Communication Methods and Measures, 8:190–206, 2014

ISSN: 1931-2458 print/1931-2466 online
DOI: 10.1080/19312458.2014.937527

Routledge
Taylor & Francis Group

Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis

Björn Burscher, Daan Odijk, Rens Vliegenthart, Maarten de Rijke, and Claes H. de Vreese
University of Amsterdam

We explore the application of supervised machine learning (SML) to frame coding. By automating the coding of frames in news, SML facilitates the incorporation of large-scale content analysis into framing research, even if financial resources are scarce. This furthers a more integrated investigation of framing processes conceptually as well as methodologically. We conduct several experiments in which we automate the coding of four generic frames that are operationalised as a set of indicator questions. In doing so, we compare two approaches to modelling the coherence between indicator questions and frames as an SML task. The results of our experiments show that SML is well suited to automate frame coding but that coding performance is dependent on the way SML is implemented.

In most framing studies, news frames are coded with indicator questions in manual Content Analysis (CA) (Matthes, 2009). Generally, measures of several indicators are combined to cover different aspects of a frame (e.g., Simon & Xenos, 2000). Human coders can be properly trained to code frame indicators, and through training their performance can be improved until accuracy and reliability reach satisfactory levels. However, human coding is a time-consuming and costly process. This limits the scope of CA in framing research. Computers, in contrast, are more naturally suited for the processing of large quantities of documents and the repetitiveness of coding frames. Therefore, we introduce a computer-aided method for indicator-based frame coding; this not only decreases the effort required for CA of news frames but also helps in addressing substantial issues in communication research.

The method we apply is based on Supervised Machine Learning (SML) (Sebastiani, 2002), a technique in which a computer learns from a set of human-coded training documents to automatically predict content-analytical variables in texts. By applying SML to the coding of four generic frames, we develop a theory of how the technique should be used to automate CA in future framing studies. We address the following issues: first, we investigate how useful it is to model indicator questions when predicting frames using SML. We compare two approaches. In the first approach, we build a classifier to automatically code frame indicators, which we then aggregate to a frame measure (indicator-based approach). In the second approach, we build a classifier to

Correspondence should be addressed to Björn Burscher, Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Postbus 15793, 1001NG Amsterdam, The Netherlands. E-mail: b.burscher@uva.nl

directly code the presence of a frame (holistic approach). Second, we test the generalizability of SML classifiers by applying them to news sources that were not in the training documents. Third, we investigate the relationship between the number of training documents used and the accuracy of computer-based codings.

We conclude that SML is well suited for frame coding and that it addresses several shortcomings of current approaches to automatic CA. Furthermore, we believe that future framing research can profit from SML theoretically as well as methodologically. SML can promote the incorporation of large-scale CA in framing research by making frame coding much faster and less expensive. This facilitates more integrated studies of framing processes (Matthes, 2012) as well as the analysis of large datasets that have become increasingly available. We discuss extensively the theoretical and methodological implications of our findings for framing research and CA in general.

AUTOMATIC FRAME CODING

According to Gamson and Modigliani (1989), news coverage can be approached as an accumulation of "interpretative packages" in which journalists depict an issue in terms of a "central organising idea," to which Gamson and Modigliani refer as a frame. Frames in news take a central position in framing models (e.g., Scheufele, 1999); they are the dependent variable when studying how frames emerge (frame building) and the independent variable when studying effects of frames on predispositions of the public (frame setting). When detecting frames in news media, CA is the most dominant research technique.

In communication research, various methods are applied to the CA of frames in news (see Matthes, 2009, for an overview). When investigating the framing of news coverage, we distinguish between frame identification and frame coding. While frame identification includes operations aimed at retrieving and defining frames adopted in the news, frame coding is the annotation of frames defined earlier as content analytical variables. Coding a frame requires an operationalization, which enables the methodological assessment of the frame and allows other scholars to reliably study its use across issues, time, and space. Currently, the two most popular frame operationalizations are human coding with indicator questions and dictionary-based computer-aided coding.

Using questions as indicators of news frames in manual CA is the most widely used approach to frame coding. Indicator questions are collected in a codebook and are answered by human coders while reading the text unit to be analyzed (e.g., Simon & Xenos, 2000; Vreese et al., 2001). Each question is designed such that it captures the semantics of a given frame. Generally, several questions are combined to cover various aspects of a frame. Human coding of frames with indicator questions is a reliable but resource-intensive process. As the volume of digitally available media content increases significantly, computer-aided methods become desirable and even a necessity.

Most computer-aided techniques for frame coding follow a dictionary-based approach. In such an approach, previously defined character strings and rules for their combination are used to code text units into content categories (Krippendorff, 2004). In some studies, search strings are used to directly code a frame (Roggeband & Vliegenthart, 2007). In other studies, search strings are used to code a set of predefined concepts (e.g., an issue), and a frame is then

revealed from the co-occurrence of these concepts (Ruigrok & Van Atteveldt, 2007; Shah et al., 2002).

Dictionary-based approaches to frame coding have several disadvantages. First, the researcher herself must manually build the model from which texts are coded into content categories. Therefore, she must design, pre-test, and refine search queries. Not only is this a time-intensive process, but it also may compromise semantic validity. This is because manually compiled classification rules are at risk of being biased by the subjective conceptions and limited domain knowledge of the researcher(s). A search that is too narrow in scope will omit relevant documents (false negatives), while one that is too broad will retrieve unwanted documents (false positives). Supervised Machine Learning (SML) is an alternative approach to computer-aided frame coding that addresses these shortcomings.

When applied to CA, the goal of SML is to automatically code large numbers of text documents into previously defined content categories (see Laver et al., 2003; Durant & Smith, 2007). Basically, the computer tries to replicate the coding decisions of humans. A precondition for the application of SML is a set of documents that are already coded for the content categories of interest. We call this the training set. SML involves three steps: First, text documents from the training set are converted so that they are accessible for computational analysis. Each document is represented as a vector of quantifiable text elements (e.g., word counts) that are called features. Second, feature vectors of all documents in the training set, together with the documents' content labels (e.g., the presence of a frame), are used to train a classifier to automatically code the content categories. In doing so, a supervised machine-learning algorithm statistically analyses features of documents from each content category and generates a predictive model to classify future documents according to the content categories. Finally, the classifier is used to code text documents outside the training set. For a detailed introduction to SML we refer to Russell and Norvig (2002) or Grimmer and Stewart (2013).

Using SML to automate the coding of frames is an improvement compared to dictionary-based methods (Hillard et al., 2008). In SML, in contrast to dictionary-based approaches, a computer automatically estimates a model that classifies texts according to content categories. This is not only more efficient but also likely to be more effective because the rules used to detect frames are based on a statistical analysis of human-coded training data. Furthermore, because manually coded material is available, one can systematically assess the accuracy of computer-based annotations.

Additionally, SML is valuable to future framing research more generally. First, it makes CA of frames more feasible. Once a classifier is trained to code a frame, it can be effortlessly employed for real-time CA of that frame. This not only leads to savings in time and costs but also promotes integrating (large-scale) CA with experimental as well as survey research.

Furthermore, SML enables scholars to easily increase the scope of framing analysis. Comprehensive CA of mass media allows investigation of news framing and its effects over the long term and also allows more nuanced, conditional and comparative research. This is relevant because more and more media content is becoming available digitally.

Finally, because one can directly study the entire population of texts, an SML approach can decrease the risk of committing sampling errors and prevent problems related to statistical accuracy as a result of limited samples.

RESEARCH QUESTIONS

We apply SML to the coding of four generic news frames: conflict frame, economic consequences frame, human-interest, and morality (Semetko & Valkenburg, 2000).[1] In doing so, we study the following questions.

First, we empirically investigate the question of the extent to which an SML approach is suitable for automatic coding of indicator-based news frames. Second, we investigate how we should model indicator-based frame coding as a SML task. When using machine-learning techniques to tackle methodological challenges in social science research, it is important to tailor its implementation to the specific research problem at hand. We test whether teaching the computer to code a frame directly (holistic approach) is more effective than teaching it to code a set of indicator questions from which the frame is derived by means of aggregation (indicator-based approach). Both approaches are described in detail in the following section.

Third, we investigate the generalizability of SML classifiers. The goal of automating frame coding is to be able to easily code large amounts of data from several sources. Therefore, we are interested in the question of whether our models are able to correctly predict the four frames in articles from news media not included in the training data.

Finally, we study the relationship between the amount of training data used to build a classifier and its performance to predict frames. Because manually coded training data are expensive and labor-intensive to obtain, it is important to know how much training data one needs to build a well-performing frame classifier. We expect that increasing the number of news articles in the training set leads to an increase in coding performance.

HOLISTIC VERSUS INDICATOR-BASED FRAME CODING

This study aims to increase our understanding of how SML should be used to effectively master CA problems in communication research. SML is a set of algorithms and approaches for automatic classification. Finding the optimal way of performing a specific classification task generally involves comparing various models. Previous studies have compared the performance of different SML algorithms (Joachims, 1998; Pang et al., 2002), feature types (Scharkow, 2013; Alm et al., 2005), feature selection mechanisms (Forman, 2003; Hillard et al., 2008), and validation techniques (Joachims, 2012). We differentiate between predicting frames directly and predicting them via indicators because we expect this particular modification to impact the performance of indicator-based frame coding. This is relevant because we want to automatically code frames in news as accurately as possible.

Before presenting details of the approaches, we first define some basic concepts. We have a collection of news articles D and a set of frames U, each of which is operationalised as a set of indicator questions V. When applying SML, we predict the probability $P(U_m|d)$ that a frame $u_m \in U$ is present in an article $d \in D$. For this task we build a classifier on the basis of a training set of news articles that humans have coded for each indicator question $v \in V$.

We try to resemble the manual coding process in the indicator-based approach. First, we train a set of classifiers to predict the answer to each indicator question. In formal terms, we estimate

[1] All four frames are introduced in detail in the next section.

$P\left(\hat{v}_n|d\right)$ for each indicator question $v_n \in V$ of the frame. As in manual CA, we then combine the predicted answers to the indicator questions into a single frame measure. We thus derive the probability $P\left(u_m|d\right)$ for each frame $u_m \in U$ from $P\left(\hat{u}_m|\hat{v}_1, \ldots, \hat{v}_N\right)$ for all indicator questions $v_m \in V$. Answers to indicator questions can be combined in various ways. In our case, we argue that all questions indicate presence of the frame by focusing on different but equally important aspects of it (Semetko & Valkenburg, 2000). Therefore, we claim the frame to be present when at least one of the indicators is coded "yes."

In the holistic approach we do not train classifiers to predict indicator questions. Rather, we try to predict the presence of frames directly. First, for each frame we aggregate coded indicator questions in the training data to a single frame measure. Again, a frame is considered present if at least one of the indicators is coded positive. Second, we use the resulting frame-level codings as training data to train a classifier for each frame that can predict the presence of the frame. Formally, we train a classifier to estimate $P\left(u_m|d\right)$ for each frame $u_m \in U$. In contrast to the former approach, here we completely ignore indicator-level codings in the SML process, but train our classifiers directly on frame-level codings.

Why exactly do we expect performance differences between the two approaches? This question brings us to the role of indicators in manual frame coding. Indicators are a means of measuring theoretical concepts in texts. In our case, they help coders to decide on the presence or absence of a frame aspect, from which we infer whether the frame is present. An SML algorithm, in contrast, bases its decision on a systematic statistical analysis of the vocabulary of the text. This leads to a complex model in which each unique word is associated with a probability of the text containing the frame. That is, while a human coder relies on a small set of questions as indicators, the computer relies on the presence of each word from the document collection as an indicator.

Therefore, we expect that the holistic approach might provide a better model to predict the frame variable. It is likely that text features include variables, which explain variation in the frame variables very well but do not explain variation in indicators. In the indicator-based approach, the predictive power of such unknown variables is not considered when predicting the frame, because the frame measure is based on the indicators only. In contrast, in the holistic approach, such variables are included in the model to predict the frame.

CLASSIFIERS AND DOCUMENT REPRESENTATION

To test these SML approaches, we need to train classifiers for predicting indicator questions and frames. In doing so, we must choose a supervised machine-learning algorithm. As we code different frames with several indicators each, the applied SML algorithm must deal with considerable variation in content characteristics. Consequently, one would expect different SML algorithms to perform better, depending on the frame and indicators considered. Therefore, we propose an approach in which we combine the strengths of various SML algorithms (Dietterich, 2000; Hillard et al., 2008; Polikar, 2012). The resulting combination of different algorithms is called an *ensemble of classifiers.*

Ensemble classifiers can be constructed in different ways. We applied a technique called stacked generalization, which involves training a learning algorithm to combine the predictions of several other learning algorithms. To do this, we first partitioned the data into a held-in and a

held-out set. We then trained each learning algorithm on the held-in set, and obtained a vector of predictions for the held-out set. Each element of the vector corresponded to a prediction of one of the individual algorithms. Next, we learned how to combine these predictions. We trained a logistic regression model with the individual classifiers' predictions of the held-out set as input, and the correct responses as output.[2] This way of combining predictions of various classifiers into a final predictive model is intended to be flexible in addressing the different complex characteristics of each of the frame-coding tasks.

In the ensemble we combined two different Linear Support Vector Machines (SVM) (Joachims, 1998), a Polynomial SVM classifier (Chang et al., 2010), and a Perceptron algorithm (Lippmann, 1987). In Appendix A we report on the relative performance of these algorithms. This helps to understand the merits of combining several learning algorithms compared with relying on a single one.

To train classifiers and apply them to frame coding, the content of each news article must be represented quantitatively as a vector of document features. Such features are variables containing quantified information about an article that is relevant to the coding task. Selecting relevant features has a significant impact on the ability of the SML algorithms to compute a good predictive model and therefore influences coding performance when predicting the presence of frames in future news articles (Sebastiani, 2002).

When selecting document features for our frame coding task, we thus need to confront the question of which elements of a news article constitute a frame. According to Entman (1993, p. 52), news frames manifest themselves in certain text attributes as "the presence or absence of certain keywords, stock phrases, (and) stereotyped images (. . .)." Therefore, we assume that it is appropriate to represent each article as a listing of the words it contains. This is referred to as the "bag-of-words" approach and has been shown to be effective in various text classification tasks (Joachims, 1998; Sebastiani, 2002). Strictly speaking, we represent each article as a vector of TF.IDF weights (Russell & Norvig, 2002). This means that each word is assigned the number of times it occurs in a document (TF) and is weighted by the inversed frequency of articles in the entire collection containing the word (IDF). The idea behind TF.IDF weighting is to evaluate the power of a word to discriminate between articles. Rare words are assumed to be more discriminating and therefore are assigned higher weight.

Formally, each article $d \in D$ is represented as a vector V containing a TF.IDF weight W for each unique word $t \in T$ in the collection of articles, $V_d = (W_{d1}, W_{d2}, \ldots, W_{dN})$. The TF.IDF weight for each word in an article is computed as follows: $W_{TD} = TF_{TD} * IDF_t = TF_{TD} * log\left(\frac{N}{n_t}\right)$, where N is the total number of articles in the collection, and N_t is the number of articles in the collection that contain word t.[3]

We used the Scikit-Learn machine learning toolkit (Pedregosa et al., 2011) for computing feature representations of documents. For training and testing classification models, we used the

[2]Instead of a single split into held-in and held-out, the vectors of predictions are obtained through 10-fold cross-validation.

[3]We also tried alternative bag-of-words transformations, for example, binary-word presence, word counts, and parsimonious language models (Hiemstra et al., 2004). Additionally, we tried representing all articles in terms of n-grams and latent topics as derived from a LDA-model (Radim & Sojka, 2010). These variations in feature representation, as well as combinations of them, did not improve on TF.IDF weighting. We suggest applying syntactic (e.g., part of speech tags) or semantic features in future research.

Orange Data Mining Toolbox (Demšar et al., 2013). Both libraries are general-purpose machine-learning modules for the Python programming language. Python, Scikit-learn, and Orange are open-source applications and are therefore free[4].

FOUR GENERIC NEWS FRAMES

We apply SML to the coding of four generic news frames. These are the conflict frame, the economic consequences frame, the human-interest frame and the morality frame. The conflict frame highlights conflict between individuals, groups or institutions. Prior research has shown that the depiction of conflict is common in political news coverage (Neuman et al., 1992; Semetko & Valkenburg, 2000) and that it has inherent news value (Galtung & Ruge, 1965; Eilders, 1997; McManus, 1994; Staab, 1990). Furthermore, several scholars have observed an increase in the portrayal of conflict in political reporting (Patterson, 1993; Blumler et al., 1995; Cappella & Jamieson, 1997; Vliegenthart et al., 2011). Within the field of political communication, the conflict frame is often employed in empirical research (e.g., Schuck et al., 2011; Vliegenthart et al., 2008).

By emphasising individual examples in the illustration of issues, the human-interest frame adds a human face to news coverage. According to Iyengar (1991), news coverage can be framed in a thematic manner, taking a macro perspective, or in an episodic manner, focusing on the role of the individual affected by an issue. Such use of exemplars in news coverage has been observed by several scholars (Semetko & Valkenburg, 2000; Neuman et al., 1992; Zillmann & Brosius, 2000) and connects to research on personalisation of political news (Iyengar, 1991).

Economic consequence framing approaches an event in terms of its economic impact on individuals, groups, countries or institutions. Covering an event with respect to its economic consequences has been argued to possess high news value (Graber, 1993; McManus, 1994) and to increase the event's pertinence among the audience (Gamson, 1992).

The morality frame puts moral prescriptions or moral tenets central when discussing an issue or event. Morality as a news frame has been the subject of several studies and is used in the context of various issues, such as gay rights (Nisbet & Huge, 2006; Nisbet et al., 2003) and biotechnology (Brewer, 2002, 2003).

We have chosen generic news frames because generic frames, as opposed to issue-specific frames, are topic-independent. This enables us to test our SML approaches with semantically distinct frames while using the same dataset. Consequently, our findings are not limited to frames and news coverage concerning one topic.

DATA

Our data consist of front-page news articles of three national Dutch daily newspapers (*De Volkskrant, NRC Handelsblad,* and *De Telegraaf*) between 1995 and 2011. All items were collected digitally via the Dutch Lexis-Nexis database. For each year, a stratified sample (13%) of

[4]The Python code used can be provided upon request.

news articles was manually coded for references to politics[5] and the presence of the four frames. Only those articles that were coded positive for references to politics were coded for the presence of the four frames. The unit of coding was the distinct news story. To measure the extent to which the four frames appeared in stories that mention politics, we used a series of 11 questions to which the coder was required to answer yes or no.[6] See Table 1 for the question wordings of all indicators[7] used. Frame codings were constructed by aggregating measures of indicator questions such that a frame was considered present when at least one of its indicators had been coded positive.

Manual coding was conducted by a total of 30 trained coders at the University of Amsterdam. All coders were native speakers of the Dutch language and received extensive training. To assess inter-coder reliability, political news articles from a random subset (N=156) were each[8] coded by two coders. We report Krippendorff's Alpha as well as pairwise agreement (in parentheses) for all frames: conflict frame = .51 (.77), morality frame = .21 (.85), economic consequences frame = .58 (.82), and human-interest frame = .29 (.64). See Table 1 for reliability measures for individual frame indicators[9]. We stress that inter-coder reliability is not optimal. Performance of the classifiers likely suffers from imperfect training data, but we consider it unlikely that this

TABLE 1
Question Wording and Inter-coder Reliability of Frame Indicators

Item	*Wording*	*Kr. Alpha*
C1	Does the item reflect disagreement between parties, individuals, groups or countries?	.47 (.72)
C2	Does the item refer to two sides or more than two sides of the problem?	.41 (.70)
E1	Is there a reference to the financial costs/degree of expense involved, or to financial losses or gains?	.61 (.83)
E2	Is there a reference to economic consequences of pursuing or not pursuing a course of action?	.37 (.85)
H1	Does the issue provide a human example or a human face to the issue?	.20 (.75)
H2	Does the item employ adjectives or personal vignettes that generate feelings of outrage empathy or caring?	.33 (.57)
H3	Does the item mention how individuals and groups are affected by the issue or problem?	.16 (.84)
M1	Does the item contain any moral message?	.35 (.91)
M2	Does the item make references to morality, God or other religious tenets?	.43 (.91)
M3	Does the item offer specific social prescriptions about how to behave?	.29 (.92)

Note. Krippendorff's alpha is reported as main measure of inter-coder reliability. Percentage agreement is reported in parentheses.

[5]Coders were required to answer 'yes' or 'no' to the following question:"Is the story political in nature?"

[6]In previous research, these questions have been shown to be reliable indicators of the four frames (e.g., Semetko & Valkenburg, 2000; Vreese et al., 2001).

[7]We performed a principal component analysis with non-orthogonal rotation to establish the coherence of the indicator questions and their relationships to the frames. As expected, we found a four-factor solution in which all indicators show significant positive loadings (>.5) on the expected frame.

[8]Nearly all coders were involved, because multiple pairs of coders were used for reliability testing.

[9]It is a well-known issue that Krippendorff's alpha measures tend to be relatively low when assessing inter-coder agreement of binary classification tasks with unbalanced class distributions. This especially is the case with the morality frame, where we observe a substantial difference between the pairwise agreement measure and Krippendorff's alpha measure.

biases the conclusions of our study. In the Discussion we elaborate on how the quality of the training data influences our findings and conclusions.

Coding was performed for a large-scale research project on the influence of media coverage on parliamentarians.[10] The final dataset consisted of 11,074 documents, of which 6,030 were political in nature. We used this set of manually coded articles to train and test our classifiers.

EVALUATION METRICS AND CROSS VALIDATION

We evaluated coding performance in terms of classification accuracy, receiver operating characteristics and Krippendorff's Alpha. Performance measures are reported for the automatic coding of indicators and frames. Accuracy (AC) is the percentage of agreement between human classifications and computer-based classifications. It indicates the number of correctly classified documents. To demonstrate a classifier's improvement over chance agreement, we compared the reported accuracy measures with a random baseline. The random baseline is a naive way of predicting the presence of an indicator or frame by chance. It randomly chooses the answer to an indicator question or whether a frame is present or not, taking into account only its prevalence in the training set. This baseline thus randomly assigns a classification without considering the document content, with a probability based only on the class distributions. Consequently, it will be more likely to randomly pick the majority class than the minority class. The classifier's accuracy improvement over the random baseline indicates its superiority to chance agreement.

Furthermore, we rely on receiver operating characteristics to evaluate classifier performance. More precisely, we report the area under the curve (AUC). AUC is a measure of how well a classifier discriminates between the presence and the absence of a frame or indicator. AUC is a commonly used evaluation method for binary choice problems (Sokolava & Laplame, 2009) that involve classifying an instance as either positive or negative. The main advantage over other evaluation methods is its insensitivity to unbalanced datasets. The AUC measure is based on the ROC curve, which is a graphical depiction showing the trade-off between increasing true positive rates and increasing false positive rates as the discrimination threshold of the classifier is varied. The AUC distils this information into a single scalar by expressing the probability that the classifier will rank a positive document above a negative document. A perfect model will score an AUC of 1, while random guessing will score an AUC of approximately 0.5. The measure thus allows us to quantify how much better than random the classifier's choices are.

Additionally, we report Krippendorff's Alpha (KA), which is a common inter-coder agreement statistic in the field of communication science. Like the AUC measure, Krippendorff's Alpha corrects for agreement by chance.

Ten-fold cross-validation was used to obtain evaluation measures of classification performance. The dataset was partitioned into ten equal parts, one of which was reserved for testing the classifier (test set). The remaining parts were used as training data (training set). We repeated this cross-validation process ten times, such that each subsample was used once as the test set. The results from all validation rounds were averaged to produce a single estimation. This

[10]The research is supported through a VENI grant from the Dutch Science Foundation, and the Dutch national program COMMIT.

way, all observations were used for training as well as evaluation of the classifiers, but training observations were always separated from the test set.[11]

To test the generalizability of our classification models, as described in the third research question, we trained classifiers on articles from two of the three available newspapers and then evaluated the classifiers' abilities to correctly code frames in articles from the third paper,[12] which were not included in the training set. We performed this test for all possible combinations of the three newspapers.

Finally, to assess the relationship between the number of training documents and classification performance, we repeatedly trained each frame classifier while increasing the number of documents in the training set. We held out a fixed set of 1,000 articles for testing. For training, we used samples of different sizes from the held-in set. In total, we performed seven iterations with the following numbers of documents in the training set: 100, 200, 500, 1,000, 2,000, 3,000, and 4,000.

RESULTS

To answer our research questions, we conducted a series of classification experiments in which we predicted four frames and their indicators. In Table 2, we report classification performance (AC, AUC and KA) per frame for the holistic and indicator-based approaches. In Table 3, we report classification performance for all indicators. Both tables include measures of the random baseline.

First, we address the random baseline. This baseline indicates agreement by chance in the classification process, based on the prevalence of frames in the training set. We observe a high variation in frame prevalence (M=41%, SD=23.01), with morality being the least prevalent frame (13%) and conflict the most prevalent frame (61%). Derived probabilities of correctly predicting the frames by chance range from .61 for the conflict frame to .87 for the morality frames (M=.69, SD=.13).

TABLE 2
Classification Performance of Frames

	Conflict			*Economic Consequences*			*Human Interest*			*Morality*		
Prevalence	*61%*			*32%*			*59%*			*13%*		
	AC	AUC	KA	AC	AUC	KA	AC	AUC	KA	AC	AUC	KA
Baseline	.61	.50		.68	.50		.59	.50		.87	.50	
Indicator	.77	.76	.52	.85	.84	.67	.74	.74	.47	.89	.62	.33
Holistic	.80	.78	.57	.89	.85	.71	.79	.78	.55	.96	.76	.63

[11]Please note that the cross-validation sample that was used to estimate weights for the ensemble of classifiers is nested in the cross-validation sample, which we used to assess coding performance.

[12]We always used a random sample of 2,000 articles as a training set and a random sample of 1,000 articles as test set.

TABLE 3
Classification Performance of Frame Indicators

	C1	*C2*	*E1*	*E2*	*H1*	*H2*	*H3*	*M1*	*M2*	*M3*
Prevalence	*52%*	*69%*	*29%*	*16%*	*21%*	*49%*	*9%*	*7%*	*6%*	*5%*
Baseline	.52	.69	.71	.84	.79	.51	.91	.93	.94	.95
CA	.77	.75	.87	.86	.82	.76	.93	.94	.96	.95
AUC	.77	.75	.85	.78	.76	.76	.64	.59	.69	.50
KA	.54	.49	.68	.50	.49	.52	.39	.26	.49	.02

Second, we turn to measures of classification performance. Accuracy (AC) and AUC scores indicate high coding performance for all four frames. Therefore, we conclude that SML is suitable for frame coding. When applying the indicator-based approach, classification accuracy ranges from .74 for the human interest frame to .89 for the morality frame (M=.81, SD=.07). When applying the holistic approach, accuracy ranges from .79 for the human interest frame to .96 for the morality frame (M=.86, SD=.08). All accuracy scores surpass the random baseline, meaning that we improve on chance agreement for each frame. Moreover, for all frames, the holistic approach outperforms the indicator-based approach in terms of classification accuracy, AUC and Krippendorff's Alpha (KA) measures. The average improvement in accuracy is about five percentage points. Therefore, we conclude that it is more effective to predict the frame variable directly, compared to predicting indicators and aggregating them afterward.

Third, we find performance differences between frames. When applying the holistic approach, AUC scores range from .76 for the morality frame to .85 for the economic consequences frame (M=.86, SD=.04). This means that, among all frames, our classifiers can most optimally differentiate between positive and negative examples of the economic consequences frame. Among the other three frames, we find little variation in AUC scores.[13]

Fourth, we investigated whether we could generalise our models to news sources that were not included in the training data. In Table 4, we report classification accuracy when training on data from two of the three newspapers and testing on articles from the third paper. The results indicate that we can generalize our classification models to other news sources. However, in most cases, classification accuracy was slightly lower compared with predicting frames in sources that were included in the training data (see Table 2).

Finally, we present findings of experiments regarding the relationship between the amount of training data and coding performance. For all frames, classification accuracy is plotted in Figure 1. As expected, measures show that increasing the number of training documents leads to increased classification performance for all classifiers. It is obvious immediately that compared to the other frames, classification accuracy of the morality frame increases more slowly when adding training documents. Most likely, this is because the morality frame is less prevalent in the training data. However, it stands out that classification accuracy for the economic consequences frame increases fastest when adding training documents, although it is not the most prevalent frame.

[13]We found the same pattern when applying the indicator-based approach.

TABLE 4
Classification Accuracy of Frames in Sources Outside the Training Set

	VK/NRC →Tel	*VK/TEL →NRC*	*NRC/TEL →VK*
Conflict	.69	.74	.75
Economic Cons.	.88	.86	.86
Human Interest	.69	.71	.67
Morality	.97	.90	.89

Note. VK = Volkskrant, NRC = NRC/Handelsblad, TEL = Telegraaf

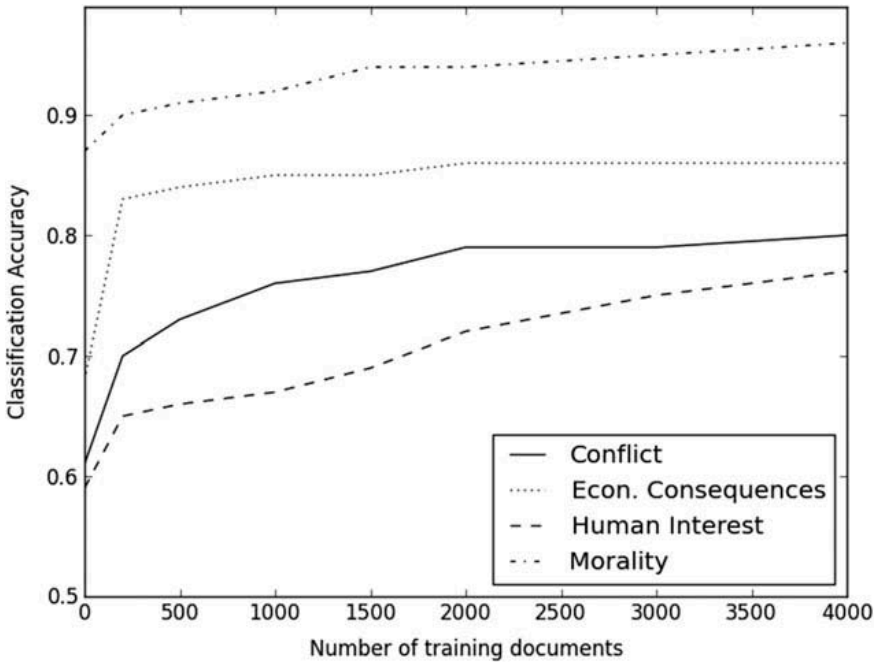

FIGURE 1 Relationship between classification accuracy and number of training documents.

This supports our finding that the SML approach works better for the economic consequences frame than for the other three frames.

DISCUSSION

In this article we explored the application of SML to frame coding. Framing is one of the key concepts in communication science, and SML can advance future framing research by easing large-scale CA. Once a classifier is trained to code a frame, it can be employed for

automatic coding of that frame in subsequent studies. Therefore, SML facilitates more integrated analyses of framing processes (Matthes, 2012; De Vreese, 2005). Several scholars have advocated studying framing processes outside the laboratory (Kinder, 2007). However, sophisticated designs, combining (panel) surveys and CA (e.g., Schuck et al., 2013; Wettstein, 2012), are expensive. SML-based frame coding not only facilitates the application of such mixed-methods designs but also allows the scale of its CA part to be easily increased. Large-scale CA, which becomes more and more attractive as the amount of digitally available media content rises (Lazer et al., 2009), helps address substantial issues in framing research. Such issues include looking at frame variation over time (Matthes & Schemer, 2012; Chong & Druckman, 2010) and the conditionality of framing processes (e.g., Chong & Druckman, 2007). To what extent is a frame repeated or challenged in the media, and how does this affect the public over time? To what extent do frame usage and framing effects depend on the topic of a message, the actors with whom frames are associated in that message, and the medium used to transmit the message? Appropriate investigation of these questions requires frame coding over a long period and across various domains, respectively. SML can help the affordability of such CA without relying on small samples.

In this study, we applied SML to the coding of four widely used journalistic frames. We observed high levels of coding performance for all four frames. Using our classifiers, we can now automatically code these frames in future studies. We conclude that SML is generally suited to automate frame coding. When investigating a new frame in future studies, manual coding can be limited to that needed for training a classifier, and the remaining documents can be coded by applying the classifier. Performance levels of SML-based CA in our study are comparable to similar attempts of employing SML to automate the coding of concepts that are relevant to communication research (see, e.g., Scharkow, 2013, for SML-based coding of news values).

Our study informs the application of SML to frame coding and CA more generally in several ways. First, we conclude that SML approaches might work even if one does not possess tens of thousands of training documents, which were available in previous studies applying SML to CA (e.g., Hillard et al., 2008). In this study, the amount of training data necessary to train a well-performing classifier varies from frame to frame. One important factor is the overall presence of a frame. When studying a frame that occurs regularly within the text corpus used, manual coding of a few hundred documents might be sufficient to automate coding of the remaining documents. When studying an uncommon frame, active sampling (Tong & Koller, 2000) of positive examples of the frame can help keep manual coding efforts manageable. Several strategies for this are discussed in the literature (e.g., Hillard et al., 2008). Furthermore, we conclude that some concepts are less difficult to predict than others. We found, for example, that classification performance of the economic consequences frame improves the most when increasing the size of the training set, although this frame is not the most prevalent one.

Second, we conclude that a trained classifier can be applied to automatic coding in sources other than those used for training. We provide evidence for this but also find that classification accuracy decreases for some frames. We believe the generalizability of a classifier strongly depends on the coding task and the training data used. Therefore, in future studies, similar experiments should be repeated (e.g., generalization from print to online media).

We might extrapolate these conclusions to several other concepts in communication research. This includes the coding of such concepts as sentiment, emotions, or news values, which have some conceptual similarity with frames. We recommend testing all of this in future research.

In this article, we also compared two approaches, indicator-based and holistic, to modeling the frame coding process. When applying SML it might seem appropriate to proceed as in manual CA, where we code indicators and aggregate them to frame measures. However, results of our experiments show that it is more effective to train a classifier to predict the presence of a frame directly. In regard to generalising this finding, we would like to mention some limitations. It is difficult to say whether performance differences would be similar with other frames or even other concepts because the pattern we found might be due to properties of the data we used and the variables we coded. We compared the approaches when using a binary frame measure. When combining indicators in such a way that one gets a continuous outcome measure (e.g., by averaging them), the holistic approach might not outperform the indicator-based approach. Predicting the strength of a frame (or other concept) in a text is most likely more complicated than simply predicting its presence. Therefore, explicitly modelling indicators in the SML process might be of greater relevance.

The fact that we find the same pattern for all studied frames, which are substantially different, gives us some confidence in the generalizability of the finding. However, future studies are needed to test this. At least we can make the following argument: In some cases it works better to predict frames directly. Although we cannot establish clear rules about when this is the case based on our findings, it is worth comparing both approaches when trying to automate a coding task using SML. In future research, similar comparisons should be made using other datasets and frames.

Another limitation of our study is that we tested the SML approach with generic frames only. We believe that it would work similarly for other types of frames, such as issue-specific frames (e.g., Rhee, 1997). The critical difference between generic frames and issue frames is that the former are used more widely and have little issue dependence. There is no reason, however, to believe that it would not work with issue-specific frames, because we expect them to be manifested in a certain vocabulary as well. One might even expect better performance, because an issue frame might be more salient in an article than a generic frame. Moreover, with issue-specific frames, the population of texts to analyse is more uniform, which might decrease the complexity of the classification problem. Then again, it might be difficult to generate good training data, as one must deal with a limited population of texts containing the issue frame.

Another question is whether SML can be applied to more complex frames. Among the frames studied, we believe the morality frame to be the most complex. Because we are able to automatically code the morality frame with performance similar to the other three frames, we believe that an SML approach generally works with more complex frames. More advanced feature representations are likely to increase performance when coding complex frames. We leave this question for future research.

Finally, an important limitation of our study concerns inter-coder reliability. First, we are aware that we should have coded each article by more than two coders when assessing reliability. Second, the quality of our training data is not optimal. In various cases, coders disagreed on the presence of frames, as indicated by the reported reliability measures. Disagreement likely results from a combination of unsystematic coding errors and systematically different interpretation of frame indicators across coders. The most relevant question is how the latter, especially, might influence our findings and conclusions. We expect classification performance to decrease as a result of inconsistencies in the training data. If texts with similar features are associated with different labels, it becomes more difficult for the SML algorithm to estimate a model that can clearly differentiate between two classes.

Although classification performance is most likely influenced by the moderate training data, we believe our conclusion to be largely unaffected. We conclude that SML is suited for automating frame coding, but the more error-prone the training data are, the more error-prone the automatic classifications. Moreover, our conclusion that trained classification models might fit texts from sources not included in the training data is unlikely to be affected. There is no reason to believe that models would be less generalizable if inter-coder agreement were higher. Despite those shortcomings, this paper is the first to apply SML to frame coding. Our study not only provides promising results but also provides important insights regarding the use of SML in future communication research.

REFERENCES

Alm, C.O., Roth, D., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Association for Computational Linguistics Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 579–586). Vancouver, Canada: ACL.

Blumler, J. G., Blumler, J., & Gurevitch, M. (1995). *The crisis of public communication*. London, UK: Routledge.

Brewer, P. R. (2002). Framing, value words and citizens' explanations of their issue opinions. *Political Communication*, *19*(3), 303–316.

Brewer, P. R. (2003). Values, political knowledge, and public opinion about gay rights: A framing-based account. *Public Opinion Quarterly*, *67*(2), 173–201.

Cappella, J. N., & Jamieson, K. H. (1997). *Spiral of cynicism: The press and the public good*. New York, NY: Oxford University Press.

Chong, D., & Druckman, J. N. (2007). A theory of framing in competitive elite environments. *Journal of Communication*, *57*(1), 99–118.

Chong, D., & Druckman, J. N. (2010). Dynamic public opinion: Communication effects over time. *American Political Science Review*, *104*(4), 663–680.

De Vreese, C., Peter, J., & Semetko, H. A. (2001). Framing politics at the launch of the euro: A cross-nation comparative study on frames in the news. *Political Communication*, *18*(2), 107–122.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roll (Eds.), *Multiple classifier systems. Lecture notes in computer science, Vol. 1857* (pp. 1–15). Berlin, Germany: Springer.

Durant, K. T., & Smith, M. D. (2007). Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. In B. Massand (Ed.), *Advances in web mining and web usage analysis. Lecture notes in computer science, Vol. 4811* (pp. 187–206). Berlin, Germany: Springer.

Eilders, C. (1997). *Nachrichtenfaktoren und Rezeption*. Opladen, Germany: Westdeutscher Verlag.

Entman, R. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51–55.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

Galtung, J., & Ruge, M. H. (1965). The structure of foreign news. *Journal of Peace Research*, *2*(1), 64–90.

Gamson, W. (1992). *Talking politics*. New York, NY: Cambridge University Press.

Gamson, W., & Modigliani, A. (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology*, *95*(2), 1–37.

Graber, D. (1993). *Mass media and American politics*. Washington, DC: CQ Press.

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.

Hiemstra, D., Robertson, S., & Zaragoza, H. (2004). Parsimonious language models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 178–185). New York, NY: ACM.

Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, *4*(4), 31–46.

Iyengar, S. (1991). *Is anyone responsible? How television frames political issues*. Chicago, IL: University of Chicago Press.

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features*. Berlin, Germany: Springer.

Joachims, T. (2012). *Learning to classify text using support vector machines: Methods, theory, and algorithms*. Dordrecht, The Netherlands: Kluwer Academic.

Kinder, D. R. (2007). Curmudgeonly advice. *Journal of Communication*, *57*(1), 155–162.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, *97*(2), 311–331.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., . . . Gutmann, M. (2009). Life in the network: The coming age of computational social science. *Science*, *323*(5915), 721.

Lippmann, R. P. (1987). An introduction to computing with neural nets. *ASSP Magazine*, *4*(2), 4–22.

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

Matthes, J. (2012). Framing politics: An integrative approach. *American Behavioral Scientist*, *56*(3), 247–259.

Matthes, J., & Schemer, C. (2012). Diachronic framing effects in competitive opinion environments. *Political Communication*, *29*(3), 319–339.

McManus, J. (1994). *Market-driven journalism: Let the citizen beware?* Thousand Oaks, CA: Sage.

Neuman, W. R., Just, M. R., & Crigler, A. N. (1992). *Common knowledge: News and the construction of political meaning*. Chicago, IL: University of Chicago Press.

Nisbet, M. C., Brossard, D., & Kroepsch, A. (2003). Framing the stem cell controversy in an age of politics. *The International Journal of Press/Politics*, *8*(2), 36–70.

Nisbet, M. C., & Huge, M. (2006). Attention cycles and frames in the plant biotechnology debate managing power and participation through the press/policy connection. *The Harvard International Journal of Press/Politics*, *11*(2), 3–40.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Association for Computational Linguistics Conference on Empirical Methods in Natural Language Processing* (pp. 79–86). New York, NY: ACM.

Patterson, T. E. (1993). *Out of order*. New York, NY: Vintage.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., & Thirion, B. (2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Polikar, R. (2012). Ensemble learning. In C. Zhang & Y. Ma (Eds.), *Ensemble machine learning* (pp. 1–34). Berlin, Germany: Springer.

Radim, R., & Sojka, P. (2010). Software framework for topic modeling with large corpora. In *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Rhee, J. W. (1997). Strategy and issue frames in election campaign coverage: A social cognitive account of framing effects. *Journal of Communication*, *47*(3), 26–48.

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

Ruigrok, N., & van Atteveldt, W. (2007). Global angling with a local angle: How British and Dutch newspapers frame global terrorist attacks. *The Harvard International Journal of Press/Politics*, *12*(1), 68–90.

Russell, S., & Norvig, P. (2002). *Artificial Intelligence: A modern approach*. Upper Saddle River, NJ: Prentice Hall.

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity*, *47*(2), 761–773.

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of Communication*, *49*(1), 103–122.

Schuck, A., Boomgaarden, H., & de Vreese, C. (2013). Cynics all around? The impact of election news on political cynicism in comparative perspective. *Journal of Communication*, *63*(2), 287–311.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47.

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.

Shah, D. V., Watts, M. D., Domke, D., & Fan, D. P. (2002). News framing and cueing of issue regimes: Explaining Clinton's public approval in spite of scandal. *Public Opinion Quarterly*, *66*(3), 339–370.

Simon, A., & Xenos, M. (2000). Media framing and effective public deliberation. *Political Communication*, *17*(4), 363–376.

Staab, J. F. (1990). *Nachrichtenwert-theorie: Formale Struktur und empirischer Gehalt*. Freiburg, Germany: Alber.

Tong, S., & Koller, D. (2000). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, *2*, 45–66.

Van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, N., & Schlobach, S. (2008). Good news or bad news? Conducting sentiment analysis on Dutch text to distinguish between positive and negative relations. *Journal of Information Technology & Politics*, *5*(1), 73–94.

Vliegenthart, R., Boomgaarden, H. G., & Boumans, J. W. (2011). *Changes in political communication: Personalization, conflict and negativity in British and Dutch newspapers*. London, UK: Palgrave Macmillan.

Vliegenthart, R., Schuck, A. R. T., Boomgaarden, H. G., & de Vreese, C. (2008). News coverage and support for European integration, 1990-2006. *International Journal of Public Opinion Research*, *20*(4), 415–439.

Wettstein, M. (2012). Frame adoption in referendum campaigns: The effect of news coverage on the public salience of issue interpretations. *American Behavioral Scientist*, *56*(3), 318–333.

Zillmann, D., & Brosius, H. B. (2000). *Exemplification in communication*. Mahwah, NJ: Erlbaum.

APPENDIX A. INDIVIDUAL PERFORMANCE OF CLASSIFIERS IN ENSEMBLE

Below we show classification accuracy for each individual classifier in the ensemble when directly predicting frames (holistic approach). For all frames, we find clear performance differences between classifiers. We conclude that combining several algorithms helps us address the different characteristics of each of the frame coding tasks. The ensemble classifier performs as well as the best individual classifier in all cases, but does not improve on it.

	Conflict	*Economic Consequences*	*Human Interest*	*Morality*
Linear SVM 1	.79	.87	.79	.87
Linear SVM 2	.75	.85	.74	.96
Polynomial SVM	.57	.76	.58	.88
Perceptron	.73	.89	.75	.90
Ensemble	.80	.89	.79	.96