomial Naïve Bayes from the R package `quanteda` (Benoit et al., 2018) was found to work best for classification (precision of 0.81, meaning 81% of irrelevant articles were classified correctly as irrelevant, and a recall of 1, meaning no relevant paragraph was categorised as irrelevant by the model).

As can be seen in Figure 2, the cleaning removed a considerable amount of articles from the initial dataset. From the 598,587 articles retrieved from LexisNexis, 27,496 remained in the set after all cleaning steps. This means that 95.4% of the original set was discarded. The reasons for this are that "protest" and "demonstration" are common words in the English language and are used

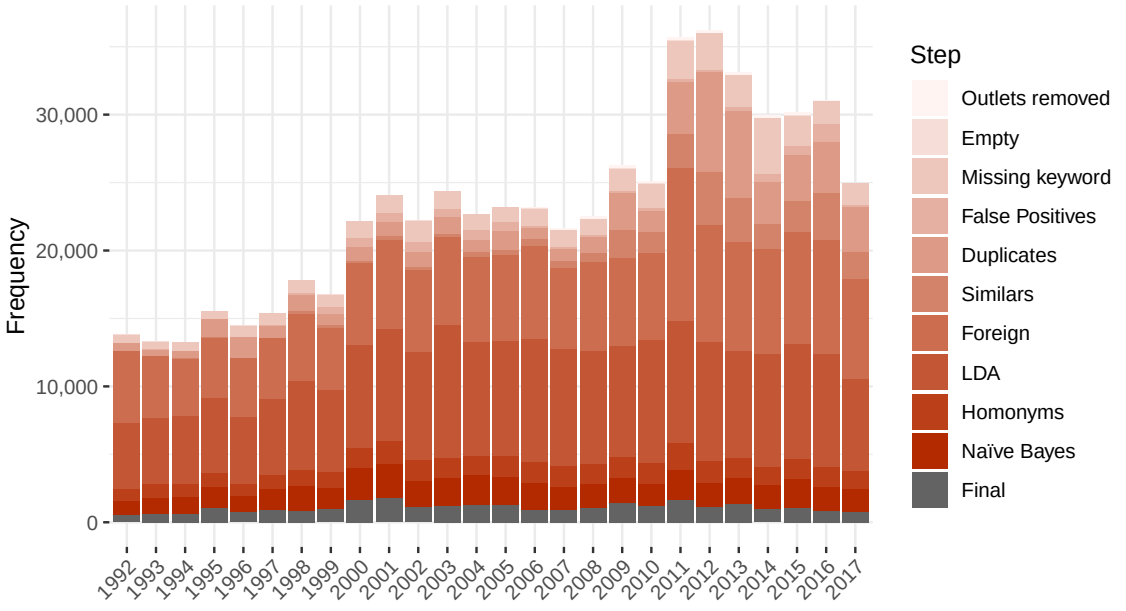


Figure 2: *Article cleaning by year*

in many different contexts, that British newspapers commonly report about foreign protest and that protest or protesters are often mentioned as a side note in other stories. What becomes clear from Figure 2 is that no valid research using automated content analysis can be done without validating integrity of the underlying dataset and possibly cleaning it.

ii. Codebook and Manual Coding

The first step of the procedure was to consult available sources and use them to construct a codebook based on the idea of frame elements (See step 1 in Figure 1). Entman's (1993) four frame elements, *problem definition*, *causal interpretation*, *moral evaluation*, and *treatment recommendation* themselves are still relatively abstract categories. Matthes and Kohring (2008) therefore suggest to further split them into coding variables. This formed the basis for developing the codebook which is shown in Table 1. For *problem definition* the idea was that it can be conceived by coding the *topic*, or central issue that is discussed, and the main *actor* described in a text. It should be noted that codes of the same elements are assigned to a paragraph mutually exclusive, which means only one *topic* and *actor* are coded on the paragraph level. *Moral evaluation* and *Causal attribution* are closely linked. The first one contains the most frequently mentioned *benefit* and *risk* described in a news item about protest, while the latter one measures who was identified to be responsible for said *benefits* and *risks*. Finally, the moral judgement is treated as a proxy for *treatment recommendation*. I assume that when protests are judged negatively, this implies an urge towards the protesters to

stop demonstrating.

This combination of codes can map the different types of stories found in media coverage. As an example, consider a paragraph in which the *topic* is the violence that occurred at a protest event. The *actor* in this case could be the police, a *risk* could be the destruction of property and this risk could be attributed to protesters, while the *benefit* of reinstating public order could be attributed to the police. This can be considered one possible framing of the story. However, a different combination of codes could alter the story significantly: Even with the same *topic* and *actor*, a totally different picture emerges if police are portrayed as responsible for the *risk* to public safety, for example if their use of force towards protesters is perceived as unprovoked.

The list of variables was developed in two stages as . The first one was to obtain suitable categories from the literature about media coverage of protests. Notably, McLeod and Hertog (1999) compiled a comprehensive set of characteristics typically employed in coverage of protest. Not only did this include elements of a story which tend to marginalise protest but also features that are deemed mixed or sympathetic towards demonstrations—assumed to be only found in alternative press outlets. These categories systematised and formed the core of what is commonly referred to as the *protest paradigm* (Chan and Lee, 1984): a set of common news coverage patterns or an implicit template that journalists are believed to follow when reporting about protest. McFarlane and Hay (2003) built on and extended the category system, as did Dardis (2006b,a) who conducted the first quantitative studies using, by then, a typology of 14 variables. Di Cicco (2010) added that over the years, the mainstream media started to portray protests mainly as an impotent nuisance, questioning the legitimacy of the practice as a whole.

The characteristics were "translated" into one or several coding variables. For example, the emphasis on clashes between protesters and the police—dubbed "violent crime story" (McLeod and Hertog, 1999, p. 312)—resulted in several variables: *Topic: Violence/Crime*, *Actor: Protesters* or *Actor: Police*, *Benefit attribution: Police*, *Risk attribution: Protesters*, *Benefits: Reinstating public order*, *Risk: Public safety* and *Judgement: Negative* (see Table 1).

The second set of variables emerged during coding (denoted with an * in Table 1). Additional categories were, for example, *risk attribution: Officials*, which emerged from reports about anti-Brexit protests where the news coverage highlighted potential risks which are caused by Britain leaving the EU; *topic: effect of protest*, which is found, for example, in the discussion of the 2000 fuel protest, that caused considerable economic damage and led the government to freeze fuel duties; and the *risk: Grievance* which highlights the negative effects caused by the issue a protest is addressing.

Manual coding was done on a random set of 500 articles. Articles were split into paragraphs and coding was done at this level. Paragraphs were coded in order of appearance and variables were assigned mutually exclusive within a paragraph. Where, for example, two topics or actors were present in a paragraph, the one that was featured more prominently was chosen. When

two variables were equally prominent, the one that was mentioned first was selected, as it is assumed that more important information is provided earlier in a news story. To accomplish a suitable data structure, codes were turned into binary variables with values being either 1, if a code was assigned to a document, or 0, if the code was not assigned. Coding the 500 random articles resulted in 4277 codes over 2534 paragraphs.

The decision to code variables on the paragraph level was made after a pilot phase showed that coding frame elements on the article level would involve demanding coding decisions as newspaper articles, especially longer ones, often switch perspectives between paragraphs. Often articles highlight what one actor has done or said in one part of the story and focus on another party in a later part. This makes it hard to pin down which is the main *topic*, *actor*, *risk* and so forth in an article. A paragraph, by comparison, usually comprises only one argument or thought. While this problem is not completely avoided by coding on the paragraph level, decisions are usually easier, which leads to less demand for subjective judgement.

For the analysis, however, it seemed to make more sense to aggregate the paragraphs back into articles since different frame elements were, most of the time, spread over multiple paragraphs. Paragraphs were aggregated by taking the mean value for each variable. The values for each could therefore range anywhere between 1, if the code was always chosen, and 0, if it was never chosen in an article. This way, information on whether an article does try to balance between different views is preserved while arguments spread over multiple paragraphs of an article are nevertheless captured. Nevertheless, not all articles contained the full set of frame elements. This poses no issue, however, as not all frames necessarily contain all elements (Matthes and Kohring, 2008).

Table 1: *Variables and Codes for manual content analysis*

Frame Element	Variable	Description
Problem Definition	Topic: Event/Spectacle	Description of the event (e.g., size, marching route or what protesters did) often highlighting the entertaining or spectacle aspects (emptying protest of its political significance).
	Topic: Violence/Crime	Violence, vandalism and destruction of public or private property surrounding a protest.
	Topic: Clash	Confrontation with the police, not necessarily violent.
	Topic: Protesters	The appearance, mental ability, visual deviance and oddities of the protesters (including pathologising their protest group, social movement or subculture) or their 'real' underlying motives.
	Topic: Confrontation	The protests are part of a confrontation between two groups.
	Topic: Public Opinion	The protest represents a minority/majority of public opinion (operationalised through polls, interviews or reference to norms).
	Topic: Cause	The cause of why a protest took place or the goals of the protesters.
	Topic: Nuisance	The protest caused inconvenience to regular citizens and the government.
	Topic: Media*	Scrutiny of the attention the media spent on a protest.
	Topic: Protest as Tool*	Discussion of the practice of protest itself (e.g. its legitimacy, tactics or potency) .
	Topic: Other Activism*	Other actions by the same group who initiated protest (like letters or action in courts)
	Topic: Policing Tactics*	Discussion about how the police or other security forces should deal with protesters
	Topic: Effect of Protest*	Discussion about the effect a protest had (e.g., starting public debate about a topic or causing disruption)
	Topic: Judicial Persecution*	Persecution of protest related actions in court (e.g., court case about violence of protesters or police)
	Actor: Protesters	The people engaging in a protest.
	Actor: Police	Members of the police force engaging with protesters.
	Actor: Officials	Representatives of government organisations.
	Actor: Business	Representatives of business organisations.
	Actor: Other Political Elite*	Other political players such as members of the opposition party or regional parliaments
	Actor: Other*	Other non-elite actors (e.g. motorists, local residents, counter-demonstrators or unidentified people)
Causal Attribution	Benefit Attribution: Police	Police is responsible for the benefit (e.g., by reinstating order).
	Benefit Attribution: Protesters	Protesters are responsible for benefit (e.g., by being entertaining).
	Benefit Attribution: Officials	Officials are responsible for benefit (e.g. by preventing chaos protesters try to inflict).
	Benefit Attribution: Other Pol. Elite*	Other political elites are responsible for benefit (e.g. by proposing a solution to a problem targeted by protesters).
	Risk Attribution: Protesters	Protesters are responsible for risk (e.g. vandalism, attacks on police or nuisance).
	Risk Attribution: Police	Police are responsible for risk (e.g. unnecessary clashes with peaceful protesters).
	Risk Attribution: Business*	Business actors are responsible for risk.
	Risk Attribution: Officials*	Officials are responsible for risk (e.g. by signing a bad law).
	Risk Attribution: Media*	The media are responsible for risk (e.g. problematic reporting).
	Risk Attribution: Other Pol. Elite*	Other political players such as members of the opposition party or regional parliaments are responsible for a risk
	Risk Attribution: Other*	Other actors are responsible for a risk, often by attacking protesters (e.g. motorists, local residents, counter-demonstrators or unidentified people)
Moral Evaluation	Benefit: Expression / Entertainment	The protest was aesthetic, artistic or emotionally expressive.
	Benefit: Reinstating Public Order	Usually attributed to police who reinstate public order after it was strained by the protest.
	Benefit: (Just) Cause*	Usually attributed to protesters when their actions are seen to as struggle for a good cause.
	Benefit: Initiated Public Debate*	Usually attributed to protesters when their actions have caused a (necessary) public debate.
	Risk: Public Safety	The protest as a risk for public safety.
	Risk: Property Destruction	Property was destroyed due to protest
	Risk: Dulling Political Discussion	The protesters spoil serious political debates with their childish, insane or uninformed arguments or false claims.
	Risk: Decay of Morals or Other Social Norms	The protesters are a fringe group of freaks who set a bad example for others and disturb the general political consensus.
	Risk: Unpatriotic Behaviour/Treason	The protest is unpatriotic or hurts the nation.
	Risk: Nuisance	Protest is bothersome to citizens and the government yet impotent as a political tool.
	Risk: Costs of Demonstrations	Clear up and police deployment costs of demonstrations burden public budget.
	Risk: Grievance*	The grievance which the protest is addressing is a risk.
	Risk: Censorship*	Usually attributed to police or officials who allegedly try to silence or discredit protest.
	Risk: Suppression*	Usually attributed to police or officials who allegedly try to undermine protest or make it impossible.
	Risk: Harm Discussion*	The protest or other actions harm the discussion about a topic
	Risk: Breaking Laws*	Breaking the law but neither destroying anything nor harming anyone (e.g. entering parliament without permission)
	Risk: Bad for Business*	Protests have negative impact on business revenues.
Treatment	Judgement: Negative	The protests were/are bad (e.g. protesters should go home).
	Judgement: Positive	The protests were/are good (e.g. protests made aware of a problem).

* Categories found inductively from the coded material instead of the literature

iii. Cluster Results

After manual coding was finished, frames were revealed through cluster analysis by grouping together articles which are similar in terms of the frame elements they employ (see Figure **??**). Cluster analysis is a statistical procedure in which multivariate data, in this case codes assigned to each document, are grouped together so that cases similar to each other belong in one group and cases dissimilar to each other belong to different groups (e.g., King 2015). This is appropriate, since the assumption of the procedure is that articles using the same frame will be similar in terms of the *problem definition*, *causal interpretation*, *moral evaluation*, and *treatment recommendation* they employ. By grouping together similar articles, recurring patterns become visible and the frames they employ can be determined empirically.

To perform the clustering, the k-means algorithm implemented in `R` (R Core Team, 2019) was chosen. K-means continues to be one of the most popular approaches for partitioning large datasets. A crucial step in k-means cluster analysis is thus to decide on an optimal number of starting centres and, hence, clusters (k). In the literature, a considerable number of indices and approaches have been suggested and explored, yet which one is the right procedure might vary significantly between datasets (Milligan and Cooper 1985). A systematic and elegant solution is offered by `NbClust` (Charrad et al., 2014): instead of relying on a single index or approach, a number of different ones are calculated. Taking stock of currently available indices and extending the list, Charrad et al. (2014) arrive at a number of 30 indices which are included in the package. Most of these indices offer clear optimal solutions as measures are either minimised or maximised over different clustering solutions, making an assessment fully automatic. It can then be evaluated where different indices agree with one another on the optimal number of clusters. Charrad et al. (2014) show that by using a simple majority rule among the different solutions, this approach outperforms any individual index in determining the best number of groups in real as well as simulated datasets where the "real" number of groups is known beforehand.

```
## Warning:  Use of ‘count$n‘ is discouraged.  Use ‘n‘ instead.
```

The result is robust as well if the maximum number of clusters is increased. The only difference in that case is that the indices which arrive at a the ten-cluster solution choose solutions with even more clusters. Note that Figure 3 does not display the optimal solutions for all 30 indices that were calculated. This is the case as some indices do not offer one clear results by design, similar to the "elbow" criterion were human judgement is needed.

Figure 3 shows the result of running `NbClust` on the article data. A majority of indices suggest that three is the optimal solution. [8] This result is robust between different cluster analysis methods and distance measures. Solutions that were taken into account ranged from two to twenty-five.

[8] Seven of twenty-four indices agreed on this solution. The other algorithms either did not converge or offer a visual result, which is not shown here.

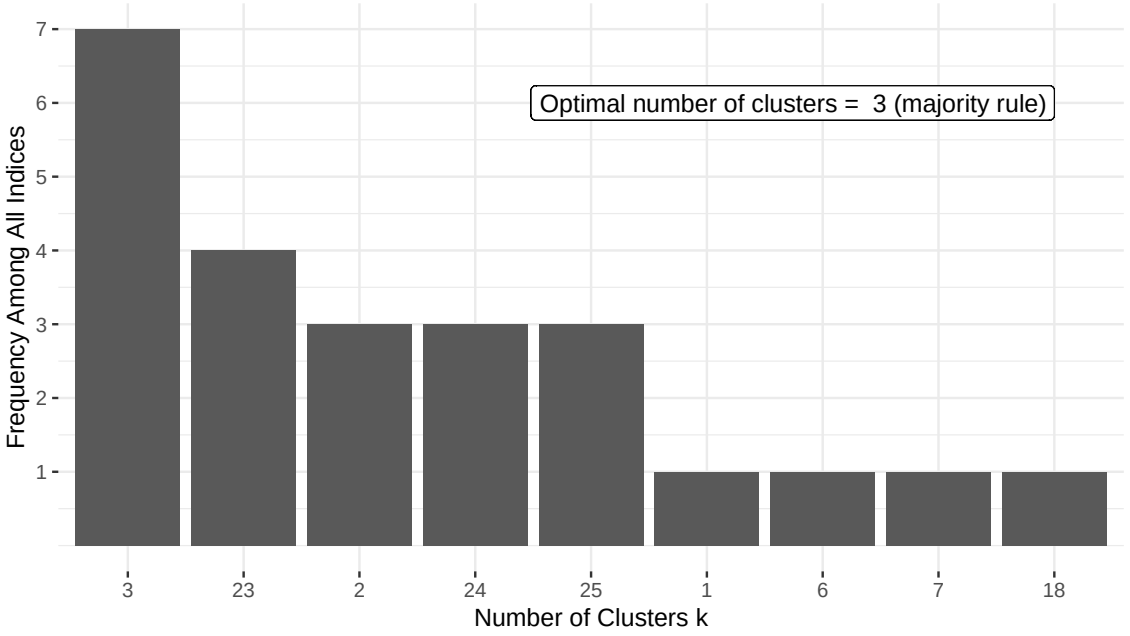


Figure 3: *Optimal number of clusters*

When the maximum was increased or lowered, three still stayed the optimal solution for most indices. Additionally, the second and third most often chosen solutions were manually assessed and eventually discarded, since the three-solution appeared to make most sense when interpreting the results.

iv. From Clusters to Frames

After clustering, the cluster means were used to describe and interpret clusters as frames. Again, this makes theoretical sense as frames are perceived as a recurring pattern of the same frame elements used in a number of different articles. The cluster means describe the average proportion of paragraphs in articles in one cluster which use a specific variable. Specifically, codes in the categories *topic*, *actor*, *benefit attribution*, *risk attribution*, *benefit*, *risk* and *judgement* are aggregated to a means between 1—if all paragraphs in all articles in the cluster use the code—or 0—if a code is not used at all in a cluster. This means for the interpretation of the clusters as frames, higher mean values indicate the most often assigned codes in a variable.

```
## Error in doColorRamp(colorMatrix, x, alpha, ifelse(is.na(na.color), "", :  could
not find function "doColorRamp"
```

Figure 4 shows the cluster means as a heatmap with high values tending towards red and low values tending towards yellow. Frames were named after their most distinct features: *Balanced*,

Figure 4: *Cluster means*

Struggle for (Just) Cause and *Troublemakers*. Three things are important when interpreting the means as frames. First, the highest values in a cluster are the most important codes for a frame. For Struggle for (Just) Cause, for example, nearly every time an actor was coded, it were the protesters (94%), indicating who was the driving force behind the struggle. Second, relatively low cluster means can still be important if they are high compared to other clusters. Negative judgement, for example only really occurs in the Troublemakers frame while it is rarely used in articles in the *Balanced* frame and mostly absent in the *Struggle for (Just) Cause frame*. Third, low values can also be significant if it distinguishes one frames from the others. In the *Troublemakers* frame, for example, the topic Cause is nearly absent (6%) compared to the other two frames.

These three frames are clearly distinguishable from each other. The frame named *Troublemakers* is the smallest cluster comprising 23.0% of the articles in the manual coding sample. The focus in articles in this cluster lies on violence and crime (mean in topic: 0.33) which is most often attributed to protesters (0.58). Either the protesters (0.49) or the police (0.23) appear as actors. The main concern is public safety (0.64). The frame contains a judgement more often than any of the other frames and it is always negative. However, the absolute values is still low with . What becomes clear from this frame though is that even without explicit judgement, the selection of codes in the variables clearly show that protests are seen negatively by articles in this frame.

The *Troublemakers* frame therefore has a striking resemblance to the *protest paradigm* mentioned above. Yet, in contrast to the literature, the role of this frame seems not that significant.

The frame *Struggle for (Just) Cause* is the second biggest cluster in the manual coding sample containing 36.7% of the articles. Most noteworthy here is that *cause* is that the most prevalent topic (0.44) and *grievance* by far the most important risk (0.36). As mentioned above, the protesters are nearly always the actor signalling a degree of agency in this frame that surpasses even the one in the Troublemakers frame. Interestingly, the topic *spectacle* has the highest proportion it has in any frame here (0.16). Contradicting previous findings, this indicates that even when reports pick up the spectacle and entertaining parts of a protest event, this does not seem to lead to detrimental coverage. The frame also focuses more than others on reporting the event itself and what protesters do, indicated by the comparatively high average for the *event* topic (0.22) (McFarlane and Hay, 2003). Overall, it seems that this frame, making up more than a third of the articles in the sample, is rather sympathetic to the protesters conduct and goals.

The final and largest frame—at least in the sample—was called *Balanced*. At first, the cluster appeared to signal a problem with the model as no clear tendency towards possible interpretations seemed sensible. In fact, the *Balanced* frame is characterised by a lack of clear dominance of one code in any of the variables. Descriptions of the *event* are the most prevalent topic, but only at a low level of just 0.15). Judicial persecution, effects of protest, Cause and Confrontation basically all range on a similar level as topic. Unlike in the other clusters, there is no clear tendency if benefits or risks overweight. Positive judgement is a little more common than negative but only at a mean

of 0.05). However, after consulting the articles in that cluster, it appears that they usually paint a rather balanced picture of the protest. Often articles devote equal attention to both sides, the cause of an event as well as the clash with police. This is in line with professional norms in journalism.

This answered the first research question: mainstream news media portrays protest using one of three frames that were identified here. One of these frames was described before in several studies as the *protest paradigm* and believed to marginalise protester's messages before they can reach the audience (Chan and Lee, 1984; McLeod and Hertog, 1992, 1999; McFarlane and Hay, 2003). It was found, however, that this frame is far from being as dominant as expected in the British press.

v. Classification results: framing over time

[This is work in progress! The results are only based on the coding of 300 articles so far. While the accuracy looks promising, more articles need to be added to the training sample. This should increase performance of the algorithms to a acceptable level.]

Table 2: *Accuracy measures for different SML algorithms*

model	Accuracy	Low	High	package
Maximum Entropy	0.64	0.44	0.81	RTextTools
SVM	0.59	0.39	0.78	quanteda.classifiers
LogitBoost	0.59	0.36	0.79	caret/caTools
bagging	0.50	0.31	0.69	RTextTools
Naive Bayes	0.48	0.29	0.68	quanteda
Random Forest	0.48	0.29	0.68	caret/ranger
NNSEQ	0.44	0.25	0.65	quanteda.classifiers
Penalised Multinomial Regression	0.44	0.25	0.65	glmnet

V. Conclusion

The promise of framing analysis is that it can uncover patterns of reporting and potential biases in a diverse collection of content. Yet scholars have pointed to a lack of reliability and validity arising from the demanding tasks to identify and code frames. Complicated coding decisions also tend to make framing analysis labour intensive and extensive training is often needed for coders before they can work on a project.

In this paper I introduce a procedure which aims to alleviate these problems: the idea is to divide coding decisions into smaller tasks by asking coders only to code compartmentalised

variables based on the frame elements introduced by Entman's (1993) definition of framing. The number of frames and their most distinct features are then revealed through statistical procedures, removing most of the subjective decision-making from the process. After applying this procedure on a small training set, it is used to train machine learning algorithms to replicate the classification on new documents This way, analysis can be scaled up to large corpora of similar texts with little or no additional manual effort.

Compared to other automated content analysis methods, the procedure has several advantages: fully automated methods, such as topicmodels, take much control from the researcher who has little impact—and often knowledge—on what is used for classification decisions. That means that any prior knowledge about a topic, collected by previous research or the researcher, can't be employed. Additionally, validation is demanding as the researcher not only needs to test if coding decisions fit human interpretation, but needs to figure out what the individual classes mean in the first place. In the proposed procedure, by contrast, the researcher has full control over the initial set of codes and variables used in the manual coding part of an analysis. Since an initial set of documents is already coded at this stage, the validity of the results can easily be tested by comparing classification by the machine with the human gold standard,

More deductive techniques of automated content analysis methods, such such as dictionary-based methods, require detailed knowledge of the categories, the material and an immense effort when first set up since hundreds or thousands of words need to be reviewed in order to determine if they can be used to distinguish frames from one another. This task can be difficult though since words can have different meanings in different contexts. That often means that even after several iterations, including certain words in one category of the dictionary can have unforeseen outcomes. Additionally, categories need to be clear ex-ante since when applied, the method will only consider words in the dictionary, not picking up related or similar terms. The procedure introduced here can also employ ex-ante categories but is not limited to these the initial knowledge of them. During the manual coding phase, unforeseen additional codes can easily be added to the coding scheme. Additionally, since machine learning considers not only a pre-defined dictionary of words but the mixture of all words in each document in the learning set at the same time, it is less susceptible omit important words that occur infrequently.

However, the method is not without limitations. One problem is that since it relies heavily on Entman's (1993) definition of framing, it inherits some of its problems. Mainly, how did Entman arrive at precisely four elements, and why are *problem definition*, *causal interpretation*, *moral evaluation*, and *treatment recommendation* the elements Entman chose? Also, how many of these elements need to be present to constitute a frame and could texts feature only parts of a certain frame? Entman (1993) states that "any particular text may not necessarily include all four [elements]" (p.52) yet in some texts, only a problem definition and actor were present. An alteration of the procedure that uses a different way of breaking up frames into smaller elements

is, however, possible.

References

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., and Matsuo, A. (2018). quanteda: An r package for the quantitative analysis of textual data. https://quanteda.io.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, http://dl.acm.org/citation.cfm?id=944919.944937.

Borah, P. (2011). Conceptual issues in framing theory: A systematic examination of a decade's literature. *Journal of Communication*, 61(2):246–263, doi: 10.1111/j.1460-2466.2011.01539.x.

Boumans, J. W. and Trilling, D. (2015). Taking stock of the toolkit. *Digital Journalism*, 4(1):8–23, doi: 10.1080/21670811.2015.1096598.

Boyle, M. P., McLeod, D. M., and Armstrong, C. L. (2012). Adherence to the protest paradigm: The influence of protest goals and tactics on news coverage in u.s. and international newspapers. *The International Journal of Press/Politics*, 17(2):127–144, doi: 10.1177/1940161211433837.

Burscher, B., Odijk, D., Vliegenthart, R., de Rijke, M., and de Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, 8(3):190–206, doi: 10.1080/19312458.2014.937527.

Callaghan, K. and Schnell, F. (2002). *Introduction: Framing Political Issues in American Politics*, pages 1–18. University of Pittsburgh Press.

Card, D., Boydstun, A. E., Gross, J. H., Resnik, P., and Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. *Proceedings of Association for Computational Linguistics Conference (ACL)*.

Card, D., Gross, J. H., Boydstun, A. E., and Smith, N. A. (2016). Analyzing framing through the casts of characters in the news. *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*.

Carragee, K. M. and Roefs, W. (2004). The neglect of power in recent framing research. *Journal of Communication*, 54(2):214–233, doi: 10.1111/j.1460-2466.2004.tb02625.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.2004.tb02625.x.

Chan, J. M. and Lee, C.-C. (1984). Journalistic 'paradigms' of civil protests: A case study of hong kong. In Arno, A., editor, *The news media in national and international conflict*, A Westview special study, pages 183–202. Westview Press, Boulder, Colo.

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6):1–36, http://www.jstatsoft.org/v61/i06/.

Chong, D. and Druckman, J. N. (2007a). Framing public opinion in competitive democracies. *American Political Science Review*, 101(04), doi: 10.1017/S0003055407070554.

Chong, D. and Druckman, J. N. (2007b). Framing theory. *Annual Review of Political Science*, 10(1):103–126, doi: 10.1146/annurev.polisci.10.072805.103054.

Cottle, S. (2008). Reporting demonstrations: The changing media politics of dissent. *Media, Culture & Society*, 30(6):853–872, doi: 10.1177/0163443708096097.

D'Angelo, P. (2002). News framing as a multiparadigmatic research program: A response to entman. *Journal of Communication*, 52(4):870–888, doi: 10.1111/j.1460-2466.2002.tb02578.x, http://dx.doi.org/10.1111/j.1460-2466.2002.tb02578.x.

D'Angelo, P. and Kuypers, J. A., editors (2010). *Doing news framing analysis: Empirical and theoretical perspectives*. Communication series. Routledge, New York and London.

Dardis, F. E. (2006a). Marginalization devices in u.s. press coverage of iraq war protest: A content analysis. *Mass Communication and Society*, 9(2):117–135, doi: 10.1207/s15327825mcs0902_1.

Dardis, F. E. (2006b). Military accord, media discord: A cross-national comparison of uk vs us press coverage of iraq war protest. *International Communication Gazette*, 68(5-6):409–426, doi: 10.1177/1748048506068719.

de Vreese, C. H., Peter, J., and Semetko, H. A. (2001). Framing politics at the launch of the Euro: A cross-national comparative study of frames in the news. *Political Communication*, 18(2):107–122, doi: 10.1080/105846001750322934.

Deacon, D. (2007). Yesterday's papers and today's technology. *European Journal of Communication*, 22(1):5–25, doi: 10.1177/0267323107073743.

Della Porta, D. (2013). Repertoires of contention. In Snow, D. A., Della Porta, D., Klandermans, B., and McAdam, D., editors, *The Wiley-Blackwell Encyclopedia of Social and Political Movements*, pages 43–46. Blackwell Publishing Ltd, Oxford, UK.

Di Cicco, D. T. (2010). The public nuisance paradigm: Changes in mass media coverage of political protest since the 1960s. *Journalism & Mass Communication Quarterly*, 87(1):135–153, doi: 10.1177/107769901008700108.

Druckman, J. (2001). The implications of framing effects for citizen competence. *Political Behavior*, 23(3):225–256, doi: 10.1023/A:1015006907312, http://dx.doi.org/10.1023/A

Eisenstein, J. (2019). *Introduction to Natural Language Processing*. The MIT Press, https://github.com/jacobeisenstein/gt-nlp-class/.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58, doi: 10.1111/j.1460-2466.1993.tb01304.x, https://doi.org/10.1111/j.1460-2466.1993.tb01304.x.

Entman, R. M., Matthes, J., and Pellicano, L. (2009). Nature, sources, and effects of news framing. In Wahl-Jorgensen, K. and Hanitzsch, T., editors, *The handbook of journalism studies*, International Communication Association handbook series, pages 163–173. Routledge, New York.

Gamson, W. A. and Modigliani, A. (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology*, 95(1):1–37, doi: 10.1086/229213.

GeoNames (2018). Geonames. http://www.geonames.org/about.html.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*, volume CN 372 of *Harper colophon books*. Harper & Row, New York.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, doi: 10.1093/pan/mps028.

Gruber, J. (2018). LexisNexisTools. An R package for working with newspaper data from 'LexisNexis'. https://github.com/JBGruber/LexisNexisTools, R package version 0.2.0.

Grün, B. and Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30, doi: 10.18637/jss.v040.i13.

Hopkins, D. J. and King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, doi: 10.1111/j.1540-5907.2009.00428.x.

Iyengar, S. (1991). *Is Anyone Responsible? How Television Frames Political Issues*. American politics and political economy series. University Of Chicago Press, Chicago.

Jacobi, C., van Atteveldt, W., and Welbers, K. (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4(1):89–106, doi: 10.1080/21670811.2015.1093271.

Jasperson, A. E., Shah, D. V., Watts, M., Faber, R. J., and Fan, D. P. (1998). Framing and the public agenda: Media effects on the importance of the federal budget deficit. *Political Communication*, 15(2):205–224, doi: 10.1080/10584609809342366.

King, G., Lam, P., and Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*, 61(4):971–988, doi: 10.1111/ajps.12291, http://j.mp/2nxUa8N.

Klüver, H. and Mahoney, C. (2015). Measuring interest group framing strategies in public policy debates. *Journal of Public Policy*, 35(02):223–244, doi: 10.1017/S0143814X14000294.

Loughran, T. I. and Mcdonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, doi: 10.1111/j.1540-6261.2010.01625.x.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., and Adam, S. (2018). Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3):93–118, doi: 10.1080/19312458.2018.1430754.

Matthes, J. (2009). What's in a frame? a content analysis of media framing studies in the world's leading communication journals, 1990-2005. *Journalism & Mass Communication Quarterly*, 86(2):349–367, doi: 10.1177/107769900908600206.

Matthes, J. (2013). Zum Gehalt der Framing-Forschung: Eine kritische Bestandsaufnahme. In Marcinkowski, F., editor, *Framing als politischer Prozess*, Schriftenreihe Politische Kommunikation und demokratische Öffentlichkeit, pages 17–28. Nomos, Baden-Baden.

Matthes, J. and Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication*, 58(2):258–279, doi: 10.1111/j.1460-2466.2008.00384.x.

McFarlane, T. and Hay, I. (2003). The battle for Seattle: Protest and popular geopolitics in the australian newspaper. *Political Geography*, 22(2):211–232, doi: 10.1016/S0962-6298(02)00090-2.

McLeod, D. M. and Hertog, J. K. (1992). The manufacture of 'public opinion' by reporters: Informal cues for public perceptions of protest groups. *Discourse & Society*, 3(3):259–275, doi: 10.1177/0957926592003003001.

McLeod, D. M. and Hertog, J. K. (1999). Social control, social change and the mass media's role in the regulation of protest groups. In Demers, D. and Viswanath, K., editors, *Mass media, social control, and social change*, pages 305–330. Iowa State Univ. Pr, Ames.

Miller, M. M. (1997). Frame mapping and analysis of news coverage of contentious issues. *Social Science Computer Review*, 15(4):367–378, doi: 10.1177/089443939701500403.

Mueller, C. (1997). International press coverage of east german protest events, 1989. *American Sociological Review*, 62(5):820–832, doi: 10.2307/2657362.

Pan, Z. and Kosicki, G. (1993). Framing analysis: An approach to news discourse. *Political Communication*, 10(1):55–75, doi: 10.1080/10584609.1993.9962963.

R Core Team (2019). R: A language and environment for statistical computing. https://www.R-project.org/.

Reese, S. D. (2010). *Framing public life: Perspectives on media and our understanding of the social world*. LEA's communication series. Routledge, New York, repr edition.

Roggeband, C. and Vliegenthart, R. (2007). Divergent framing: The public debate on migration in the Dutch parliament and media, 1995–2004. *West European Politics*, 30(3):524–548, doi: 10.1080/01402380701276352, https://doi.org/10.1080/01402380701276352.

Ruigrok, N. and van Atteveldt, W. (2007). Global angling with a local angle: How U.S., British, and Dutch newspapers frame global and local terrorist attacks. *Harvard International Journal of Press/Politics*, 12(1):68–90, doi: 10.1177/1081180X06297436, https://doi.org/10.1177/1081180X06297436.

Semetko, H. and Valkenburg, P. (2000). Framing european politics: a content analysis of press and television news. *Journal of Communication*, 50(2):93–109, doi: 10.1111/j.1460-2466.2000.tb02843.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.2000.tb02843.x.

Shah, D. V., Watts, M. D., Domke, D., and Fan, D. P. (2002). News framing and cueing of issue regimes. *Public Opinion Quarterly*, 66(3):339–370, doi: 10.1086/341396.

Thesaurus.com (2018). Synonyms of protest. http://www.thesaurus.com/browse/protest, (last accessed 20.05.2018).

Tilly, C. (2012). *Contentious performances*. Cambridge studies in contentious politics. Cambridge University Press, Cambridge and New York, second edition edition.

Tversky, A. and Kahneman, D. (1987). Rational choice and the framing of decisions. In *Rational choice: The contrast between economics and psychology*, Rational choice: The contrast between economics and psychology, pages 67–94. University Of Chicago Press, Chicago, IL, US.

van Atteveldt, W., Sheafer, T., and Shenhav, S. (2013). Automatically extracting frames from media content using syntacting analysis. In Davis, H., Halpin, H., Pentland, A., Bernstein, M., and Adamic, L., editors, *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*, pages 423–430, New York, New York, USA. ACM Press.

van Gorp, B. (2005). Where is the frame?: Victims and intruders in the belgian press coverage of the asylum issue. *European Journal of Communication*, 20(4):484–507, doi: 10.1177/0267323105058253.

van Gorp, B. (2007). The constructionist approach to framing: Bringing culture back in. *Journal of Communication*, 57(1):60–78, doi: 10.1111/j.0021-9916.2007.00329.x.

Wikipedia (2018). Protest. https://en.wikipedia.org/wiki/Protest, (last accessed 20.05.2018).