Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 4.2 (2022) 528-570
*HTTPS://DOI.ORG/10.5117/CCR2022.2.006.KROO*

# Beyond Counting Words

*Assessing Performance of Dictionaries, Supervised Machine Learning, and Embeddings in Topic and Frame Classification*

Anne C. Kroon
*Amsterdam School of Communication Research* (*ASCoR*)*, University of Amsterdam*

Toni van der Meer
*Amsterdam School of Communication Research* (*ASCoR*)*, University of Amsterdam*

Rens Vliegenthart
*Wageningen University*

**Abstract**

Topics and frames are at the heart of various theories in communication science and other social sciences, making their measurement of key interest to many scholars. The current study compares and contrasts two main deductive computational approaches to measure policy topics and frames: Dictionary (lexicon) based identification, and supervised machine learning. Additionally, we introduce domain-specific word embeddings to these classification tasks. Drawing on a manually coded dataset of Dutch news articles and parliamentary questions, our results indicate that supervised machine learning outperforms dictionary-based classification for both tasks. Furthermore, results show that word embeddings may boost performance at relatively low cost by introducing relevant and domain-specific semantic information to the classification model.

**Keywords:** automated text analysis, dictionaries, supervised machine learning, word embeddings, frames, policy topics

The application of techniques that find their origin in computer science helps scholars in the social sciences to better handle the sheer volume of "big data" and trace its unique characteristics. Consequently, and over the past years, scholars are increasingly adopting advanced automated text analysis techniques to capture and quantify predefined concepts at the heart of various communication theories, notably the identification of policy topics and frames, which are considered key content features of communicative texts and central in many media effects studies (e.g., Albaugh et al., 2014; Ruigrok & Atteveldt, 2007). In particular, *dictionary-based analysis* and *supervised machine learning* have become in vogue among scholars that have a clear preconception of their concepts of interest: Where the first method assigns class membership based on matches between articles and a predefined lexicon, the latter typically learns to identify patterns in texts by training an algorithm on manually annotated data.

Although these automated techniques have apparent advantages when it comes to minimizing costs and time while maximizing breadth, it is not always clear how researchers inform their choice for a specific technique. With the increasing popularity of cutting-edge computational methods in the field of communication science, it becomes more and more important to fully appreciate the impact of methodological choices on subsequent findings, conclusions, and theorizing. Especially as computational analysis is thought to bring "hard evidence" to the table (see for example Boumans & Trilling, 2016), it is crucial to gauge their ability to do just that.

Consequently, an increasing body of research is devoted to the comparison or implementation of lexicon and supervised techniques for text classification (Barberá et al., 2021; Chan et al., 2021; Hailong et al., 2014; Tulkens et al., 2016). Such comparisons typically focus on the classification of sentiment using English-language textual data (but see Al-Azani & El-Alfy, 2017; Stoll et al., 2020). The current study makes three unique contributions to this literature. First, we will compare the efficacy of dictionary approaches and supervised machine learning to measure key concepts of communication research *beyond* the context of English-language textual data. This is important as the availability of large, labelled training datasets, pre-trained classification tools and validated dictionaries is less self-evident when working with non-English textual corpora. At the same time, this represents a scenario that many communication researchers find themselves in.

Second, we specifically consider the benefits and drawbacks of competing classification techniques in terms of implementation, validation, and costs. Dictionary-based and supervised machine learning techniques have been widely applied in recent years, but not often conjointly, nor have their applications been subject to much comparative evaluation in the context of non-English languages and the classification of policy topics and frames.

Third, we investigate the usefulness of introducing word embeddings to supervised classification algorithms. Word embeddings, a state-of-the-art technique from the field of natural language processing and computer science, have transformed the ability of computers to understand the semantic and syntactic meaning of language (Le & Mikolov, 2014; Mikolov et al., 2013). Embedding models represent words in a high dimensional vector space, such that similar words occupy similar positions. When applied to a classification task, these models can help understand words in the application text even if they do not appear in the training data set. As the popularity of these methods increases, it becomes more and more important for researchers to understand the specific benefits and drawbacks associated with these techniques. Nonetheless, recognition of the potential of embeddings to improve classification tasks in the realm of communication research has remained limited (except for Rudkowsky et al., 2018).

Altogether, this contribution aims to provide guidelines for when communication researchers should opt for dictionary-based or supervised machine learning techniques, how they may implement these techniques, and with what consequences. We aim to move automated text analysis forward by demonstrating how communication scholars can avoid comprising quality and depth when automating and escalating the breadth of their research. We scrutinize the usefulness of computational techniques in two typical communication science tasks: the identification of *policy topics* (e.g., Albaugh, Sevenans, Soroka, & Loewen, 2013) and *frames* (Semetko & Valkenburg, 2000). We do so by utilizing a manually coded dataset derived from two key agendas: political (i.e., parliamentary questions) and newspaper coverage in the Dutch context. The final findings of this study will help scholars to make an informed decision regarding the selection of the appropriate method to describe, analyze, and understand communication phenomena.

## Automated Text Analysis in Communication Science

In an area of evolving online politics and digital media, studying the dynamics and framing of issue agendas has become increasingly challenging

and complex (Guo & Vargo, 2015). Daily interactions between politicians, policymakers, journalists and (interest) organizations have moved to online and openly accessible spaces. At the same time, traditional media outlets remain important in today's media landscape, shaping interactions in both online and offline settings (Djerf-Pierre & Shehata, 2017; King, Schneer, et al., 2017). The sheer amount of digital and traditional sources of data available for analysis has therefore increased substantially (Boumans & Trilling, 2016; Grimmer & Stewart, 2013).

Communication scholars are increasingly borrowing from computer linguists' toolkit to process and analyze this wealth of information. Generally speaking, established computational methods can be classified along a continuum of deductive, or so-called 'top-down' approaches, and inductive, 'bottom-up' approaches (Boumans & Trilling, 2016; Günther & Quandt, 2016). Using predefined categories, word list, and rules, deductive approaches aid researchers that already have a clear sense of key data characteristics. Most notably in this regard is the use of supervised machine learning to study overtime agenda convergence between predefined issues in diverse contexts, such as social and traditional media (Vargo et al., 2014). On the other end of the spectrum, inductive approaches allow the computer to extract meaning from specific datasets. Examples are inductive issue and frame analysis, using techniques such as Latent Dirichlet Allocation (LDA) (see Grimmer & Stewart, 2013), and assessment of document similarity (such as Cosine or Levenshtein distance) (e.g., Boumans, 2017) to trace agenda-setting effects across diverse domains.

The current contribution scrutinizes two regularly applied deductive automated approaches: dictionary-based text analysis and supervised machine learning. Communication scholars adopting these methods tend to have similar aims: quantifying established concepts with inherent meaning to the field of communication science, including particularly policy topics and frames. We borrow from the Comparative Agendas Project (CAP) community where computational methods are applied to reduce countless hours of policy issue coding, especially when conducting large-scale cross-country and over-time analyses (Albaugh et al., 2014, 2013). Additionally, following previous serious attempts (Burscher, Odijk, Vliegenthart, de Rijke, & de Vreese, 2014), we explore the automated measurement of frequently-coded *generic news frames* (Semetko & Valkenburg, 2000). As analyses based on dictionaries and supervised machine learning are geared towards measuring equivalent concepts, results yielded by these methods should—from a *theoretical* perspective—be largely comparable among each other as well as with human evaluations. The extent to which we can expect this to be the case will be discussed next.

## Dictionary-Based Text Analysis

Dictionaries describe the categories of interest employing lists of indicator words, after which these words are sought within the documents of interest (Günther & Quandt, 2016). Dictionary-based analysis exists by virtue of the assumption that individual words carry meaning *beyond* context (Taboada et al., 2011) —at least to some extent—making their application promising across several domains. Its adoption is particularly prevalent among studies aiming to classify sentiment, where words such as "happy" or "good" carry a positive connotation *regardless* of a specific context.

Nonetheless, the meaning and valence of many words *does* differ across contexts, as is, for example, the case for words as "legit" or "bad". Consequently, dictionaries developed to identify and categorize social media data—characterized by relatively high levels of subjectivity (e.g., Welbers & Opgenhaffen, 2019)—might not straightforwardly generalize to financial (Loughran & McDonald, 2011), editorial or political texts. Ultimately, this poses a challenge to scholars aiming to apply a dictionary in a different domain than the one it was developed for. Altogether, the performance of dictionaries structurally varies (Boukes et al., 2019) and largely depends on the fit between domain and genre of the application data and the data used to generate the dictionaries (Loughran & McDonald, 2011; Ribeiro et al., 2015).

One may decide to create *tailor-made* dictionaries to meet particular research purposes (e.g., Damstra & Boukes, 2018; Neuman, Guggenheim, Jang, & Bae, 2014), by creating wordlists that aim to represent theoretical concepts from scratch. Yet, scholars also have a broad set of *off-the-shelf dictionaries* at their direct disposal (for overviews; see Boukes et al., 2019; Ribeiro et al., 2015), typically tested and validated by human coders within specific content domains (such as social media: Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010). As off-the-shelf dictionaries do not straightforwardly generalize, validating and tailoring off-the-shelf dictionaries to the domain or genre of interest (e.g., Kroon & van der Meer, 2018; Vargo et al., 2014), or hand-crafting a dictionary to meet the specific research aim, might be a necessary requirement (Boukes et al., 2019).

Dictionary-based analysis has the advantage of being financially and computational cheap as it does not require supervision, explaining its widely accepted practice. Yet, dictionary-based approaches have been criticized for several reasons. Dictionaries are generally not very well equipped to deal with multi-word phrases, lexical patterns, and semantical context. In its most basic form, dictionary approaches assign a binary weight to each term—which may not be sufficient to capture the complexity of specific categories. Most pertinently, dictionaries are *manually* developed word lists

that are hoped to identify the underlying constructs they represent (Guo et al., 2016). It is, however, challenging for humans—if not impossible—to arrive at a dictionary capturing all relevant words or word-combinations to identify relevant categories, meeting the criteria of *inclusiveness* (i.e., avoiding false negatives) while remaining *discriminating* enough (i.e., avoiding false positives). The manual construction of dictionaries is not neutral but likely shaped by the domain knowledge and personal conceptions of the researcher (Burscher et al., 2014), causing it to be highly unreliable: Two human experts will most likely arrive at keyword lists that overlap only marginally (King, Lam, et al., 2017). Especially with thousands or millions of articles to analyze it is hard to arrive at a representative dictionary, making it "very likely that the predetermined list of categories will narrow or bias the potential areas to be analyzed" (Guo et al., 2016), ultimately compromising semantic validity.

Although the performance of several off-the-shelf dictionaries has been compared among each other (Boukes et al., 2019; Ribeiro et al., 2015), less is known about how they compare to the performance of another popular deductive tool to automate human coding: Supervised machine learning. We will discuss this next.

**Supervised Machine Learning**

At heart, the algorithms underlying supervised learning supersede the human element in dictionary construction by teaching the computer how to construct a dictionary themselves (Günther & Quandt, 2016). In contrast to dictionary-based approaches, supervised machine learning requires a more "expensive" and time-consuming process, as a manually coded dataset is needed. Typically, in a first step, a codebook will be constructed after which data will be manually coded. Then, in a second step, these manual annotations are used as training data for the algorithms. During training, the supervised classifiers "learn" to decipher rules about the relation between textual features and classes that underlie human decisions. These rules will then be used to predict the class membership of unseen documents (such as social media posts, press releases or news articles).

Researchers interested in applying supervised machine learning techniques have a broad number of different classifiers to choose from. As part of the family of Bayesian algorithms, *Naïve Bayes Classifier* is a simple, probabilistic algorithm that tries to construct rules based on (co)-occurrences of features to predict class membership. Owing to the naïve assumption that features independently contribute to class probability, this algorithm is relatively fast and computational affordable. Regardless, Naïve

Bayes classifiers belong to the group of most effective machine learning classifiers (Kübler et al., 2017). *Support-Vector Machines* (SVM) aim to identify a hyperplane in a *n* – dimensional space (whereby *n* represents the number of features) that distinctly categorizes the data points. As a large-margin (rather than probability) classifier, SVM often outperforms Naïve Bayes. The *Passive Aggressive classifier* (PA) is an online learning algorithm that resembles the SVM algorithm. PA classifiers use the margin to improve the classification. The *passive* part of the classifier keeps the model in case a correct classification is made, while the *aggressive* part updates the model weights in the case of incorrect classification. In this way, the algorithm can classify texts in a highly effective manner (as demonstrated in a similar classification task as the one presented in this study: Burscher et al., 2015). *Stochastic Gradient Descent* (*SGD*) is an optimization technique aiming at finding a local optimum given a starting point[1]. The ability of this algorithm to effectively classify media content has been confirmed (Budak et al., 2016). Finally, decision trees, such as *Extra Trees* (*ET*) are among the most popular classifier algorithms. Consisting of a large number of decision trees, ET is an ensemble learning method that randomizes decisions and data subsets to avoid overfitting. It returns the class that received most votes.

The popularity of these and other supervised methods to automate—parts of—the analysis of dynamics in social, news, and political, and media content is increasing rapidly in the field of political communication and the social sciences more generally (Grimmer & Stewart, 2013). More in particular, these techniques have been used to identify issue publics (Yuan et al., 2019), measure elements of deliberative quality in the public sphere (e.g., Colleoni, Rozza, & Arvidsson, 2014; Su et al., 2018), and quantify policy topics (or, *issues*) (Albaugh et al., 2014; Burscher et al., 2015).

Going even a step further, a set of studies have explored the ability of computer-assisted methods to identify *frames*. Due to its "abstract" (Matthes & Kohring, 2008, p. 258) and "elusive" (Maher, 2001, p. 83) nature, frames are generally considered difficult to identify and quantify, especially by computers that are generally seen as unfit to understand subtle nuances and complexities in language. Regardless—as demonstrated by a sequence of experiments using Dutch data—generic news frames (based on the codebook of Semetko & Valkenburg, 2000) can be effectively and accurately coded using supervised classifiers (Burscher et al., 2014; see also: Opperhuizen, Schouten, & Klijn, 2019).

Based on the above-mentioned discussion, we expect supervised classifiers to outperform dictionary-based text analysis. As the subjective human influence in dictionary construction is made redundant in supervised

machine learning, the risk of researcher bias (in terms of limited domain knowledge and concept subjectivity) will be reduced. On the other hand, obtaining a good golden standard for training is not always self-evident. We scrutinize two research problems with different levels of difficulty: The identification of policy topics (considered a relatively manifest and easy to classify concept) and frames (considered a relatively abstract and hard to classify concept). We expect that supervised machine learning algorithms perform better in the classification of policy topics and frames when compared to dictionary-based analysis.

**Introducing Word Embeddings to Text Classification**

Despite that human coding is far from flawless, it is typically considered superior when it comes to the appreciation of semantic and syntactic word meaning. Computers have typically been criticized as being "unable to understand human language in all its richness, complexity, and subtlety as can a human coder" (Simon, 2001, p. 87). Yet, advances in the field of Artificial Intelligence (AI) and computational linguistics have made it questionable to what extent this criticism is still valid. Particularly, the introduction of word embeddings has radically transformed computers' ability to interpret and understand human language (Le & Mikolov, 2014; Mikolov et al., 2013).

The basic idea of word embeddings is captured in the famous quote of linguist Firth (1935, p. 37): "[t]he complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously." Inspired by Firth's ideas, embedding models acknowledge that word meaning differs across context. By mapping vocabulary words to its context across its many occurrences in the training dataset, embedding models learn the meaning of language.

Embedding models are typically trained on large volumes of text, often derived from online sources such as Wikipedia or news corpora. These models learn distributed vector representations of words on continuous-bag-of-words or skip-gram algorithms. During training, embedding models learn to represent the semantic and syntactic meaning of language in an unsupervised manner. After training, each word is mapped to a multidimensional vector that captures information about word relationships. Similar words will receive similar vector representations while unrelated words will be far apart in the *n*-dimensional vector space. For example, the vector of *France* will be close to the vector of *Austria* and *Belgium,* while the vector of *Jesus* will be close to that of *God* (Collobert et al., 2011). Ultimately, word vectors should preserve relevant information about a text, which might allow for better classification results than one-hot word encodings.

Using word embeddings in a text classification task may result in better performing and generalizable models. In traditional supervised methods, words that are not present in the training dataset (i.e.., "unseen" or "out-of-context" words) are treated as plain nuisance. Word embedding models allow the detection and classification of such unseen words based on their relationship with other words if these words were included in the training-set of the embedding model (Rudkowsky, Haselmayer, Wastian, Jenny, Emrich, Sedlmair, et al., 2018). This is particularly useful when training data are scarce—and thus the number of unseen words is large—which is often the case as supervised classifiers need thousands of annotated examples to perform well. This logic applies when the embedding are pre-trained or learned separately from the supervised classifier.

Embeddings can also be part of advanced supervised algorithms. More specifically, embeddings are often used as a first layer in deep-learning models. In the current paper, we do not consider the configuration and performance of such deep-learning models as traditional classification algorithms might be more appropriate when the labelled dataset is limited (e.g., Tamara & Milićević, 2018). Using word embeddings as a document representation method may help communication scholars to add contextual information to their model, also under the conditions of limited training data and without knowledge and skills needed to configure deep learning models.

Several off-the-shelf word embeddings are freely available. Such models are typically trained on generic web data, such as Wikipedia articles. Although off-the-shelf models may be useful for several Natural Language Processing (NLP) tasks, we expect the best results from an embedding model that is trained on *relevant* and *domain-specific* media content. Particularly, the broader population of social media posts, news articles or parliamentary questions from which the labeled dataset is derived represents a vital source of semantic and syntactic information crucial to the research question under investigation. By training an embedding model on the broader population of media content, relevant and domain-specific information about the semantic and syntactic meaning of words in the corpus of interest is retrieved. We, therefore, expect that domain-specific word embeddings will boost the performance of supervised classification of policy topics and frames.

## Method

The current study compares dictionary-based and machine learning classification approaches to the gold standard of human coding across the measurement of policy topics and frames. More specifically, we compare

and contrast existing dictionaries for the classification of policy topics with custom-made dictionaries for the classification of frames. In addition, we test a set of machine learning classification algorithms for both policy topics and frames. Finally, we investigate whether the inclusion of semantic vectors (i.e., word embeddings) can aid these classification problems. The complete code for this project was written in Python. Both the data and code are openly accessible at: https://github.com/annekroon/dictionaries-vs-sml.

The current study relies on two main datasets. The first dataset is used to train and test the classifiers. The second dataset is used to train word embedding models that will be used to transform the training data.

**Training Dataset**

The current study relies on data from the political and news media agenda within the period 1995 till 2017. First, and regarding the political agenda, a stratified random sample was taken from written parliamentary questions in the Dutch parliament to obtain an equal number of documents for each year ($N$=1,694). These written questions provide the most comprehensive presentation of the parliament's agenda and, unlike questions asked during the weekly question hour, are not biased by selection processes by for example the chair of parliament (Van Aelst & Vliegenthart, 2014). Selected items were downloaded from the Dutch government's official website. Subsequently, the HTML pages were parsed so to extract relevant content in tabular form (e.g., date, relevant text).

Second, and regarding the news media agenda, an extensive search string identifying political news[2] was used to select and download political news from *LexisNexis*. We restricted our search to two prominent Dutch newspapers: One of the main-left-wing newspapers *de Volkskrant* ($n$= 667) and the most popular right-wing tabloid-like *Telegraaf* ($n$= 446), totalling to a stratified random sample of $N$=1,113 news articles.

The sample size ($N$=2807) is relatively small for a challenging machine learning problem consisting of multiple classes and labels. Although clear rules for the optimal training size do not exist (but rather, depend on for example task type and input features), generally the quality and amount of data determines quality classification. At the same time, we believe it represents a realistic sample size to acquire for communication scholars, who often have limited time and financial resources. Furthermore, we believe this dataset represents a more conservative test of the difference between dictionary and machine learning approaches, as dictionary approaches might prove more useful in the condition of restricted training data.

**Table 1. Sources of News Content Used to Train Embedding Model**

| Source | *N* |
|---|---|
| ANP (print) | 1718459 |
| NOS (www) | 82221 |
| Telegraaf (www) | 348803 |
| Tubantia (www) | 66807 |
| BN DeStem (www) | 75923 |
| ED (www) | 77065 |
| Gelderlander (www) | 52781 |
| bd (www) | 88779 |
| Trouw (www) | 52133 |
| Zwarte Waterkrant (www) | 1794 |
| NU.nl (www) | 168057 |
| Metro (print) | 169460 |
| NRC (print) | 719626 |
| PZC (www) | 64507 |
| Trouw (print) | 623446 |
| De stentor (www) | 70686 |
| Telegraaf (print) | 895478 |
| AD (www) | 158132 |
| Spits (www) | 41481 |
| Metro (www) | 104291 |
| AD (print) | 861902 |
| Volkskrant (print) | 726556 |
| Volkskrant (www) | 137007 |
| Parool (www) | 46751 |
| Friesch Dagblad (www) | 797 |
| FD (print) | 452968 |
| NRC (www) | 88546 |

We preprocessed the news articles and parliamentary questions in the following manner: After tokenization and lower-casing all words, we removed single-letter words, punctuation, and Dutch stop words[3].

## Word Embedding Dataset

For the current study, a domain-specific embedding model was trained on a diverse, large and representative population of news articles and parliamentary questions from which our training sample was drawn. More specifically, we train an embedding model on a corpus of news articles ($n$=7,894,456) (2000-2018) and the total population of parliamentary questions (1995—2017) ($n$=57,892), totaling to $N$=7.9M documents. The news articles are derived from a diverse set of online and print news sources in the Netherlands,

published between 2000 and 2018. Table 1 displays an overview of the outlets that were included. The parliamentary questions were downloaded from the official website of the Dutch parliament. As a consequence, our model will learn word meaning and semantic relationships between words in the corpus of interest; resulting in word vectors that capture the dominant or general meaning of words in the Dutch news and political domains, which should be useful for subsequent classification (see Rudkowsky et al., 2018).

The news articles and parliamentary questions were split into sentences. All sentences were lowercased and punctuation was removed. We rely on the word2vec algorithm from the Python library Gensim to train a baseline model using the continuous-bag-of-words architecture (dimensions=300, window size=10, negative sampling=15). We expected that higher levels of the size of vector (which may express the complexity of discourse) and window size (which may account for semantics or grammar) may affect the performance of our models. We, therefore, trained three additional models with alternative configurations (dimensions varying between 100-300 and windows sizes varying between 10-15).

**Manual Coding and Training**

A team of six coders was trained to identify topics and frames in the selected parliamentary questions and news articles. For the coding of the Policy Topics, we follow the general instructions as provided by the Comparative Agendas Project (CAP). For the coding of the news frames, we use the coding instructions provided by Semetko and Valkenburg, (2000). Coders received instructions and asked to code several news articles and parliamentary questions as part of their coder training. All coders received extensive feedback on their performance. Following several rounds of instructions and feedback, satisfactory levels of intercoder reliability were obtained—afterwards the final coding could start. The sample used to calculate the intercoder reliably consisted of the final training dataset ($n$=19) and random sample drawn during the coding process ($n$=376 for policy topics, $n$=11 for news frames). Intercoder reliability of policy topics was satisfactory (Krippendorff's $\alpha = 0.75$). Agreement among coders for the frames ranged from relatively low to acceptable: Attribution of responsibility (Krippendorff's $\alpha = 0.34$), conflict (Krippendorff's $\alpha = 0.52$), human interest (Krippendorff's $\alpha = 0.49$), and economic consequences (Krippendorff's $\alpha = 0.68$). Comparable levels of agreement have been reported by previous studies analyzing media data with multiple categories and coders (Burscher et al., 2014; Van der Pas, 2013). For the final sample, each unit was coded by one of the six coders.

## Classification of Policy Topics

In our attempt to classify policy topics, we are faced with a *multiclass classification* problem. In line with the general approach of the Comparative Agendas Project to allow for a single topic per document coders are asked to identify the most salient topic per news article / parliamentary question. This means that each news article or parliamentary question in our dataset can be classified as a single policy issue; the categories are mutually exclusive. For the classification of policy topics, we rely on the entire dataset of news articles and parliamentary questions ($N$=2807).

### Manual coding of Policy topics

For the manual coding of policy topics, we used the Dutch version of the Belgian Policy Agenda codebook (also used in, for example, Vliegenthart et al., 2016). The codebook consisted of a total of 28 main topics that could be coded for. For analysis, however, we decided to focus only on policy topics representing a substantial share of the final dataset. Niche topics that occurred less than 80 times in the dataset were assigned to the residue category 'other issue'. This leaves us with the following 14 categories: *Banking, finance, & commerce; Civil right; Defense; Education; Environment; Governmental operations; Health; Immigration & integration; Int. affairs & foreign aid; Labor & employment; Law & crime; Social welfare; Transportation;* and the residual category *Other issue* (combining articles that were assigned to niche topics occurring in low shares of the sample).

### Dictionary Approach to Classify Policy Topics

We use the Dutch Policy Agenda Topic Dictionary (Albaugh et al., 2013): A validated lexicon used to measure policy topics in Dutch-language media content and party manifestos from Belgium. The original study used the software program *Lexicoder* to map the dictionaries to the text. Using Python and adding several pre-processing steps to the analysis, such as removing punctuation and lower casing all words, our approach deviates slightly from the original approach—but resembles that of other validation studies (Albaugh et al., 2014).

Each newspaper article or parliamentary question was classified as a topic when at least two words from a topic category occurred in the text. Although this is a fairly arbitrary number, this threshold worked well in previous research using the same dictionary (Albaugh et al., 2014). In addition to mapping the original version of the dictionary, we also applied a stemmed version. To this aim, we stemmed both the dictionary and the text corpus

using the Dutch Snowball Stemmer from the nltk package in Python. In the stemming process, words are reduced to their root. We expect that adding stemming will aid the dictionary in identifying relevant words that signify policy topics. Again, we set the threshold for identifying a topic at two dictionary words. Smaller categories were first identified on their own and then merged into the residual category *Other issue* in order to make sure that topics matched by the dictionary mapped those identified by the supervised classifier.

Both the original and stemmed versions of the dictionary did not classify the documents in a mutually exclusive manner, meaning that some news articles and parliamentary questions were classified into multiple topics. As this conflicts with human coding and supervised learning classification, we selected the prominent topic per document using the following two approaches. First, we use an *index-based* approach; for each hit of a dictionary word the position of the match in the respective document (i.e., index location) was returned, so that the first word in a document receives the number one and ascends to the last word. Following the logic that more salient topics will most likely be discussed upfront of a newspaper article or parliamentary question, we argue that dictionary words occurring at the start of a document carry more weight compared to words that occur at the end. In case of conflict between multiple topics, we selected the topic associated with the lowest index number. Second, and using a *count-based* approach, we selected topics with the highest count of dictionary terms per document.

### Machine Learning Classification of Policy Topics

**Bag of Words (BoW) vectorizers.** Under the hood, machine learning algorithms operate on *vectors* (arrays of numbers) rather than textual data. Before applying supervised machine learning models, one should, therefore, transform textual data to numerical representations (i.e., vectorize the textual data). We convert text to numerical matrices using *count* and *tfidf vectorizers*: After tokenizing a collection of documents, count vectorizers create a matrix of token counts, using the number of times a vocabulary word occurs in a document as its weight**.** Term Frequency-Inverse Document Frequency (tf-idf) vectorizers help to reflect the importance or uniqueness of a word to a document. Weights assigned to tokens are calculated based on both recurrences of the term in the entire corpus in addition to counts within particular documents. Tf-idf values increase proportionally to the frequency of a word in a document while being offset by the frequency of word occurrence in all documents (Bilbro et al., 2018).

**Embedding Vectorizers.** Word vectors derived from the baseline domain-specific word embedding model are used to vectorize the training data. For each word in a document (i.e., news article or parliamentary question) the associated word vectors in the embedding model are retrieved to build input features readable for machine learning algorithms. In the next step, one may simply average the retrieved word vectors for all words in the document. The inclusion of weighting, however, based on average word frequency or tf-idf can boost performance (Corrêa et al., 2017; Ferrero et al., 2017). The current study uses both the *count* and *tf-idf* weighted word vectors of the words in a document.

**Classifiers.** In this study, we included the following commonly used classifiers: *Support-Vector Machines* (SVM)*, Passive Aggressive classifier* (PA), *Stochastic Gradient Descent* (*SGD*) and *Extra Trees* (*ET*). For hyperparameter tuning of the diverse classifiers, we relied on the grid search technique using 5-fold cross-validation.

## Classification of Frames

The classification of frames represents a *multilabel classification* problem: News articles can and may contain multiple frames, meaning that the categories are *not* mutually exclusive. Here, we rely on the newspaper dataset (*N*=1,113) only, as the focus of the analysis is on the classification generic *news* frames.

### Manual Coding of Frames

We focus on a set of generic frames that are well-established in the literature: *Attribution of responsibility, Conflict, Human Interest,* and *Economic Consequences* (Semetko & Valkenburg, 2000). Coders were allowed to code multiple frames per news article. It is quite common that multiple frames are present per news article: In 547 news articles, two or more frames were identified. The coding criterion was adopted from Semetko & Valkenburg (2000). To determine whether a frame is present, coders were asked to respond to a set of questions per frame. One positive response was sufficient for a frame to be present.

***Attribution of Responsibility.*** The presence of this frame indicates that the news article addresses some level of responsibility of a governmental body or representative for alleviating or causing an issue/ problem. Four items were used to measure the presence of attribution of responsibility (e.g., *Does the story suggest an individual responsible for the problem?*).

***Human Interest.*** The presence of this frame indicates that the news article gives a human face to the discussed issue or problem. Four items measured the presence of human-interest frame (e.g., *Does the story provide a "human face" or example on the issue?*).

***Conflict.*** The conflict frame indicates that story reflects some level of disagreement between parties, groups, or countries. Three items measured conflict (e.g., *Does the story refer to two sides or more sides of the problem or issue?*).

***Economic Consequences.*** This frame is indicated to be present when the news article mentions financial losses or gains. Three items measured the presence of economic consequences frame (e.g., *Does the story mention the costs/ degree of expenses, now or in the future?*).

### Dictionary Approach to Classify Frames

To measure the presence of frames, a custom-made dictionary was developed for the purpose of this study, created during the round of manual coding of news articles. Specifically, each time coders encountered a frame, they were asked to indicate which set of words in the news story signified the presence of the frame. They could submit between two and seven words per frame. This resulted in extensive word lists. We assess performance of the dictionaries at different training sizes. We selected the 30 most frequently mentioned words per frame. We opted for this number, as during several pilot tests we noticed that best results were yielded with dictionaries of this length (i.e., longer lists increased the number false positives). The complete tailor-made dictionaries measuring attribution of responsibility (example words: *minister, cabinet, municipality*), human interest (example words: *people, woman, children*), economic consequences (example words: *euro, million, billion*), and conflict (example words: *criticism, struggle, problems*) can be found in Appendix A. Again, we also created a stemmed version of the dictionary to apply to the stemmed version of the test dataset.

### Machine Learning Classification of Frames

**Vectorizers**. For the supervised machine learning approach, we largely rely on the same approach as was used for the classification of policy topics. The data was transformed using both the BoW vectorizers (i.e., *count* and *tf-idf*), as well as using the vectorizers based on four embedding models with different configurations (i.e., dimensions varying between 100 and 300, and window sizes of 10 and 15). These models were used to retrieve vectors for each word that are *mean, max,* and *sum* weighted by word frequencies

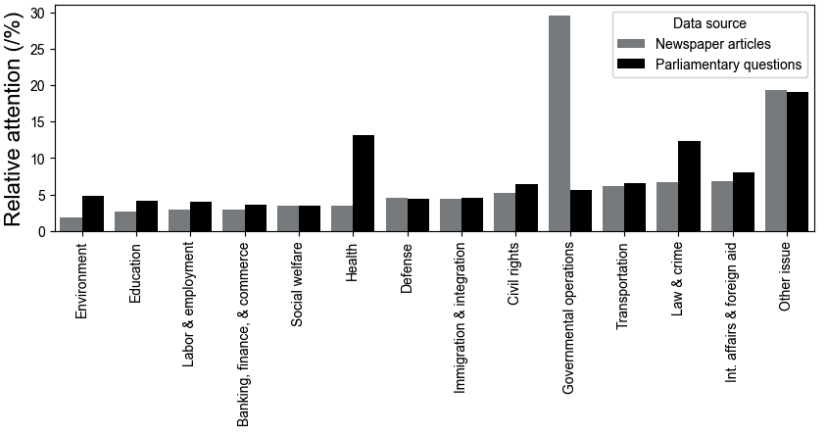



**Figure 1.** Relative Attention for Policy Topics in Newspapers and Parliamentary Questions

and tfidf (Giatsoglou et al., 2017; Rudkowsky, Haselmayer, Wastian, Jenny, Emrich, Sedlmair, et al., 2018).

**Classifiers.** Regarding the implementation of classification algorithms, a different approach is taken. To account for the multi-label structure of the data, one-vs-rest strategies are implemented. This approach fits each class (i.e., frame) against all other classes, essentially converting our multi-label issue to a binary classification problem. We use this approach in tandem with the same set of classification algorithms used to predict policy issue membership (SVM, PA, SGD, and ET).

## Analysis: Evaluating classification effectiveness

To evaluate the classification effectiveness of the different classifiers, we rely on the following performance metrics: precision, recall, and f1-score. These metrics are based on four prime parameters: *True positives* (TP): Correctly predicted positive values, *false positives* (FP): Incorrectly predicted positive classes, *true negatives* (TN): Correctly predicted negative values, and *false negatives* (FN): Incorrectly predicted positive classes.

Based on these four parameters, we calculate the evaluation metrics: ***Precision*** indicates how many of the identified instances are relevant (TP / TP+FP). ***Recall*** indicates the proportion of the true positives that were found or recalled by the model, informing us about how many relevant items were selected (TP / TP + FN). The ***f1-score*** represents the harmonic mean of the precision and recall values and is calculated as follows: 2*(Recall * Precision) / (Recall + Precision).

**Table 2. Attention for Policy topics across News and Political Agendas**

| | Newspaper articles | | Parliamentary questions | | Total | |
|---|---|---|---|---|---|---|
| | *N* | % | *N* | % | *N* | % |
| Banking, finance, & commerce | 33 | 3.0 | 62 | 3.7 | 95 | 3.4 |
| Civil rights | 58 | 5.2 | 108 | 6.4 | 166 | 5.9 |
| Defense | 51 | 4.6 | 74 | 4.4 | 125 | 4.5 |
| Education | 30 | 2.7 | 69 | 4.1 | 99 | 3.5 |
| Environment | 20 | 1.8 | 82 | 4.8 | 102 | 3.6 |
| Governmental operations | 329 | 29.6 | 94 | 5.5 | 423 | 15.1 |
| Health | 39 | 3.5 | 223 | 13.2 | 262 | 9.3 |
| Immigration & integration | 49 | 4.4 | 76 | 4.5 | 125 | 4.5 |
| Int. affairs & foreign aid | 76 | 6.8 | 136 | 8.0 | 212 | 7.6 |
| Labor & employment | 33 | 3.0 | 68 | 4.0 | 101 | 3.6 |
| Law & crime | 74 | 6.6 | 210 | 12.4 | 284 | 10.1 |
| Social welfare | 38 | 3.4 | 59 | 3.5 | 97 | 3.5 |
| Transportation | 68 | 6.1 | 111 | 6.6 | 179 | 6.4 |
| Other issue | 215 | 19.3 | 322 | 19.0 | 537 | 19.1 |
| Total | 1113 | 100.0 | 1694 | 100.0 | 2807 | 100.0 |

## Results

We will start with a presentation of our results pertaining to the multi-class classification of policy topics. Table 2 shows the absolute attention for the policy topics in news media and parliamentary questions, while Figure 1 displays the relative attention to the different policy topics. Relative attention for the topics is largely comparable across both domains, with that *governmental operations* are more frequently discussed in news media compared to parliament, while the issue *health* receives comparably little attention in the news environment.

It was expected that supervised machine learning algorithms perform better in the classification of (a) policy topics when compared to dictionary-based text analysis and that this classification would improve when vectorizing the data using a domain-specific embedding model. Table 3 presents an overview of the evaluation metrics assessing the identification of policy topics across classifiers, while Figure 2 visualizes the f1-scores. As can be seen the supervised classifiers outperform the stemmed and not stemmed count-based and index-based dictionary-based analysis. Overall, and as indicated by the f1-score, results show that SGD classifier using the count embedding vectorizer represents the best classifier. An inspection of Figure 2 informs us that the performance of the classifiers was boosted when BoW

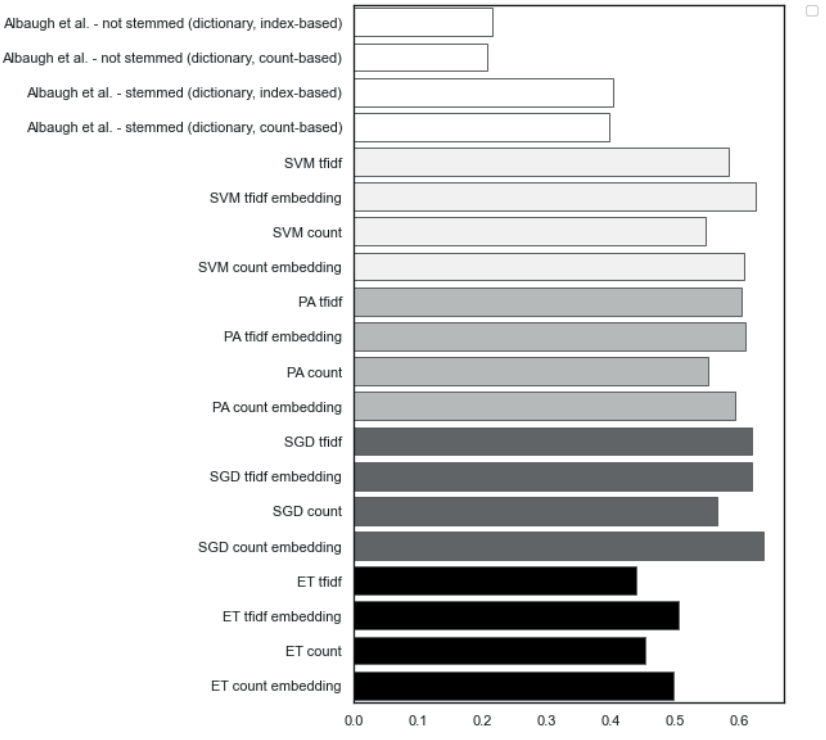



**Figure 2.** Effectiveness (F1-Scores) of Policy Issue Classification across Classifiers

vectorizers were replaced for vectorizers based on word vectors. Inspection of the difference in actual and predicted policy topics when opting for an analysis based on the here-used dictionary versus SGD classification indicates that the 'other category' was often classified by the dictionary, while the 'environment' topic was missed by the supervised algorithm. It should be noted that even the best classifier leaves much room for improvement.

We proceed to the multi-label classification challenge of frame identification. It was expected that supervised machine learning algorithms perform better in the classification of (b) frames when compared to dictionary-based analysis. Also, we expected that this classification would improve when vectorizing the data using the domain-specific embedding model. Table 4 summarizes the performance results across classifiers. Specifically, we have listed f1-score, precision and recall of the top three performing classifiers per type of vectorizer (i.e., BoW and based on the embeddings), as well as the dictionaries. Overall, our tailored-made dictionary performs poorer than the supervised machine learning

**Table 3. Performance of Policy Issue Classification across Classifiers**

| | precision | recall | Weighted f1-score |
|---|---|---|---|
| Albaugh et al. – not stemmed (dictionary, index-based) | 0.25 | 0.24 | 0.21 |
| Albaugh et al. – not stemmed (dictionary, count-based) | 0.26 | 0.23 | 0.21 |
| Albaugh et al. – stemmed (dictionary, index-based) | 0.44 | 0.42 | 0.40 |
| Albaugh et al. – stemmed (dictionary, count-based) | 0.45 | 0.41 | 0.40 |
| SVM tfidf | 0.62 | 0.57 | 0.58 |
| SVM tfidf embedding | 0.65 | **0.61** | 0.62 |
| SVM count | 0.57 | 0.54 | 0.55 |
| SVM count embedding | 0.63 | 0.60 | 0.61 |
| PA tfidf | 0.63 | 0.59 | 0.60 |
| PA tfidf embedding | 0.65 | 0.59 | 0.61 |
| PA count | 0.57 | 0.54 | 0.55 |
| PA count embedding | 0.65 | 0.58 | 0.59 |
| SGD tfidf | 0.66 | 0.60 | 0.62 |
| SGD tfidf embedding | 0.67 | 0.60 | 0.62 |
| SGD count | 0.59 | 0.56 | 0.57 |
| SGD count embedding | **0.69** | **0.61** | **0.64** |
| ET tfidf | 0.51 | 0.43 | 0.44 |
| ET tfidf embedding | 0.53 | 0.49 | 0.50 |
| ET count | 0.50 | 0.44 | 0.45 |
| ET count embedding | 0.53 | 0.48 | 0.50 |

*Note.* Largest value in each column is bolded.

algorithms, however: There is quite some variation in *how much* worse it performs. When inspecting the f1-score, we see that for the identification of the economic consequences frame differences not very large. Here, a pre-defined list of words seems to work well for identifying an economic consequences perspective in the news. Words such as *financial* and *euro* seem to capture the economic consequences frame quite well. However, for the identification of the other frames (attribution of responsibility, conflict, and human-interest frame) supervised algorithms outperformed the dictionary-based approach by far (.26, .1, and .32 points difference in the f1-score respectively). For these frames, it proved relatively hard to come up with words that indicate attribution of responsibility, conflict or a human-interest perspective *beyond* the specific context for which they were developed.

**Table 4. Performance of Frame Classification across Classifiers: The Top 3 Best Performing Classifiers for the Embedding and BoW vectorizers and the Dictionaries are Listed.**

| Precision | Recall | F1-score | Classifier | Vectorizer |
|---|---|---|---|---|
| | | | **Attribution of responsibility** | |
| 0.69 | 0.68 | 0.67 | SVM tfidf embedding sum, d=300, s=15 | embedding vectorizer |
| 0.69 | 0.68 | 0.67 | SVM tfidf embedding sum, d=300, s=10 | embedding vectorizer |
| 0.67 | 0.66 | 0.65 | SVM tfidf embedding sum, d=100, s=10 | embedding vectorizer |
| 0.58 | 0.59 | 0.58 | SGD tfidf | baseline vectorizer |
| 0.55 | 0.56 | 0.55 | PA count | baseline vectorizer |
| 0.55 | 0.55 | 0.55 | SGD count | baseline vectorizer |
| 0.56 | 0.48 | 0.41 | Dictionary – stemmed | Dictionary – stemmed |
| 0.64 | 0.49 | 0.39 | Dictionary – not stemmed | Dictionary – not stemmed |
| | | | **Conflict** | |
| 0.63 | 0.63 | 0.62 | SGD tfidf embedding mean, d=300, s=10 | embedding vectorizer |
| 0.62 | 0.62 | 0.62 | PA count embedding mean, d=300, s=10 | embedding vectorizer |
| 0.61 | 0.61 | 0.61 | PA tfidf embedding mean, d=300, s=10 | embedding vectorizer |
| 0.60 | 0.60 | 0.60 | ET count | baseline vectorizer |
| 0.57 | 0.57 | 0.56 | SGD tfidf | baseline vectorizer |
| 0.56 | 0.56 | 0.56 | SGD count | baseline vectorizer |
| 0.58 | 0.57 | 0.55 | Dictionary – not stemmed | Dictionary – not stemmed |
| 0.56 | 0.55 | 0.52 | Dictionary – stemmed | Dictionary – stemmed |
| | | | Economic Consequences | |
| 0.81 | 0.81 | 0.80 | SVM tfidf embedding max, d=300, s=15 | embedding vectorizer |
| 0.81 | 0.81 | 0.80 | SVM count embedding max, d=300, s=15 | embedding vectorizer |
| 0.81 | 0.81 | 0.80 | PA count | baseline vectorizer |
| 0.78 | 0.78 | 0.78 | SGD tfidf embedding max, d=300, s=15 | embedding vectorizer |
| 0.79 | 0.79 | 0.78 | PA tfidf | baseline vectorizer |
| 0.79 | 0.79 | 0.78 | SGD tfidf | baseline vectorizer |
| 0.78 | 0.74 | 0.74 | Dictionary – not stemmed | Dictionary – not stemmed |
| 0.77 | 0.70 | 0.71 | Dictionary – stemmed | Dictionary – stemmed |

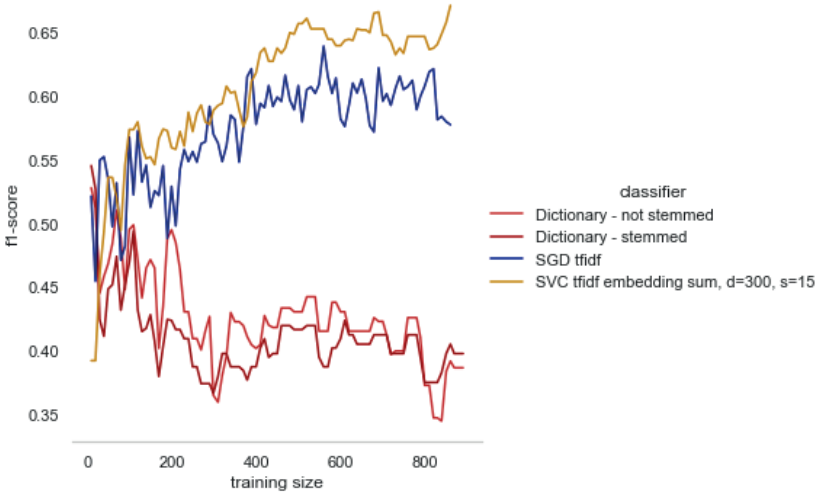
| | | | Human Interest | |
|---|---|---|---|---|
| 0.78 | 0.80 | 0.77 | ET count embedding mean, d=300, s=15 | embedding vectorizer |
| 0.77 | 0.79 | 0.77 | SGD tfidf | baseline vectorizer |
| 0.77 | 0.79 | 0.77 | SGD tfidf embedding sum, d=300, s=10 | embedding vectorizer |
| 0.79 | 0.81 | 0.77 | ET count embedding mean, d=300, s=10 | embedding vectorizer |
| 0.76 | 0.76 | 0.76 | SGD count | baseline vectorizer |
| 0.76 | 0.78 | 0.76 | PA tfidf | baseline vectorizer |
| 0.70 | 0.43 | 0.45 | Dictionary – not stemmed | Dictionary – not stemmed |
| 0.77 | 0.41 | 0.41 | Dictionary – stemmed | Dictionary – stemmed |

*Note.* D = dimensionality, S = window size.

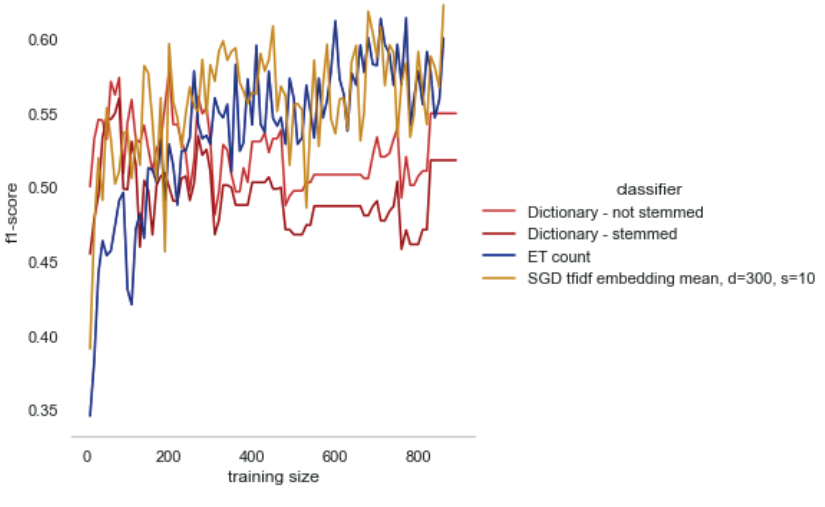
Furthermore, Table 4 indicates that for the identification of the attribution of responsibility and the conflict frame the combination of supervised classification algorithms (respectively SVM and SGD) combined with embedding-based tfidf vectorizers proved most beneficial. Specifically, the embedding models configured with 300 dimensions proved to boost performance to the most competitive level.

For the classification of the economic consequences and human interest frame, f1-scores slightly improved classification effectiveness. When inspecting the top three classifiers for the economic consequences and human interest frame, it becomes clear that traditional BoW vectorizers and embedding based vectorizers are highly competitive: The differences in performance are only very marginal. In conclusion, using embedding-based vectorizers may boost performance of the classification of frames—although traditional BoW-vectorizers may suffice.

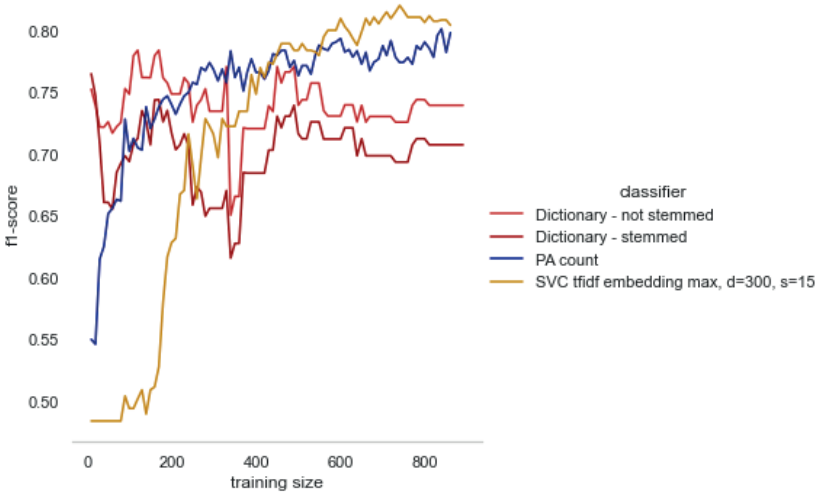
Figure 3 displays the learning curves of the best performing supervised algorithm using the BoW-vectorizer, the best performing algorithm using the embedding-based vectorizer, and the stemmed and unstemmed versions of the dictionaries. Results are based on random train-test splits, with an incremental increase of the train set of 10 documents. Although the cut-off points differ across frames, it becomes evident that the performance of the tailor-made dictionaries does not improve substantially when adding more than 300-400 random training examples. For example, performance of the dictionary aiming to capture the conflict frame fluctuates across different training sizes, but does not substantially increase. This can potentially be explained by the fact that multi-word phrases, semantic and syntactic context, and lexical patterns are not considered—which hampers the ability of the dictionaries to become



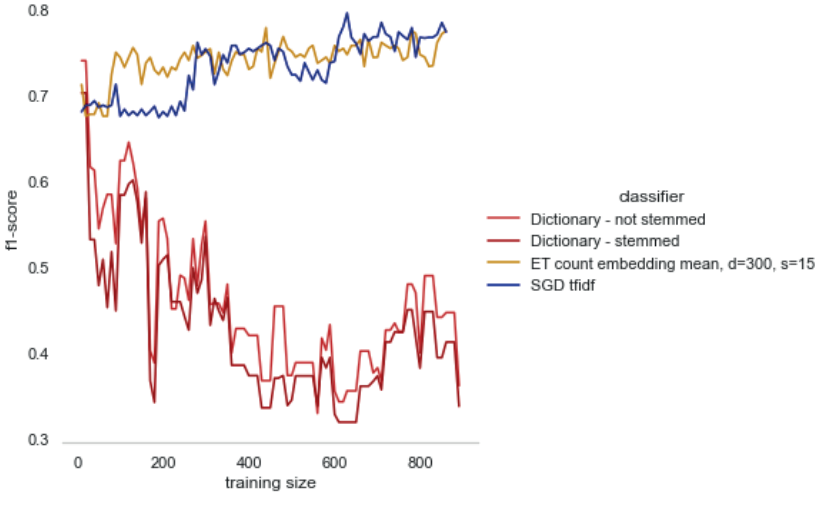

Attribution of responsibility




Conflict




Economic consequences




Human Interest

**Figure 3.** Effectiveness (F1-Scores) of Frame Classification across Classifiers for Different Sizes of Training Sets. D = dimensionality, S = window size.

better at learning how to classify the news articles. Hence, the additional costs associated with coding more news articles to arrive at more inclusive dictionaries does not seem to pay off here. On the other hand, performance of the supervised algorithms with BoW or embedding-based vectorizers *does* improve when the number of labelled examples increases. For supervised algorithms, investing time and resources in creating a labelled dataset actually results in better performance. One may expect that performance would further improve if we had collected more labelled examples ourselves, as the learning curves for most of our frames did not flatten out yet.

## Discussion

"Validate, validate, validate", reads one of the key principles for automated content analysis in the social sciences (Grimmer& Stewart, 2013, p. 271). Yet, despite this strong incitement, only few attempts have been made to scrutinize agreement in classifying *equivalent* concepts at the heart of communication research using *different* deductive (i.e., "top-down") computational approaches. Accordingly, this study was set out to compare and contrast the efficacy of dictionary approaches and supervised machine learning algorithms to measure key concepts in communication research, and to scrutinize the usefulness of word embeddings to these tasks. Although a comparison between bag-of-words or lexicon analyses to supervised machine learning approaches is not new, the current study's contribution lies in demonstrating how such methods might be implemented and validated by communication scholars.

The here-presented findings confirm that scholars' choice of method has substantial implications for the quality of their findings. Supervised machine learning proved to outperform dictionary-based classification for both the identification of policy topics and frames. These results confirm previous findings (e.g., Hailong et al., 2014). Furthermore, the results show that both the classification of policy topics and frames may—but not necessarily—benefits from including word embeddings that capture information about the broader semantical context from which the training data is derived. Our findings suggest that traditional bag-of-words document representation models (i.e., count and tfidf vectorizers) have advantages and may be sufficient to reach the most optimal level of performance. Yet, vectorizing the textual data using a self-trained domain-specific word embedding model proved to boost performance for several classifiers and tasks. Specifically, we found that our supervised algorithms using embedding based vectorizer

reached highest performance for the classification of topics and two news frames. By preserving relevant information about the semantic meaning of words in the larger population of media, political, or organizational content at relatively low dimensionality and costs, embedding models represent a promising tool for communication scholars aiming to improve the performance of their classification task. In sum, we extend Rudkowsky et al.'s (2018) conclusion that a word embeddings approach has merits for classification tasks at the heart of communication science and the social sciences more broadly.

It should be noted that the dictionary used to classify policy topics were developed in a neighboring, yet different political context than the Netherlands (Albaugh et al., 2013). Regardless, the finding that dictionary-based approaches performed relatively poorly resonates with concerns voiced in the literature. Particularly, scholars argued that some words do not frequently occur whilst being essential to the meaning of a text (Hertog & McLeod, 2001; Matthes & Kohring, 2008). Following from this, it is hard to set specific rules for how many words, or word-combinations, should be present in a text for a category to be present. This is an inherent problem related to dictionary-based analysis (e.g., Burscher et al., 2014; Günther & Quandt, 2016).

Communication scholars are often confronted with limited research resources, data of inferior quality, and high time pressure. While explicitly acknowledging this daily reality, the current study aimed to help communication scholars make informed decisions about which tool to select from the computational communication toolkit. In light of these limitations, one may want to reflect on some of the specific advantages and disadvantages of the here-discussed methods in terms of performance and costs. First, and although limited performance should be acknowledged, off-the-shelf dictionaries are competitive when budgets are limited and researchers strive towards a transparent method of classification. An additional benefit of dictionary-approaches is that the quantity and quality of the training dataset does not affect the final results (Hailong et al., 2014). However, as performance might be below par (as was the case in our exploration)—manual validation for specific domain, genre and language of the research project is key (Boukes et al., 2019). Such validation efforts may very well signal the need to adapt or refine the dictionary to the context at hand: Tailoring an off-the-shelf dictionary to the dataset under investigation likely burdens researchers with additional costs and time investments but might be necessary to reach satisfactory performance.

An important lesson learnt from our efforts to manually construct a tailor-made dictionary capturing news frames is that such an approach

may be competitive to supervised classifiers when training examples are limited. This means that when resource budgets constrain the manual labelling of large collections of texts (>300-400 random articles), one may opt for manually constructing or tailoring existing dictionaries. Yet, we found that investing time and resources in creating a larger labelled dataset actually resulted in better performance for our supervised classifiers, and should therefore be preferred.

Last, our conclusions that adding word embeddings to the supervised machine learning pipeline is especially interesting for communication scholars with (1) limited resources to manually annotate a large dataset, but (2) *do* have access to the larger population of documents (e.g., social media posts, news articles, or press releases) that the training set was drawn from. More specifically, our findings suggest that vectorizing documents using a word embedding model trained on the larger population of news articles and parliamentary questions from which our training sample was drawn helped boost performance *without* requesting additional financial resources. More specifically, because we had access to the larger population of textual data, training embedding models with different settings was a mere additional step in the computational analyses that allowed us to introduce information about the semantic relation between words in the population of interest to our models. It should be noted, however, that learning a word embedding model may be time-consuming and likely only benefits performance when one has access to a large population of documents. On the other hand, one may try to use pre-trained word embedding models (such as Word2Vec and GloVE) that also exists for smaller languages.

A crucial drawback of computer-assisted content analysis is that words or phrases are typically assigned a single meaning, disregarding the ambiguities, complexities, and manifold meaning interlaced in language (Matthes & Kohring, 2008). Word embeddings, however, directly address this limitation. Capturing the meaning of words in a *n*-dimensional space, embedding models are better equipped in capturing nuances and complexities in language. Adding these nuances and complexities to the classification model significantly boosted performance in our classification models. Importantly, the embedding model helped understanding the meaning for words not occurring in the training data. For example, even if the word "quarrel" did not occur in the training data, a sound embedding model should still be able to capture part of its meaning—and grasp its close relation to words such as "argument" or "controversy"—ultimately enabling the identification of, in our case, the *conflict frame*.

Some nuances and limitations of the here-reported findings should be acknowledged. First of all, it should be noted that the current study

explored the effectiveness of two dictionaries. We cannot draw conclusions about the effectiveness of other dictionaries than the ones tested here, especially as large variation exists among their quality and validity. Yet, in line with previous research (Boukes et al., 2019) questioning the validity of off-the-shelf dictionaries—especially when applied out-of-context—we believe it is important for researchers to be aware of the possible drawbacks associated with this form of textual analysis. Second, the current study found, in line with Rudkowsky et al. (2018), that word embeddings help boost performance in classification challenges central to communication theory. It should be noted, however, that access to a high-quality embedding model is a prerequisite for achieving these benefits. If researchers do not have access to the broader population of news articles, press releases, or parliamentary questions from which their training data is drawn, off-the-shelf embedding models (in the language of interest) may offer a solution. Such models, however, may not straightforwardly or to the same extent improve classification performance. This may particularly be the case when there is a mismatch in terms of the domain from which the training dataset and embedding model originate, such that meaning in the embedding model does not translate well to the context of interest. Ultimately, what "works" best is an empirical question that should be tested rigorously and empirically. Third, it should be noted that intercoder reliability of some of the frames were not optimal. Supervised machine learning algorithms will likely suffer from such flawed training data (cf. Burscher et al., 2014). This is a serious limitation for the quality of the classification as training was not possible on a reliable basis. Future studies may want to use standard datasets with proven annotation quality to overcome this problem.

Finally, although the across-the-board performance of the dictionaries tested in this study was lower compared to that of most supervised algorithms, they might be useful in the first phase of architecting an effectively supervised classifier: data sampling for training data. More specifically, dictionaries may help create a more *balanced sample* than would be the case when taking a random sample (Albaugh et al., 2014). This decreases the likelihood of class imbalance, a common issue in supervised text classification: Researchers can better assure those infrequent occurring categories are well-represented in the training dataset.

Altogether, this study has demonstrated that scholars aiming to automate the measurement of policy topics or frames are well-advised to use high-quality and domain-specific word embeddings in a supervised classification task. The accuracy of their results is likely to increase by adding complexity

and nuance to their models, and herewith move beyond simple bag-of-words models and accompanying inferior classification results.

## Author Note

Correspondence concerning this article should be addressed to Anne Kroon, Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Nieuwe Achtergracht 166, 1018WV, Amsterdam.
E-mail: A.C.Kroon@uva.nl.

## Appendix A

### Tailor-Made Dictionaries Measuring Frames

*Attribution of responsibility*
minister, kabinet, gemeente, staatssecretaris, verantwoordelijk, plannen, kamer, verantwoordelijkheid, wetsvoorstel, maatregelen, gemeenteraad, onderzoek, overheid, oplossing, wet, nieuwe, plan, politieke, burgemeester, opheldering, ministerie, regering, voorstel, wethouder, besluit, advies, politiek, aanpakken, beleid, bijdrage

*Human Interest*
mensen, vrouw, kinderen, persoonlijke, vrouwen, echtgenote, zoon, ouders, voormalig, meisje, werk, politiek, moeder, slachtoffer, kwetsbare, man, stem, jonge, turkse, familie, jongen, zoontje, balkenende, zwaar, verleden, jongeren, overleden, militair, vrienden, vriend

*Conflict*
kritiek, problemen, strijd, grote, fel, probleem, verwijt, conflict, crisis, steun, boos, woede, ruzie, voorstanders, verzet, verlies, discussie, afstand, beleid, boze, fout, meerderheid, tegenstanders, opheldering, eist, tegenstander, winnaar, slecht, zeer, bezwaren

*Economic Consequences*
euro, miljoen, miljard, geld, gulden, kosten, betalen, financiële, bedrag, bezuinigingen, economische, budget, miljarden, financieel, belasting, goedkoper, begroting, euros, bedragen, miljoenen, belastingbetaler, vergoeding, prijs, extra, eur, subsidies, honderden, financiering, duur, rekening

**Table A1. Performance of Policy Issue Classification across Classifiers Per Topico**

| Topic | precision | recall | f1-score | classifier |
|---|---|---|---|---|
| Banking, finance, & commerce | 0.25 | 0.08 | 0.13 | Albaugh et al. – not stemmed (dictionary, index-based) |
| Banking, finance, & commerce | 0.21 | 0.08 | 0.13 | Albaugh et al. – not stemmed (dictionary, count-based) |
| Banking, finance, & commerce | 0.35 | 0.33 | 0.34 | Albaugh et al. – stemmed (dictionary, index-based) |
| Banking, finance, & commerce | 0.27 | 0.34 | 0.34 | Albaugh et al. – stemmed (dictionary, count-based) |
| Banking, finance, & commerce | 0.24 | 0.80 | 0.36 | SVM, BoW tfidf vectorizer |
| Banking, finance, & commerce | 0.29 | 0.71 | 0.42 | SVM, embedding tfidf vectorizer |
| Banking, finance, & commerce | 0.24 | 0.67 | 0.35 | SVM, BoW count vectorizer |
| Banking, finance, & commerce | 0.29 | 0.71 | 0.42 | SVM, embedding count vectorizer |
| Banking, finance, & commerce | 0.24 | 0.67 | 0.35 | Passive Agressive, BoW tfidf vectorizer |
| Banking, finance, & commerce | 0.18 | 1.00 | 0.30 | Passive Agressive, embedding tfidf vectorizer |
| Banking, finance, & commerce | 0.24 | 0.67 | 0.35 | Passive Agressive, BoW count vectorizer |
| Banking, finance, & commerce | 0.12 | 0.67 | 0.20 | Passive Agressive, embedding count vectorizer |
| Banking, finance, & commerce | 0.24 | 0.67 | 0.35 | Stochastic Gradient Descent (SGD), BoW tfidf vectorizer |
| Banking, finance, & commerce | 0.12 | 0.40 | 0.18 | Stochastic Gradient Descent (SGD), embedding tfidf vectorizer |
| Banking, finance, & commerce | 0.24 | 0.67 | 0.35 | Stochastic Gradient Descent (SGD), BoW count vectorizer |
| Banking, finance, & commerce | 0.18 | 0.50 | 0.26 | Stochastic Gradient Descent (SGD), embedding count vectorizer |
| Banking, finance, & commerce | 0.24 | 0.50 | 0.32 | ExtraTrees, BoW tfidf vectorizer |
| Banking, finance, & commerce | 0.24 | 0.50 | 0.32 | ExtraTrees, embedding tfidf vectorizer |
| Banking, finance, & commerce | 0.24 | 0.57 | 0.33 | ExtraTrees, BoW count vectorizer |
| Banking, finance, & commerce | 0.29 | 0.63 | 0.40 | ExtraTrees, embedding count vectorizer |
| Civil rights | 0.36 | 0.02 | 0.05 | Albaugh et al. – not stemmed (dictionary, index-based) |
| Civil rights | 0.29 | 0.02 | 0.05 | Albaugh et al. – not stemmed (dictionary, count-based) |
| Civil rights | 0.41 | 0.05 | 0.10 | Albaugh et al. – stemmed (dictionary, index-based) |
| Civil rights | 0.28 | 0.08 | 0.10 | Albaugh et al. – stemmed (dictionary, count-based) |
| Civil rights | 0.38 | 0.43 | 0.40 | SVM, BoW tfidf vectorizer |
| Civil rights | 0.46 | 0.39 | 0.42 | SVM, embedding tfidf vectorizer |
| Civil rights | 0.25 | 0.26 | 0.26 | SVM, BoW count vectorizer |

| | | | | |
|---|---|---|---|---|
| Civil rights | 0.42 | 0.37 | 0.39 | SVM, embedding count vectorizer |
| Civil rights | 0.42 | 0.31 | 0.36 | Passive Agressive, BoW tfidf vectorizer |
| Civil rights | 0.54 | 0.31 | 0.39 | Passive Agressive, embedding tfidf vectorizer |
| Civil rights | 0.33 | 0.25 | 0.29 | Passive Agressive, BoW count vectorizer |
| Civil rights | 0.46 | 0.38 | 0.42 | Passive Agressive, embedding count vectorizer |
| Civil rights | 0.25 | 0.32 | 0.28 | Stochastic Gradient Descent (SGD), BoW tfidf vectorizer |
| Civil rights | 0.25 | 0.33 | 0.29 | Stochastic Gradient Descent (SGD), embedding tfidf vectorizer |
| Civil rights | 0.46 | 0.31 | 0.37 | Stochastic Gradient Descent (SGD), BoW count vectorizer |
| Civil rights | 0.17 | 0.33 | 0.22 | Stochastic Gradient Descent (SGD), embedding count vectorizer |
| Civil rights | 0.21 | 0.24 | 0.22 | ExtraTrees, BoW tfidf vectorizer |
| Civil rights | 0.25 | 0.19 | 0.21 | ExtraTrees, embedding tfidf vectorizer |
| Civil rights | 0.13 | 0.15 | 0.14 | ExtraTrees, BoW count vectorizer |
| Civil rights | 0.33 | 0.36 | 0.35 | ExtraTrees, embedding count vectorizer |
| Defense | 0.59 | 0.19 | 0.29 | Albaugh et al. – not stemmed (dictionary, index-based) |
| Defense | 0.54 | 0.20 | 0.29 | Albaugh et al. – not stemmed (dictionary, count-based) |
| Defense | 0.67 | 0.25 | 0.36 | Albaugh et al. – stemmed (dictionary, index-based) |
| Defense | 0.54 | 0.27 | 0.36 | Albaugh et al. – stemmed (dictionary, count-based) |
| Defense | 0.50 | 0.72 | 0.59 | SVM, BoW tfidf vectorizer |
| Defense | 0.65 | 0.81 | 0.72 | SVM, embedding tfidf vectorizer |
| Defense | 0.35 | 0.56 | 0.43 | SVM, BoW count vectorizer |
| Defense | 0.65 | 0.81 | 0.72 | SVM, embedding count vectorizer |
| Defense | 0.69 | 0.67 | 0.68 | Passive Agressive, BoW tfidf vectorizer |
| Defense | 0.77 | 0.69 | 0.73 | Passive Agressive, embedding tfidf vectorizer |
| Defense | 0.65 | 0.71 | 0.68 | Passive Agressive, BoW count vectorizer |
| Defense | 0.77 | 0.67 | 0.71 | Passive Agressive, embedding count vectorizer |
| Defense | 0.73 | 0.63 | 0.68 | Stochastic Gradient Descent (SGD), BoW tfidf vectorizer |
| Defense | 0.77 | 0.74 | 0.75 | Stochastic Gradient Descent (SGD), embedding tfidf vectorizer |
| Defense | 0.62 | 0.73 | 0.67 | Stochastic Gradient Descent (SGD), BoW count vectorizer |
| Defense | 0.88 | 0.77 | 0.82 | Stochastic Gradient Descent (SGD), embedding count vectorizer |
| Defense | 0.35 | 0.53 | 0.42 | ExtraTrees, BoW tfidf vectorizer |

| | | | | |
|---|---|---|---|---|
| Defense | 0.31 | 0.50 | 0.38 | ExtraTrees, embedding tfidf vectorizer |
| Defense | 0.19 | 0.56 | 0.29 | ExtraTrees, BoW count vectorizer |
| Defense | 0.15 | 0.33 | 0.21 | ExtraTrees, embedding count vectorizer |
| Education | 0.33 | 0.22 | 0.27 | Albaugh et al. – not stemmed (dictionary, index-based) |
| Education | 0.30 | 0.22 | 0.27 | Albaugh et al. – not stemmed (dictionary, count-based) |
| Education | 0.41 | 0.62 | 0.49 | Albaugh et al. – stemmed (dictionary, index-based) |
| Education | 0.39 | 0.62 | 0.49 | Albaugh et al. – stemmed (dictionary, count-based) |
| Education | 0.56 | 0.83 | 0.67 | SVM, BoW tfidf vectorizer |
| Education | 0.72 | 0.68 | 0.70 | SVM, embedding tfidf vectorizer |
| Education | 0.50 | 0.60 | 0.55 | SVM, BoW count vectorizer |
| Education | 0.72 | 0.81 | 0.76 | SVM, embedding count vectorizer |
| Education | 0.72 | 0.68 | 0.70 | Passive Agressive, BoW tfidf vectorizer |
| Education | 0.44 | 0.62 | 0.52 | Passive Agressive, embedding tfidf vectorizer |
| Education | 0.56 | 0.59 | 0.57 | Passive Agressive, BoW count vectorizer |
| Education | 0.83 | 0.38 | 0.53 | Passive Agressive, embedding count vectorizer |
| Education | 0.83 | 0.68 | 0.75 | Stochastic Gradient Descent (SGD), BoW tfidf vectorizer |
| Education | 0.78 | 0.58 | 0.67 | Stochastic Gradient Descent (SGD), embedding tfidf vectorizer |
| Education | 0.50 | 0.60 | 0.55 | Stochastic Gradient Descent (SGD), BoW count vectorizer |
| Education | 0.78 | 0.56 | 0.65 | Stochastic Gradient Descent (SGD), embedding count vectorizer |
| Education | 0.44 | 0.89 | 0.59 | ExtraTrees, BoW tfidf vectorizer |
| Education | 0.39 | 0.64 | 0.48 | ExtraTrees, embedding tfidf vectorizer |
| Education | 0.50 | 0.53 | 0.51 | ExtraTrees, BoW count vectorizer |
| Education | 0.33 | 0.46 | 0.39 | ExtraTrees, embedding count vectorizer |
| Environment | 0.29 | 0.05 | 0.08 | Albaugh et al. – not stemmed (dictionary, index-based) |
| Environment | 0.26 | 0.06 | 0.08 | Albaugh et al. – not stemmed (dictionary, count-based) |
| Environment | 0.38 | 0.06 | 0.10 | Albaugh et al. – stemmed (dictionary, index-based) |
| Environment | 0.30 | 0.09 | 0.10 | Albaugh et al. – stemmed (dictionary, count-based) |
| Environment | 0.27 | 0.50 | 0.35 | SVM, BoW tfidf vectorizer |
| Environment | 0.14 | 0.50 | 0.21 | SVM, embedding tfidf vectorizer |
| Environment | 0.32 | 0.58 | 0.41 | SVM, BoW count vectorizer |

| | | | | |
|---|---|---|---|---|
| Environment | 0.14 | 0.38 | 0.20 | SVM, embedding count vectorizer |
| Environment | 0.27 | 0.50 | 0.35 | Passive Agressive, BoW tfidf vectorizer |
| Environment | 0.36 | 0.40 | 0.38 | Passive Agressive, embedding tfidf vectorizer |
| Environment | 0.32 | 0.58 | 0.41 | Passive Agressive, BoW count vectorizer |
| Environment | 0.27 | 0.38 | 0.32 | Passive Agressive, embedding count vectorizer |
| Environment | 0.27 | 0.60 | 0.37 | Stochastic Gradient Descent (SGD), BoW tfidf vectorizer |
| Environment | 0.18 | 0.27 | 0.22 | Stochastic Gradient Descent (SGD), embedding tfidf vectorizer |
| Environment | 0.27 | 0.55 | 0.36 | Stochastic Gradient Descent (SGD), BoW count vectorizer |
| Environment | 0.00 | 0.00 | 0.00 | Stochastic Gradient Descent (SGD), embedding count vectorizer |
| Environment | 0.18 | 0.40 | 0.25 | ExtraTrees, BoW tfidf vectorizer |
| Environment | 0.18 | 0.31 | 0.23 | ExtraTrees, embedding tfidf vectorizer |
| Environment | 0.18 | 0.33 | 0.24 | ExtraTrees, BoW count vectorizer |
| Environment | 0.23 | 0.33 | 0.27 | ExtraTrees, embedding count vectorizer |
| Governmental operations | 0.65 | 0.03 | 0.06 | Albaugh et al. – not stemmed (dictionary, index-based) |
| Governmental operations | 0.57 | 0.04 | 0.06 | Albaugh et al. – not stemmed (dictionary, count-based) |
| Governmental operations | 0.53 | 0.16 | 0.24 | Albaugh et al. – stemmed (dictionary, index-based) |
| Governmental operations | 0.50 | 0.16 | 0.24 | Albaugh et al. – stemmed (dictionary, count-based) |
| Governmental operations | 0.74 | 0.51 | 0.60 | SVM, BoW tfidf vectorizer |
| Governmental operations | 0.73 | 0.59 | 0.65 | SVM, embedding tfidf vectorizer |
| Governmental operations | 0.68 | 0.48 | 0.56 | SVM, BoW count vectorizer |
| Governmental operations | 0.75 | 0.57 | 0.65 | SVM, embedding count vectorizer |
| Governmental operations | 0.77 | 0.53 | 0.63 | Passive Agressive, BoW tfidf vectorizer |
| Governmental operations | 0.80 | 0.58 | 0.67 | Passive Agressive, embedding tfidf vectorizer |
| Governmental operations | 0.73 | 0.54 | 0.62 | Passive Agressive, BoW count vectorizer |
| Governmental operations | 0.57 | 0.66 | 0.61 | Passive Agressive, embedding count vectorizer |
| Governmental operations | 0.80 | 0.52 | 0.63 | Stochastic Gradient Descent (SGD), BoW tfidf vectorizer |
| Governmental operations | 0.80 | 0.54 | 0.64 | Stochastic Gradient Descent (SGD), embedding tfidf vectorizer |
| Governmental operations | 0.72 | 0.55 | 0.62 | Stochastic Gradient Descent (SGD), BoW count vectorizer |
| Governmental operations | 0.73 | 0.63 | 0.68 | Stochastic Gradient Descent (SGD), embedding count vectorizer |
| Governmental operations | 0.73 | 0.34 | 0.46 | ExtraTrees, BoW tfidf vectorizer |

| | | | | |
|---|---|---|---|---|
| Governmental operations | 0.68 | 0.47 | 0.56 | ExtraTrees, embedding tfidf vectorizer |
| Governmental operations | 0.69 | 0.37 | 0.48 | ExtraTrees, BoW count vectorizer |
| Governmental operations | 0.72 | 0.46 | 0.56 | ExtraTrees, embedding count vectorizer |
| Health | 0.38 | 0.10 | 0.15 | Albaugh et al. – not stemmed (dictionary, index-based) |
| Health | 0.35 | 0.10 | 0.15 | Albaugh et al. – not stemmed (dictionary, count-based) |
| Health | 0.74 | 0.33 | 0.46 | Albaugh et al. – stemmed (dictionary, index-based) |
| Health | 0.67 | 0.34 | 0.46 | Albaugh et al. – stemmed (dictionary, count-based) |
| Health | 0.68 | 0.84 | 0.75 | SVM, BoW tfidf vectorizer |
| Health | 0.75 | 0.82 | 0.78 | SVM, embedding tfidf vectorizer |
| Health | 0.63 | 0.82 | 0.71 | SVM, BoW count vectorizer |
| Health | 0.73 | 0.85 | 0.79 | SVM, embedding count vectorizer |
| Health | 0.71 | 0.79 | 0.75 | Passive Agressive, BoW tfidf vectorizer |
| Health | 0.86 | 0.68 | 0.76 | Passive Agressive, embedding tfidf vectorizer |
| Health | 0.62 | 0.65 | 0.63 | Passive Agressive, BoW count vectorizer |
| Health | 0.73 | 0.75 | 0.74 | Passive Agressive, embedding count vectorizer |
| Health | 0.75 | 0.71 | 0.73 | Stochastic Gradient Descent (SGD), BoW tfidf vectorizer |
| Health | 0.78 | 0.78 | 0.78 | Stochastic Gradient Descent (SGD), embedding tfidf vectorizer |
| Health | 0.63 | 0.71 | 0.67 | Stochastic Gradient Descent (SGD), BoW count vectorizer |
| Health | 0.81 | 0.80 | 0.80 | Stochastic Gradient Descent (SGD), embedding count vectorizer |
| Health | 0.59 | 0.66 | 0.62 | ExtraTrees, BoW tfidf vectorizer |
| Health | 0.70 | 0.69 | 0.69 | ExtraTrees, embedding tfidf vectorizer |
| Health | 0.59 | 0.79 | 0.67 | ExtraTrees, BoW count vectorizer |
| Health | 0.65 | 0.68 | 0.67 | ExtraTrees, embedding count vectorizer |
| Immigration & integration | 0.31 | 0.04 | 0.07 | Albaugh et al. – not stemmed (dictionary, index-based) |
| Immigration & integration | 0.39 | 0.06 | 0.07 | Albaugh et al. – not stemmed (dictionary, count-based) |
| Immigration & integration | 0.65 | 0.19 | 0.30 | Albaugh et al. – stemmed (dictionary, index-based) |
| Immigration & integration | 0.56 | 0.29 | 0.30 | Albaugh et al. – stemmed (dictionary, count-based) |
| Immigration & integration | 0.43 | 0.82 | 0.56 | SVM, BoW tfidf vectorizer |
| Immigration & integration | 0.43 | 0.50 | 0.46 | SVM, embedding tfidf vectorizer |
| Immigration & integration | 0.48 | 0.59 | 0.53 | SVM, BoW count vectorizer |

| | | | | |
|---|---|---|---|---|
| Immigration & integration | 0.43 | 0.56 | 0.49 | SVM, embedding count vectorizer |
| Immigration & integration | 0.52 | 0.73 | 0.61 | Passive Agressive, BoW tfidf vectorizer |
| Immigration & integration | 0.43 | 0.56 | 0.49 | Passive Agressive, embedding tfidf vectorizer |
| Immigration & integration | 0.57 | 0.52 | 0.55 | Passive Agressive, BoW count vectorizer |
| Immigration & integration | 0.48 | 0.56 | 0.51 | Passive Agressive, embedding count vectorizer |
| Immigration & integration | 0.52 | 0.55 | 0.54 | Stochastic Gradient Descent (SGD), BoW tfidf vectorizer |
| Immigration & integration | 0.57 | 0.52 | 0.55 | Stochastic Gradient Descent (SGD), embedding tfidf vectorizer |
| Immigration & integration | 0.52 | 0.50 | 0.51 | Stochastic Gradient Descent (SGD), BoW count vectorizer |
| Immigration & integration | 0.52 | 0.48 | 0.50 | Stochastic Gradient Descent (SGD), embedding count vectorizer |
| Immigration & integration | 0.19 | 0.40 | 0.26 | ExtraTrees, BoW tfidf vectorizer |
| Immigration & integration | 0.38 | 0.57 | 0.46 | ExtraTrees, embedding tfidf vectorizer |
| Immigration & integration | 0.43 | 0.60 | 0.50 | ExtraTrees, BoW count vectorizer |
| Immigration & integration | 0.33 | 0.37 | 0.35 | ExtraTrees, embedding count vectorizer |
| Int. affairs & foreign aid | 0.41 | 0.03 | 0.06 | Albaugh et al. – not stemmed (dictionary, index-based) |
| Int. affairs & foreign aid | 0.32 | 0.04 | 0.06 | Albaugh et al. – not stemmed (dictionary, count-based) |
| Int. affairs & foreign aid | 0.39 | 0.22 | 0.28 | Albaugh et al. – stemmed (dictionary, index-based) |
| Int. affairs & foreign aid | 0.39 | 0.28 | 0.28 | Albaugh et al. – stemmed (dictionary, count-based) |
| Int. affairs & foreign aid | 0.43 | 0.45 | 0.44 | SVM, BoW tfidf vectorizer |
| Int. affairs & foreign aid | 0.57 | 0.52 | 0.55 | SVM, embedding tfidf vectorizer |
| Int. affairs & foreign aid | 0.45 | 0.49 | 0.47 | SVM, BoW count vectorizer |
| Int. affairs & foreign aid | 0.48 | 0.43 | 0.45 | SVM, embedding count vectorizer |
| Int. affairs & foreign aid | 0.40 | 0.50 | 0.45 | Passive Agressive, BoW tfidf vectorizer |
| Int. affairs & foreign aid | 0.31 | 0.54 | 0.39 | Passive Agressive, embedding tfidf vectorizer |
| Int. affairs & foreign aid | 0.38 | 0.55 | 0.45 | Passive Agressive, BoW count vectorizer |
| Int. affairs & foreign aid | 0.31 | 0.81 | 0.45 | Passive Agressive, embedding count vectorizer |
| Int. affairs & foreign aid | 0.45 | 0.61 | 0.52 | Stochastic Gradient Descent (SGD), BoW tfidf vectorizer |
| Int. affairs & foreign aid | 0.36 | 0.60 | 0.45 | Stochastic Gradient Descent (SGD), embedding tfidf vectorizer |
| Int. affairs & foreign aid | 0.40 | 0.52 | 0.45 | Stochastic Gradient Descent (SGD), BoW count vectorizer |
| Int. affairs & foreign aid | 0.33 | 0.64 | 0.44 | Stochastic Gradient Descent (SGD), embedding count vectorizer |
| Int. affairs & foreign aid | 0.38 | 0.40 | 0.39 | ExtraTrees, BoW tfidf vectorizer |

| | | | | |
|---|---|---|---|---|
| Int. affairs & foreign aid | 0.40 | 0.47 | 0.44 | ExtraTrees, embedding tfidf vectorizer |
| Int. affairs & foreign aid | 0.50 | 0.39 | 0.44 | ExtraTrees, BoW count vectorizer |
| Int. affairs & foreign aid | 0.43 | 0.45 | 0.44 | ExtraTrees, embedding count vectorizer |
| Labor & employment | 0.43 | 0.13 | 0.20 | Albaugh et al. – not stemmed (dictionary, index-based) |
| Labor & employment | 0.35 | 0.14 | 0.20 | Albaugh et al. – not stemmed (dictionary, count-based) |
| Labor & employment | 0.53 | 0.31 | 0.39 | Albaugh et al. – stemmed (dictionary, index-based) |
| Labor & employment | 0.36 | 0.31 | 0.39 | Albaugh et al. – stemmed (dictionary, count-based) |
| Labor & employment | 0.40 | 0.46 | 0.43 | SVM, BoW tfidf vectorizer |
| Labor & employment | 0.40 | 0.40 | 0.40 | SVM, embedding tfidf vectorizer |
| Labor & employment | 0.47 | 0.44 | 0.45 | SVM, BoW count vectorizer |
| Labor & employment | 0.47 | 0.47 | 0.47 | SVM, embedding count vectorizer |
| Labor & employment | 0.40 | 0.38 | 0.39 | Passive Agressive, BoW tfidf vectorizer |
| Labor & employment | 0.47 | 0.37 | 0.41 | Passive Agressive, embedding tfidf vectorizer |
| Labor & employment | 0.40 | 0.30 | 0.34 | Passive Agressive, BoW count vectorizer |
| Labor & employment | 0.47 | 0.35 | 0.40 | Passive Agressive, embedding count vectorizer |
| Labor & employment | 0.27 | 0.36 | 0.31 | Stochastic Gradient Descent (SGD), BoW tfidf vectorizer |
| Labor & employment | 0.53 | 0.32 | 0.40 | Stochastic Gradient Descent (SGD), embedding tfidf vectorizer |
| Labor & employment | 0.40 | 0.40 | 0.40 | Stochastic Gradient Descent (SGD), BoW count vectorizer |
| Labor & employment | 0.53 | 0.38 | 0.44 | Stochastic Gradient Descent (SGD), embedding count vectorizer |
| Labor & employment | 0.07 | 0.08 | 0.07 | ExtraTrees, BoW tfidf vectorizer |
| Labor & employment | 0.20 | 0.20 | 0.20 | ExtraTrees, embedding tfidf vectorizer |
| Labor & employment | 0.20 | 0.18 | 0.19 | ExtraTrees, BoW count vectorizer |
| Labor & employment | 0.27 | 0.22 | 0.24 | ExtraTrees, embedding count vectorizer |
| Law & crime | 0.29 | 0.15 | 0.19 | Albaugh et al. – not stemmed (dictionary, index-based) |
| Law & crime | 0.26 | 0.15 | 0.19 | Albaugh et al. – not stemmed (dictionary, count-based) |
| Law & crime | 0.45 | 0.32 | 0.37 | Albaugh et al. – stemmed (dictionary, index-based) |
| Law & crime | 0.42 | 0.36 | 0.37 | Albaugh et al. – stemmed (dictionary, count-based) |
| Law & crime | 0.73 | 0.62 | 0.67 | SVM, BoW tfidf vectorizer |
| Law & crime | 0.73 | 0.66 | 0.69 | SVM, embedding tfidf vectorizer |
| Law & crime | 0.64 | 0.58 | 0.61 | SVM, BoW count vectorizer |

| | | | | |
|---|---|---|---|---|
| Law & crime | 0.66 | 0.64 | 0.65 | SVM, embedding count vectorizer |
| Law & crime | 0.76 | 0.62 | 0.68 | Passive Agressive, BoW tfidf vectorizer |
| Law & crime | 0.59 | 0.57 | 0.58 | Passive Agressive, embedding tfidf vectorizer |
| Law & crime | 0.68 | 0.55 | 0.61 | Passive Agressive, BoW count vectorizer |
| Law & crime | 0.54 | 0.70 | 0.61 | Passive Agressive, embedding count vectorizer |
| Law & crime | 0.76 | 0.63 | 0.69 | Stochastic Gradient Descent (SGD), BoW tfidf vectorizer |
| Law & crime | 0.80 | 0.62 | 0.70 | Stochastic Gradient Descent (SGD), embedding tfidf vectorizer |
| Law & crime | 0.71 | 0.61 | 0.66 | Stochastic Gradient Descent (SGD), BoW count vectorizer |
| Law & crime | 0.81 | 0.58 | 0.68 | Stochastic Gradient Descent (SGD), embedding count vectorizer |
| Law & crime | 0.46 | 0.56 | 0.50 | ExtraTrees, BoW tfidf vectorizer |
| Law & crime | 0.58 | 0.51 | 0.54 | ExtraTrees, embedding tfidf vectorizer |
| Law & crime | 0.51 | 0.48 | 0.50 | ExtraTrees, BoW count vectorizer |
| Law & crime | 0.58 | 0.49 | 0.53 | ExtraTrees, embedding count vectorizer |
| Other issue | 0.21 | 0.86 | 0.33 | Albaugh et al. – not stemmed (dictionary, index-based) |
| Other issue | 0.21 | 0.83 | 0.33 | Albaugh et al. – not stemmed (dictionary, count-based) |
| Other issue | 0.26 | 0.82 | 0.39 | Albaugh et al. – stemmed (dictionary, index-based) |
| Other issue | 0.27 | 0.74 | 0.39 | Albaugh et al. – stemmed (dictionary, count-based) |
| Other issue | 0.66 | 0.49 | 0.56 | SVM, BoW tfidf vectorizer |
| Other issue | 0.68 | 0.58 | 0.63 | SVM, embedding tfidf vectorizer |
| Other issue | 0.64 | 0.51 | 0.57 | SVM, BoW count vectorizer |
| Other issue | 0.69 | 0.57 | 0.62 | SVM, embedding count vectorizer |
| Other issue | 0.59 | 0.61 | 0.60 | Passive Agressive, BoW tfidf vectorizer |
| Other issue | 0.59 | 0.76 | 0.66 | Passive Agressive, embedding tfidf vectorizer |
| Other issue | 0.55 | 0.55 | 0.55 | Passive Agressive, BoW count vectorizer |
| Other issue | 0.74 | 0.56 | 0.64 | Passive Agressive, embedding count vectorizer |
| Other issue | 0.60 | 0.63 | 0.61 | Stochastic Gradient Descent (SGD), BoW tfidf vectorizer |
| Other issue | 0.55 | 0.71 | 0.62 | Stochastic Gradient Descent (SGD), embedding tfidf vectorizer |
| Other issue | 0.60 | 0.56 | 0.58 | Stochastic Gradient Descent (SGD), BoW count vectorizer |
| Other issue | 0.68 | 0.61 | 0.65 | Stochastic Gradient Descent (SGD), embedding count vectorizer |
| Other issue | 0.45 | 0.41 | 0.43 | ExtraTrees, BoW tfidf vectorizer |

| | | | | |
|---|---|---|---|---|
| Other issue | 0.55 | 0.50 | 0.52 | ExtraTrees, embedding tfidf vectorizer |
| Other issue | 0.45 | 0.42 | 0.43 | ExtraTrees, BoW count vectorizer |
| Other issue | 0.55 | 0.52 | 0.53 | ExtraTrees, embedding count vectorizer |
| Social welfare | 0.40 | 0.02 | 0.04 | Albaugh et al. – not stemmed (dictionary, index-based) |
| Social welfare | 0.41 | 0.07 | 0.04 | Albaugh et al. – not stemmed (dictionary, count-based) |
| Social welfare | 1.00 | 0.02 | 0.04 | Albaugh et al. – stemmed (dictionary, index-based) |
| Social welfare | 0.53 | 0.08 | 0.04 | Albaugh et al. – stemmed (dictionary, count-based) |
| Social welfare | 0.14 | 0.60 | 0.23 | SVM, BoW tfidf vectorizer |
| Social welfare | 0.05 | 0.17 | 0.07 | SVM, embedding tfidf vectorizer |
| Social welfare | 0.14 | 0.30 | 0.19 | SVM, BoW count vectorizer |
| Social welfare | 0.05 | 0.13 | 0.07 | SVM, embedding count vectorizer |
| Social welfare | 0.14 | 0.43 | 0.21 | Passive Agressive, BoW tfidf vectorizer |
| Social welfare | 0.05 | 0.25 | 0.08 | Passive Agressive, embedding tfidf vectorizer |
| Social welfare | 0.14 | 0.38 | 0.21 | Passive Agressive, BoW count vectorizer |
| Transportation | 0.47 | 0.20 | 0.28 | Albaugh et al. – not stemmed (dictionary, index-based) |
| Transportation | 0.40 | 0.20 | 0.28 | Albaugh et al. – not stemmed (dictionary, count-based) |
| Transportation | 0.68 | 0.37 | 0.48 | Albaugh et al. – stemmed (dictionary, index-based) |
| Transportation | 0.57 | 0.39 | 0.48 | Albaugh et al. – stemmed (dictionary, count-based) |
| Transportation | 0.62 | 0.78 | 0.69 | SVM, BoW tfidf vectorizer |
| Transportation | 0.71 | 0.69 | 0.70 | SVM, embedding tfidf vectorizer |
| Transportation | 0.59 | 0.69 | 0.63 | SVM, BoW count vectorizer |
| Transportation | 0.65 | 0.73 | 0.69 | SVM, embedding count vectorizer |
| Transportation | 0.68 | 0.70 | 0.69 | Passive Agressive, BoW tfidf vectorizer |
| Transportation | 0.79 | 0.57 | 0.67 | Passive Agressive, embedding tfidf vectorizer |
| Transportation | 0.56 | 0.61 | 0.58 | Passive Agressive, BoW count vectorizer |

## Notes

1. It does so by updating the *cost function* with each iteration; The cost function measures the model's ability to estimate the relationship between X and y (typically by expressing the distance between the predicted and actual value).
2. This search string contains references to all existing and active political parties within a specific time frame.
3. Using a comprehensive Dutch stopword (https://github.com/stopwords-iso/stopwords-nl)

## References

Al-Azani, S., & El-Alfy, E. S. M. (2017). Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short Arabic text. *Procedia Computer Science*, *109*, 359–366. https://doi.org/10.1016/j.procs.2017.05.365

Albaugh, Q., Sevenans, J., Soroka, S., & Loewen, P. J. (2013). The automated coding of policy agendas: A dictionary-based approach. *6th Annual Comparative Agendas Conference*, 1–22.

Albaugh, Q., Soroka, S., Joly, J., Loewen, P., Sevenans, J., & Walgrave, S. (2014). Comparing and combining machine learning and dictionary-based approaches to topic coding. *7th Annual Comparative Agendas Project (CAP) Conference, March 2017*, 1–18.

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, *29*(1), 19–42. https://doi.org/10.1017/pan.2020.8

Bilbro, R., Ojeda, T., & Bengfort, B. (2018). *Applied text analysis with Python*. O'Reilly Media, Incorporated. https://books.google.nl/books?id=IrBqswEACAAJ

Boukes, M., van de Velde, B., Araujo, T., & Vliegenthart, R. (2019). What's the tone? Easy doesn't do It: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures*, *00*(00), 1–22. https://doi.org/10.1080/19312458.2019.1671966

Boumans, J. (2017). Subsidizing the news?: Organizational press releases' influence on news media's agenda and content. *Journalism Studies*, *0*(0), 1–19. https://doi.org/10.1080/1461670X.2017.1338154

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, *80*(Specialissue1), 250–271. https://doi.org/10.1093/poq/nfw007

Burscher, Björn, Odijk, D., Vliegenthart, R., de Rijke, M., & 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. https://doi.org/10.1080/19312458.2014.937527

Burscher, Bjorn, Vliegenthart, R., & De Vreese, C. H. (2015). Using supervised machine learning to code policy issues: Can classifiers generalize across contexts? *Annals of the American Academy of Political and Social Science*, *659*(1), 122–131. https://doi.org/10.1177/0002716215569441

Chan, C., Bajjalieh, J., Auvil, L., Wessler, H., Althaus, S., Welbers, K., van Atteveldt, W., & Jungblut, M. (2021). Four best practices for measuring news sentiment using 'off-the-shelf' dictionaries: a large-scale p-hacking experiment. *Computational Communication Research*, *3*(1), 1–27. https://doi.org/10.5117/ccr2021.1.001.chan

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*, *64*(2), 317–332. https://doi.org/10.1111/jcom.12084

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglue, K., & Kuksa, P. (2011). Natural Language Processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537. https://doi.org/10.1109/CIC.2017.00050

Corrêa, E. A., Marinho, V. Q., Borges, L., & Santos, D. (2017). A multi-view ensemble for twitter sentiment analysis. *Proceedings Ofthe 11th International Workshop on Semantic Evaluations* (*SemEval-2017*), 611–615. https://business.twitter.com/en/basics.html

Damstra, A., & Boukes, M. (2018). The economy, the news, and the public: A longitudinal study of the impact of economic news on economic evaluations and expectations. *Communication Research*, 009365021775097. https://doi.org/10.1177/0093650217750971

Djerf-Pierre, M., & Shehata, A. (2017). Still an agenda setter: Traditional news media and public opinion during the transition from low to high choice media environments. *Journal of Communication*, *67*(5), 733–757. https://doi.org/10.1111/jcom.12327

Ferrero, J., Besacier, L., Agnes, F., & Schwab, D. (2017). Using word embedding for cross-language plagiarism detection. *Proceedings Ofthe 15th Conference Ofthe European Chapter Ofthe Association for Computational Linguistics: Volume 2, Short Papers*, *2*, 415–421.

Firth, J. R. (1935). The techniques of semantics. *Transactions of Philological Society*, *34*(1), 36–77.

Firth, J. R. (1957). *Papers in linguistics, 1934-1951*. Oxford University Press. https://books.google.nl/books?id=yxZZAAAAMAAJ

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, *69*, 214–224. https://doi.org/10.1016/j.eswa.2016.10.043

Grimmer, J., & 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. https://doi.org/10.1093/pan/mps028

Günther, E., & Quandt, T. (2016). Word counts and topic models: Automated text analysis methods for digital journalism research. *Digital Journalism*, *4*(1), 75–88. https://doi.org/10.1080/21670811.2015.1093270

Guo, L., & Vargo, C. (2015). The power of message networks: A big-data analysis of the network agenda setting model and issue ownership. *Mass Communication and Society*, *18*(5), 557–576. https://doi.org/10.1080/15205436.2015.1045300

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism and Mass Communication Quarterly*, *93*(2), 322–359. https://doi.org/10.1177/1077699016639231

Hailong, Z., Wenyan, G., & Bo, J. (2014). Machine learning and lexicon based methods for sentiment classification: A survey. *Proceedings – 11th Web Information System and Application Conference, WISA 2014*, 262–265. https://doi.org/10.1109/WISA.2014.55

Hertog, J. K., & McLeod, D. M. (2001). A multiperspectival approach to framing analysis: A field guide. In *Framing public life* (pp. 157–178). Routledge.

King, G., Lam, P., & Roberts, M. E. (2017). Computer-Assisted Keyword and Document Set Discovery from Unstructured Text. *American Journal of Political Science*, *61*(4), 971–988. https://doi.org/10.1111/ajps.12291

King, G., Schneer, B., & White, A. (2017). How the news media activate public expression and influence national agendas. *Science*, *358*(6364), 776–780.

Kroon, A. C., & van der Meer, T. G. L. A. (2018). Who Takes the Lead? Investigating the Reciprocal Relationship Between Organizational and News Agendas. *Communication Research*. https://doi.org/10.1177/0093650217751733

Kübler, R. V., Wieringa, J. E., & Pauwels, K. H. (2017). Machine learning and big data. In P. K. Leeflang P., Wieringa J., Bijmolt T. (Ed.), *Advanced Methods for Modeling Markets. International Series in Quantitative Marketing.* Springer.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on International Conference on Machine Learning – Volume 32*, 1188--1196. http://dl.acm.org/citation.cfm?id=3044805.3045025

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Maher, M. (2001). Framing: An Emerging Paradigm or A Phase of Agenda Setting? *Framing Public Life: Perspectives on Media and Our Understanding of the Social World*, *1972*, 83–94.

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

Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv*, 1–12.

Neuman, W. R., Guggenheim, L., Jang, S. M., & Bae, S. Y. (2014). The dynamics of public attention: Agenda-setting theory meets big data. *Journal of Communication*, *64*(2), 193–214. https://doi.org/10.1111/jcom.12088

Opperhuizen, A. E., Schouten, K., & Klijn, E. H. (2019). Framing a conflict! How media report on earthquake risks caused by gas drilling: A longitudinal analysis using machine learning techniques of media reporting on gas drilling from 1990 to 2015. *Journalism Studies*, *20*(5), 714–734. https://doi.org/10.1080/1461670X.2017.1418672

Ribeiro, F., Araújo, M., Gonçalves, P., Gonçalves, M., & Benevenuto, F. (2015). SentiBench – A benchmark comparison of state-of-the-practice sentiment analysis methods. *Arxiv*, *9*(4), 1–32. http://arxiv.org/abs/1512.01818

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2–3), 140–157. https://doi.org/10.1080/19312458.2018.1455817

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

Semetko, H. A., & Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of Communication*, *50*(2), 93–109. https://doi.org/10.1111/j.1460-2466.2000.tb02843.x

Simon, A. F. (2001). A unified method for analyzing media framing. *Communication in US Elections: New Agendas*, 75–89.

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting Impoliteness and Incivility in Online Discussions. *Computational Communication Research*, *2*(1), 109–134. https://doi.org/10.5117/ccr2020.1.005.kath

Su, L. Y. F., Xenos, M. A., Rose, K. M., Wirz, C., Scheufele, D. A., & Brossard, D. (2018). Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets. *New Media and Society*, *20*(10), 3678–3699. https://doi.org/10.1177/1461444818757205

Taboada, M., Brooke, J., & Voll, K. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2).

Tamara, K., & Milićević, N. (2018). Comparing sentiment analysis and document representation methods of Amazon reviews. *SISY 2018 – IEEE 16th International Symposium on Intelligent Systems and Informatics, Proceedings*, 283–288. https://doi.org/10.1109/SISY.2018.8524814

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment in short strength detection informal text. *Journal of the American Society for Information Science and Technology*, *6*(12), 2544–2558. https://doi.org/10.1002/asi.21416

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., & Daelemans, W. (2016). *A dictionary-based approach to racism detection in Dutch social media*. https://doi.org/None

Van Aelst, P., & Vliegenthart, R. (2014). Studying the Tango: An analysis of parliamentary questions and press coverage in the Netherlands. *Journalism Studies*, *15*(4), 392–410. https://doi.org/10.1080/1461670X.2013.831228

Van der Pas, D. (2013). Making hay while the sun shines: Do parties only respond to media attention when the framing is right? *The International Journal of Press/Politics*, *19*(1), 42–65. https://doi.org/10.1177/1940161213508207

Vargo, C. J., Guo, L., Mccombs, M., & Shaw, D. L. (2014). Network issue agendas on Twitter during the 2012 U.S. presidential election. *Journal of Communication*, *64*(2), 296–316. https://doi.org/10.1111/jcom.12089

Vliegenthart, R., Walgrave, S., Baumgartner, F. R., Bevan, S., Breunig, C., Brouard, S., Bonafont, L. C., Grossman, E., Jennings, W., Mortensen, P. B., Palau, A. M., Sciarini, P., & Tresch, A. (2016). Do the media set the parliamentary agenda? A comparative study in seven countries. *European Journal of Political Research*, *55*(2), 283–301. https://doi.org/10.1111/1475-6765.12134

Welbers, K., & Opgenhaffen, M. (2019). Presenting News on Social Media: Media logic in the communication style of newspapers on Facebook. *Digital Journalism*, *7*(1), 45–62. https://doi.org/10.1080/21670811.2018.1493939

Yuan, X., Schuchard, R. J., & Crooks, A. T. (2019). Examining Emergent Communities and Social Bots Within the Polarized Online Vaccination Debate in Twitter. *Social Media + Society*, *5*(3), 205630511986546. https://doi.org/10.1177/2056305119865465