Qual Quant (2013) 47:761–773
DOI 10.1007/s11135-011-9545-7

Thematic content analysis using supervised machine learning: An empirical evaluation using German online news

Michael Scharkow

Published online: 28 July 2011

Abstract In recent years, two approaches to automatic content analysis have been introduced in the social sciences: semantic network analysis and supervised text classification. We argue that, although less linguistically sophisticated than semantic parsing techniques, statistical machine learning offers many advantages for applied communication research. By using manually coded material for training, supervised classification seamlessly bridges the gap between traditional and automatic content analysis. In this paper, we briefly introduce the conceptual foundations of machine learning approaches to text classification and discuss their application in social science research. We then evaluate their potential in an experimental study in which German online news was coded with established thematic categories. Moreover, we investigate whether and how linguistic preprocessing can improve classification quality. Results indicate that supervised text classification is generally robust and reliable for some categories, but may even be useful when it fails.

Keywords Content analysis · Machine learning · Online news · Bayesian classifier

1 Introduction

The ever-growing amount of publicly available content from traditional media, web sites or messages on platforms like Twitter or Facebook has challenged traditional methods of content analysis for more than a decade (Weare and Lin 2000). Apart from the largely unsolved problems in representative sampling or the effective handling of multi-modal analysis of textual and audiovisual material, the sheer quantity of potentially important text data calls for automated solutions in collecting, preparing and coding. At the same time, many research questions concerning new ways of public and/or interpersonal communication on the Internet require large-scale analyses which cannot be conducted using manual coding techniques. The

M. Scharkow (✉)
Institute of Communication Studies, University of Hohenheim,
Wollgrasweg 23, 70599 Stuttgart, Germany
e-mail: michael.scharkow@uni-hohenheim.de

need for reliable and scalable solutions in analyzing messages has led to a renewed interest in automatic or computer-aided methods of content analysis (Popping 2000; Krippendorff 2004a).

While traditional computer-aided methods like dictionary-based coding (Stone et al. 1966) or co-occurrence-analysis (Doerfel and Barnett 1996) are still used for the majority of all applied research in this area (Alexa and Zuell 2000), a number of new approaches to automatic text analysis have recently found their way into the social sciences (Hillard et al. 2007; Monroe and Schrodt 2008), most notably semantic network analysis (van Atteveldt 2008) and supervised text classification (Sebastiani 2002).

We argue that supervised text classification, which uses superficial statistical algorithms from machine learning, has the potential to become a standard method for quantitative content analysis. By using manually coded material for training, supervised classification seamlessly bridges the gap between traditional thematic and automatic content analysis. Unlike other automatic approaches, supervised classification does not require a completely different way of conducting content analyses. Rather, it can be added with little extra effort to any manual, thematic content analysis (Roberts 2000), which is still the workhorse of communication research.

In this paper, we will introduce the basics of machine learning-based text coding and provide an empirical evaluation of its potential for applied content analysis. In the following section, we will discuss the merits of supervised learning compared to other techniques. We will then present the results of a feasibility study in which German news articles were first coded manually and then automatically using a Naive Bayesian classifier. The principal research question of this evaluation is how well supervised classification works and how it can be improved by using various preprocessing techniques that are common in computer-aided content analysis.

2 Why use machine learning for content analysis?

2.1 Traditional computer-aided content analysis

Since the seminal work of Stone et al. (1966) as well as Iker and Harway (1969), a number of different approaches to automatic text analysis have been developed within and outside of the social sciences. These techniques can be classified according to different criteria, such as being thematic versus semantic (Roberts 1997) or supervised versus unsupervised (Hillard et al. 2007). For the purpose of this paper, we will not discuss unsupervised approaches like text statistics, stylometry, co-occurrence analysis or document clustering because most content analyses are hypothesis driven rather than purely descriptive or exploratory. As unsupervised methods neither require nor allow the researcher to specify the rules according to which content is coded, they cannot be used to enforce a specific interpretation or coding behavior. This directed reception of messages, however, is the main premise under which content analyses are conducted (Krippendorff 2004a).

The most frequently used approaches to automatic text coding using dictionaries or syntactic parsers are rule based: the researcher specifies either words or parsing rules that are deterministically applied to a given document (Stone et al. 1966; Schrodt et al. 1994). In a thematic analysis, for example, an article may be coded as sports if it contains words like referee, play-off or foul. In syntactic-semantic analyses on the proposition level, a parser may extract named entities as actors or verbs as links between actors (van Cuilenburg et al. 1988; King and Lowe 2003; van Atteveldt 2008).

While dictionary-based approaches have been virtually unchanged since the development of the General Inquirer in the 1960s (Stone et al. 1966), recent developments in computational linguistics have had significant impact on the use of automated semantic network analysis (van Atteveldt 2008). Syntactic parsers are available for many languages, as are additional preprocessing tools like part-of-speech-tagger or lemmatizer (Manning and Schütze 1999; Hotho et al. 2005). This enables researchers to quickly and reliably extract actors and concepts from large text corpora (Schrodt 2010).

Dictionary- and parser-based approaches follow a deductive approach to operationalization: In order to use them effectively, a researcher has to develop a complete and coherent theory of how theoretical concepts of interest manifest themselves in natural language. Since the coding by a computer is completely deterministic, researchers must carefully test and refine the rules of classification or text extraction. This leads to a content analytic process that is far removed from traditional manual coding—the result is a widening gap between researchers who apply manual procedures and those who use elaborate computer algorithms. While the development of valid dictionaries is already a challenging task that has often taken much effort (Lasswell and Namenwirth 1968), the development or even application of advanced parsers requires special knowledge that is rarely available in social science departments. This is especially problematic since, in most cases, the parsing and coding needs to be tailored for any domain-specific research question. While the general framework of semantic network analysis is language- and topic-agnostic (van Atteveldt et al. 2010), the actual computerized document processing is not (van Atteveldt 2008).

2.2 Reconciling manual and automatic coding with supervised text classification

In contrast to the procedures described above, machine learning uses an inductive approach of knowledge acquisition. A machine learning algorithm is trained with preceded data and derives the rules by which the given decisions can be reproduced. In supervised text classification, an algorithm takes documents and their correct category assignments (classes) as inputs, derives a "probabilistic dictionary" (Pennings and Keman 2002) from this data, and uses this information for the classification of new documents. The training process for the classifier bears a strong resemblance to conventional coder training, which is heavily based on example documents (Krippendorff 2004a). The computer classifier is basically treated like any human coder, albeit one with limited language skills and no contextual knowledge. Since supervised learning is a purely statistical approach, it can be used in any language and with any topic category (Hillard et al. 2007). Moreover, any manual thematic content analysis can be automated by using the documents coded by humans as training and test data for one or more supervised classifiers. Whether or not a category can successfully be coded by a computer is, of course, an empirical question. For this purpose, it is straightforward to conduct a coder-computer reliability test which uses the same metrics as conventional intercoder reliability tests.[1] Given these advantages, supervised text classification can be seen as a natural extension to conventional content analyses. Thereby, it becomes immediately appealing for social science applications.

Unlike in semantic analyses, supervised classification makes no assumptions about syntax but treats any text as a simple bag of words (Manning and Schütze 1999; Sebastiani 2002). Consequently, the machine learning approach to content analysis is solely based on superficial, i.e. lexical, features of a text and the assumption that single words or word combinations

[1] By convention, disagreement between a human coder and a classifier is almost always interpreted as a misclassification by the computer, indicating a lack of validity. However, since both training and test documents are likely to include wrong category assignments, disagreement can also be seen as a reliability issue.

(N-grams) provide enough information for thematic categorization. Therefore, inductive text classification is not suited for research questions that focus on the structure of individual assertions, nor can it be used to answer open-ended questions to the text which is the target of algorithmic information extraction (King and Lowe 2003).

Supervised text classification has been one of the most extensively researched areas in machine learning for more than a decade (Hillard et al. 2007). As a result, a variety of different classification algorithms have been developed and evaluated, often in applications like spam filtering (Cormack and Lynam 2007), opinion mining and sentiment analysis (Pang and Lee 2008). In the social sciences, machine learning has been employed in the topical classification of legislative and other legal documents (Purpura and Hillard 2006; Evans et al. 2007), political blog postings (Durant and Smith 2007) and party manifestos (Laver et al. 2003). However, very few studies have used supervised classification in communication research, i.e. on regular media content (Leopold and Kindermann 2002; van Atteveldt 2008). Furthermore, many evaluation studies used categories that are rarely employed in day-to-day content analysis, which makes inferences about the real-world performance of this approach somewhat difficult.

Another important research topic in text classification focusses on the application of automatic linguistic preprocessing and its effects. This includes includes both statistical and algorithmic procedures that aim at reducing irrelevant content which in turn should lead to increased classification quality (Manning and Schütze 1999). Comparatively little research has been conducted in this area, especially from a social science perspective (van Atteveldt 2008). Moreover, many common recommendations that preprocessing is necessary were made at a time when coding was the only automatic step in an otherwise manually conducted research process (Iker and Harway 1969; Landmann and Züll 2008). Thus, we argue that further empirical research is needed and try to offer some answers in this paper.

In sum, machine learning promises to be an ideal complement and extension to classic thematic content analysis because (1) it directly uses manually coded documents as training data, (2) is language and topic-agnostic, (3) can be used and evaluated in the same way as conventional analyses and (4) requires little to no extra effort because data collected by hand-coding can be used to quietly train and test a classifier in the background. Compared to traditional methods of automated content analysis, supervised learning does not require different operationalization strategies. Unlike traditional deductive approaches, the initial effort to get started with automated coding is very low, which makes machine learning equally attractive for small- and large-scale projects. All these arguments, however, are purely conceptual and it remains to be tested whether supervised learning does actually work for thematic content analysis.

3 Feasability study

3.1 Research questions and hypotheses

In the remainder of this paper, we present the findings from an empirical evaluation of using supervised classifier for thematic content analysis. The principal research question for this study is if—and to what extent—the text classification algorithms in machine learning are suitable for social science applications. Consequently, we investigate the reliability and validity of a supervised approach compared to manual coding.

Since all inductive methods of text classification rely on (hand-coded) training data, it is obvious that automatic classification cannot outperform manual coding in terms of reliability and validity. Any automatic classification is only as good as its training material. However, it is yet unclear if there is a linear relationship between the reliability of the manual and the automatic text classification. Two arguments can be made to support such a notion: First, categories that rely mainly on the lexical and syntactical content of a message are more readily recognized by human coders *and* the machine learning algorithm, whereas coding semantic or even pragmatic aspects of texts is often subject to coders' judgments (Potter and Levine-Donnerstein 1999; Krippendorff 2004a). This is a conceptual argument that is inherent to the way content analysis works. Second, if the training data is less polluted by misclassifications, it is easier to develop a statistical model for classification. Sheng et al. (2008) demonstrate that in many situations, it is desirable to invest in the quality rather than the amount of the training data. This leads us to our first hypothesis:

H1 Categories that are more reliable when coded manually will also be more reliable in supervised classification.

The second part of our primary research question is: how can classification performance be optimized by preprocessing the text data? Many scholars in the field recommend and perform various statistical or linguistic preprocessing steps (Popping 2000; van Atteveldt 2008; Landmann and Züll 2008). Often, the central aim of these steps is the reduction of features, often by removing words or parts of words considered noise. The most frequently used techniques are lemmatization or stemming, i.e. the substitution of inflected word forms with their stems or lemmas, text filtering, i.e. the removal of extremely frequent (stop words) or infrequent words in a corpus. Both stemming and stop word removal are relatively easy to implement and do not require complex linguistic machinery.[2] Although empirical findings concerning the benefits of stemming and stop-word removal are mixed (Leopold and Kindermann 2002), we follow common recommendations and therefore expect a positive effect on classification quality because of the potential noise reduction:

H2 Both stemming and the removal of frequent (stop) words improves classification quality.

Finally, we ask whether the supervised classification of online content can be enhanced by a different kind of preprocessing: the extraction of the relevant text from a complex web page. In manual analyses of online content, web sites are often downloaded more or less completely using mirroring software and then hand-coded with the help of a common browser software. Since the HTML document is displayed graphically using different styles, colors or supporting images, human coders have little difficulty in recognizing the relevant document parts for any category. Automatic text classification, on the other hand, often only extracts textual features like words or chars with no regard for structure or layout. Therefore, it is far more difficult for the computer to recognize the relevant part of a document for classification. Since web pages can vary greatly in their composition, it would be desirable if text classification worked irrespective of any additional markup. If that were the case, no further preprocessing would be required. Alternatively, classification quality for online content could be improved by filtering out the irrelevant text before proceeding with the analysis. Recent studies indicate that cleaning up complex documents leads to better classification quality (Li and Ezeife 2006), which leads to our final hypothesis:

[2] More linguistically challenging procedures include part-of-speech-tagging, real lemmatization or anaphora resolution (Hotho et al. 2005; van Atteveldt 2008). These require comparatively sophisticated and language-specific software algorithms, which makes them less attractive for conventionally trained communication researchers.

H3 The automatic extraction of the main text from a HTML document improves thematic classification.

If any of the mentioned preprocessing steps do not significantly improve the reliability and validity of the text classification, it will be even more straightforward to integrate the machine learning approach into common content analytic procedures. Since most of the preprocessing tasks except the body text extraction are language-specific, much effort in multilingual automatic content analyses could be saved by their omission. However, all this rests on the premise that machine learning is suited for thematic content analysis.

3.2 Method

Sample In order to conduct the evaluation study in a realistic setting, we used a sample of articles from 12 German news web sites.[3] The data collection was fully automated using a custom software that permanently retrieves all news published in the RSS feed of the site and stores them in a database together with meta data like the publication date or the original URL. Using simple text replacement rules, the tool downloads a printable version of an article in order to minimize undesired content like navigation or advertisements.

For the actual coding, we drew a random sample of 1,000 documents from 208,000 articles that were published between June 2008 and May 2009. After removing documents that contained little or no textual content, a total of 933 documents were then coded by eight trained annotators.

Classification algorithm and software In the past decade, many supervised classification algorithms have been developed and tested in various settings. Since this study is not about comparing algorithms, we chose a simple Naive Bayes (NB) classifier with enhanced feature selection. NB classifiers have been shown to be fast, robust and well suited for many classifications tasks (Durant and Smith 2007; Hillard et al. 2008). The NB model follows the assumption that the category or class of a document can be derived from the conditional probability of being in category c given it contains word (or more generally feature) w. The feature occurrences in the training data are used to compute the conditional probability $P(w|c)$, e.g. the probability that the word referee is found in a sports article. With these individual probabilities and a given document, it is possible to compute the probability that it is about sports.

$$P(c|w) \propto P(c)P(w|c) \tag{1}$$

From the vast amount of software implementations, we selected the OSBF-Lua library developed by Assis (2006) which has won multiple competitions in spam filtering, is open source and has a simple front end which consists mainly of the *train* and *classify* commands. OSBF-Lua uses sparse bigrams instead of single word features which significantly improves classification accuracy (Siefkes et al. 2004).

Measures Apart from the use of actual news material, a second important step in order to ensure the ecological validity of the evaluation is the choice of content-analytic measures. The codebook used in this study is comprised of categories which have repeatedly been applied in traditional manual analyses. Specifically, we used eight variables that cover either

[3] This included the online versions of seven daily newspapers such as sueddeutsche.de oder faz.net, three weeklies such as spiegel.de and zeit.de, and the two public service broadcasting news tagesschau.de and heute.de.

Table 1 Inter-coder reliability for the manual content analysis

Variable	Percent agr.	CI_{Acc}	Kripp. α	CI_{α}	n_{Art}
National politics	0.90	0.87–0.92	0.69	0.60–0.76	373
International politics	0.93	0.90–0.95	0.76	0.65–0.82	373
Pol. economics	0.93	0.91–0.95	0.74	0.65–0.83	373
Sports	0.99	0.98–1.0	0.98	0.93–0.99	395
Disasters, accidents	0.95	0.93–0.97	0.67	0.54–0.80	373
Crime	0.92	0.89–0.95	0.67	0.56–0.77	373
Controversy	0.69	0.65–0.74	0.49	0.40–0.56	373
Prominence	0.71	0.66–0.75	0.72	0.66–0.77	373

Confidence intervals are bias-corrected percentile intervals from bootstrapping (Hayes and Krippendorff 2007)

the topic of an article or different *news factors* which are frequently used in analyses of news (Eilders 2006). For the thematic variables, we used codebooks by Bruns and Marcinkowski (1997) and GöFAK Medienforschung (2010); news factors were operationalized according to Fretwurst (2008). Most variables are dichotomous, with news factors *Controversy* and *Prominence* being coded on a three-point ordinal scale.

About one third of all documents were coded independently by at least two persons in order to ensure proper coverage for the reliability tests. The selection of test documents happened during the normal coding, so that inferences can be made for the whole coding process. Table 1 summarizes the reliability of the coding for all variables. All thematic variables show high levels of inter-coder agreement that are in line with the reliability reported in the original studies. However, since simple percent agreement (Holsti 1969) is often positively biased, we prefer the chance-corrected coefficient alpha developed by Krippendorff (2004b). A closer look at the columns for α reveals that most variables have acceptable levels of reliability, with the notable exception of *Controversy*.

Procedure The empirical evaluation of the classification quality follows a typical train-classify-compare process (Manning and Schütze 1999), which is then varied according to a factorial treatment plan. A single evaluation run is comprised of the following steps:

1. For every document in the sample, a correct or gold-standard category or label is selected according to the manual coding. In case a document has been annotated by more than one coder, the category is randomly chosen from the codes given. This is necessary in order to account for the imperfect reliability of the manual coding.
2. The sample is partitioned into ten equal folds, with one fold reserved for testing, the other nine for training the classifier. This is repeated for all ten sets, so that we have a tenfold cross validation, with every document used once as a test case.
3. The classifier is trained using all documents (in random order) in the training set. If necessary, different treatments are applied to the documents before training (and testing).
4. The documents in the test set are automatically classified, and the classifications compared against the known manual categories. This yields a simple misclassification table which can be summarized afterwards.

The whole process is repeated according to a full factorial design with replications. Three different treatments are applied to the documents before training and testing. These treatments are:

Table 2 Classification quality for supervised learning

Variable	CR$_a$	CR$_a$–CR$_m$	α_a	α_a–α_m	Precision	Recall
National politics	0.86	−0.04	0.55	−0.14	0.65	0.63
International politics	0.89	−0.04	0.61	−0.15	0.77	0.60
Pol. economics	0.90	−0.03	0.61	−0.13	0.65	0.69
Sports	0.96	−0.03	0.84	−0.14	0.94	0.80
Disasters/accidents	0.93	−0.02	0.17	−0.50	0.78	0.12
Crime	0.86	−0.06	0.36	−0.31	0.66	0.32
Controversy	0.62	−0.07	0.30	−0.19	0.62	0.52
Prominence	0.60	−0.11	0.45	−0.27	0.73	0.63

Subscripts a and m indicate manual and automatic coding. Precision and recall only apply to automatic coding

Stemming The removal of common suffixes, using the algorithm by Porter (1980) adapted for the German language.

Stop word removal The 1,000 most common words in German are removed from the text, using a list from the Wortschatz project.[4]

Text extraction The body text is extracted from raw HTML using an algorithm by Finn et al. (2001).

In order to account for the variability in the results induced by (a) the choice of true categories in documents with multiple codings and (b) the composition of the folds, each experimental condition is replicated eight times for all eight variables coded. The analyses are therefore based on $n = 512$ evaluation runs for the full factorial design.

3.3 Results

In general, the results of this study indicate that supervised classification is a viable option for simple thematic content analyses but does not work reliably in all categories. Table 2 displays the reliability scores for the automatic classification as well as differences between manual and automatic coding quality. The most reliable category is *Sports*, which can be classified with very high accuracy (0.96). The categories for various political topics are also quite suited to supervised classification. This is not surprising, given that topical classification heavily relies on lexical features without much need for syntactic or semantic analysis. Looking at the simple accuracy measures, one could easily come to the conclusion that supervised classification is almost as good as manual coding. The average difference in accuracy is only about 5%, which is acceptable for automatic coding.

As argued by Krippendorff (2004b) and others (Eugenio and Glass 2004; Reidsma and Carletta 2008), simple accuracy measures are often upwardly biased when categories are unevenly distributed. Since a Bayesian classifier uses the class proportions as prior probabilities, high percent agreement could be a result of guessing according to the class distribution rather than successful learning. In order to assess the actual coding quality, chance corrected measures such as Krippendorff's α are therefore preferable.

Looking at these coefficients, one can see that automatic coding is still reliable and valid for *Sports* and acceptable for *Politics*. However, the limitations of the supervised learning approach become clear when we look at the results for the more difficult variables, notably *Controversy* and *Crime*. The chance-corrected reliability of the classification of 0.3 and 0.36

[4] http://wortschatz.uni-leipzig.de/Papers/top1000de.txt.

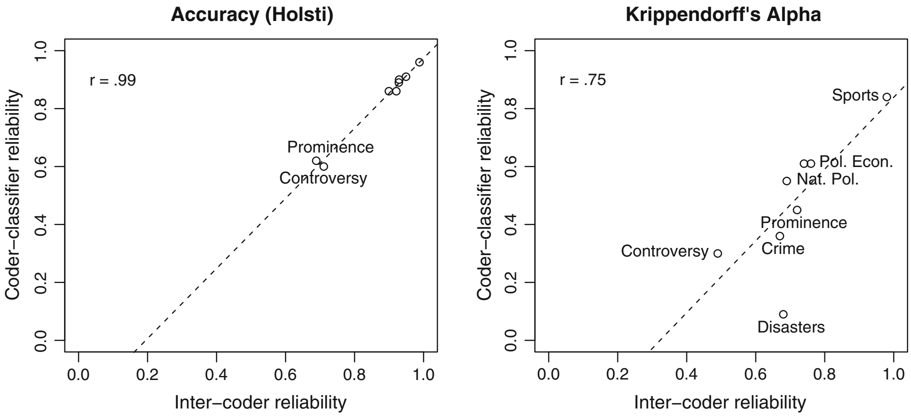

Fig. 1 Relationship between inter-coder reliability and classification quality

is clearly insufficient. The variable *Disasters/Accidents* cannot be reliably classified at all. A closer examination of the classification quality in terms of *Precision* and *Recall* (Manning and Schütze 1999; Krippendorff 2004a) reveals that supervised classification is systematic. In most cases, precision is higher than recall, i.e. the number of false negative classifications often exceeds the false positives.[5] In other words, the classifier rarely misclassifies a true crime-related article but overlooks quite many of them. There are two plausible explanations for these results: First, the decision of whether an article refers to a crime or controversy requires quite a lot of contextual knowledge, i.e. what is a crime, disaster or what constitutes a controversy. Human coders who lack this knowledge also struggle to correctly classify these articles. Second, *Disasters/Accidents* and *Crime* are comparatively rare categories (the former more so than the latter, with a prevalence of only 7% while crime occurs in 16% of all documents), and the classifier may simply not have enough training material to develop a stable statistical model.

In order to inspect this issue we analyzed the relationship between inter-coder and classification reliability. Figure 1 displays the data for both percent agreement (accuracy) and Krippendorff's α coefficient. Looking at simple agreement statistics, we see a nearly perfect linear relationship between manual and classification accuracy. This, however, is mostly due to the fact that there is very little variation in this coefficient across the different categories. Again, a look at the chance-corrected coefficient provides some more information about the classification quality. For most variables, the automatic coding is about 20% less reliable than the manual annotation, with a strong linear relationship between the two (see also Table 2). The variable *Disasters* is clearly less reliably automated than expected, while *Controversy* is somewhat better suited for supervised classification.[6] Given these results, we suspect that the bad performance of the classifier in recognizing news about accidents and disasters can be explained by the insufficient quantity of training material rather than its quality.

A more detailed look into the supervised classification process is given in Table 3 which summarizes the results for a regression analysis of the factorial experiment. The table displays the main and interaction effects of the three preprocessing steps on classification quality. In addition to the simple accuracy measure and Krippendorff's α, the effects of the

[5] All variables were dichotomized before computing precision and recall.

[6] In fact, the correlation between manual and classification reliability is $r = 0.96$ if both variables are omitted.

Table 3 Effects of preprocessing on classification quality for different categories, unstandardized regression coefficients and standard errors

	Kripp. α	Accuracy	Precision	Recall
Stemming	−1.5 (0.5)	−0.5 (0.2)	−0.7 (1.6)	−1.2 (0.8)
Stop word removal	−5.2 (0.5)	−2.6 (0.2)	−4.1 (1.6)	1.1 (0.8)
Text extraction	9.8 (0.5)	3.6 (0.2)	15.6 (1.6)	0.7 (0.8)
Stemming × Stop	0.4 (0.6)	−0.2 (0.3)	−2.0 (1.8)	1.5 (0.9)
Stemming × Text Extr.	−0.4 (0.6)	0.3 (0.3)	−2.3 (1.8)	−0.2 (0.9)
Stop × Text Extr.	−0.2 (0.6)	1.2 (0.3)	0.7 (1.8)	−4.5 (0.9)

Full factorial design with replications, $n = 512$

preprocessing procedures on precision and recall were also estimated. For improved readability, all dependent variables have been rescaled, i.e. multiplied by 100.

Since the experiment was conducted in a highly controlled setting, and the regressions included intercept terms for every variable in the codebook, nearly all variance in the reliability and validity coefficients can be explained by the model. It is interesting to note that our study corroborates the old saying that content analysis is all about the categories. Nearly 92% of all variance in the data is between the different categories, with the experimental treatments accounting for about 5%. The residual variance is then due to the partitioning of the folds and the selection of correct labels for documents that we annotated by more than one coder.

The results of the experimental evaluation show that neither stemming nor the removal of common stop words improves classification accuracy. On the contrary, removing the most frequent words significantly and consistently decreases the performance of the supervised classifier. This can be explained by the fact that OSBF-Lua, like many advanced classifiers, uses bigrams instead of single words as features. Consequently, much of the semantics in negations or other common idioms that contain stop words is lost if these are removed in advance. Since these features may be highly discriminant for the classes, their omission leads to worse performance. Contrary to common recommendations and the findings of Braschler and Ripplinger (2004), the use of stemming has a very small negative effect on classifier performance of German texts. According to our results, both preprocessing steps may be omitted in supervised classification tasks.

The algorithmic extraction of the main text from a complex HTML document is the one preprocessing step which significantly improves classification performance. Removing unnecessary features from documents especially enhances the precision of classification, so that fewer false positives occur. This means that when applying text extraction, one can be quite certain that a document classified as political news or controversial does really belong in this category. The positive effect of text extraction varies across the categories and is as large as 20% for *Crime* and 50% for *Disasters/Accidents*. Apart from the improved classification quality, automatic text extraction significantly speeds up the whole classification process as the average document size is reduced by more than 90%. We can therefore recommend this practice for the classification of online content.

One final important finding from our evaluation is that the recall measure cannot be improved by using any preprocessing. Recall is consistently lower than precision under all conditions in the experiment. Consequently, one cannot reliably use supervised classifiers as a filtering tool for manual content analyses as many important articles may never be seen by a human coder. Of course, one can modify the classification algorithm in order to optimize the recall, but given a limited training set, automatic classifiers will often fail to recognize all important aspects of a topic such as crime or disaster.

4 Discussion

Our evaluation has demonstrated that machine learning is a viable option when researchers want to apply codebooks developed for manual analysis to large document corpora. It serves its purpose particularly well in situations where categories are well defined and reliably coded by hand. In general, supervised automatic coding is about 15% less reliable than human judgement, although there are both positive and negative exceptions to this rule. The computer has difficulties with categories that rely on contextual knowledge, like some news factors. However, these categories are often difficult for human coders as well. Supervised learning is also unsuited for rare categories because it takes a few dozen or even hundreds of positive training documents to establish a stable statistical model for classification. Nonetheless, there are many situations in which a researcher would gladly accept a certain loss of reliability and validity in order to be able to code thousands of documents quickly without additional effort. Answering Schrodt (2010), we can safely say that there is no need to filter out sports or business news by hand anymore.

The experimental study also showed that most preprocessing steps do not improve supervised classification, with the specific exception of text extraction for web content. Put differently, machine learning is a very robust tool which can deal with noisy documents. Since all language-specific preprocessing steps did not improve classification, the supervised approach can basically be used with every topic model and for every language without any modifications to the software. This means that it is possible to develop a unified content analytic workflow that incorporates manual coding and automatic text classification without losing generality. In this setting, machine learning is useful even if the actual classification is not reliable and valid enough for serious inferences. Since the classifier software behaves like a naive coder, many misclassifications actually help the researcher to refine the codebook by indicating vague or unclear coding instructions. In that way, using automatic techniques can improve traditional content analyses without extra cost.

Supervised classification is certainly not a silver bullet solution to all content analytic problems. It covers only one (but an important) type of analysis—thematic coding at the document rather than the sentence level. We agree with van Atteveldt et al. (2010) that the analytic potential and reusability of data generated in semantic network analysis cannot be matched by machine learning approaches. On the other hand, supervised classification is applicable in a wider variety of settings, can be used without extensive adaptation and is strongly centered on human judgment rather than linguistic or technical issues. It is therefore neither difficult nor costly to further test its potential and limits in empirical research.

References

Alexa, M., Zuell, C.: Text analysis software: Commonalities, differences and limitations: The results of a review. Qual. Quant. **34**, 299–321 (2000)

Assis, F.: OSBF-Lua-A text classification module for Lua—The importance of the training method. In: Fifteenth TREC, Citeseer, Gaithersburg (2006)

Braschler, M., Ripplinger, B.: How effective is stemming and decompounding for German text retrieval? Inf. Retrieval **7**, 291–316 (2004)

Bruns, T., Marcinkowski, F.: Politische information im Fernsehen [Political information in television]. Leske + Budrich, Opladen (1997)

Cormack, G.V., Lynam, T.R.: Online supervised spam filter evaluation. ACM Trans. Inf. Sys. **25**(3), 11 (2007)

Doerfel, M., Barnett, G.: The use of Catpac for text analysis. Cult. Anthropol. Methods J. **8**(2), 4–7 (1996)

Durant K, Smith, M.: Predicting the political sentiment of Web Log posts using supervised machine learning techniques coupled with feature selection. In: Advances in Web Mining and Web Usage Analysis: 8th International Workshop on Knowledge Discovery on the Web, Webkdd 2006, pp. 187–206. Springer-Verlag, New York (2007)

Eilders, C.: News factors and news decisions. Theoretical and methodological advances in Germany. Communications **31**(1), 5–24 (2006)

Eugenio, B.D., Glass, M.: The kappa statistic: A second look. Comput. Ling. **30**, 95–101 (2004)

Evans, M., McIntosh, W., Lin, J., Cates, C.: Recounting the courts? Applying automated content analysis to enhance empirical legal research. J. Emp. Legal Stud. **4**(4), 1007–1039 (2007)

Finn, A., Kushmerick, N., Smyth, B.: Fact or fiction: Content classification for digital libraries. In: DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries, Dublin (2001)

Fretwurst, B.: Nachrichten im Interesse der Zuschauer. Eine konzeptionelle und empirische Neubestimmung der Nachrichtenwerttheorie [News in the viewer's interest. A conceptual and empirical re-evaluation of news values theory]. UVK Verlag, Konstanz (2008)

GÖFAK Medienforschung: Fernsehanalyse zum Bundestagswahlkampf 2009. Methodenbericht GLES1401 der German Longitudinal Election Study [content analysis of tv coverage for the bundestag election 2009]. http://www.gesis.org/fileadmin/upload/dienstleistung/forschungsdatenzentren/gles/SecureDownload/frageboegen/GLES1401_Pre1.0%20-%20Methodenbericht.pdf (2010)

Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Commun. Methods Meas. **1**(1), 77–89 (2007)

Hillard, D., Purpura, S., Wilkerson, J.: An active learning framework for classifiying political text. In: Annual Meeting of the Midwest Political Science Association, Chicago (2007)

Hillard, D., Purpura, S., Wilkerson, J.: Computer-assisted topic classification for mixed-methods social science research. J. Inf. Technol. Polit. **4**(4), 31–46 (2008)

Holsti, O.: Content Analysis for the Social Sciences and Humanities. Addison-Wesley, Reading (1969)

Hotho, A., Nürnberger, A., Paaß, G.: A brief survey of text mining. LDV Forum GLDV J. Comput. Ling. Lang. Technol. **20**(1), 19–62 (2005)

Iker, H., Harway, N.: A computer systems approach toward the recognition and analysis of content. In: Gerbner, G. (ed.) The Analysis of Communication Content. Developments in Scientific Theories and Computer Techniques., pp. 381–405. Wiley, New York (1969)

King, G., Lowe, W.: An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. Int. Organ. **57**(3), 617–642 (2003)

Krippendorff, K.: Content Analysis: An Introduction to Its Methodology, 2nd edn. Sage, London (2004a)

Krippendorff, K.: Reliability in content analysis. Human Commun. Res. **30**, 411–433 (2004b)

Landmann, J., Züll, C.: Identifying events using computer-assisted text analysis. Soc. Sci. Comput. Rev. **26**(4), 483–497 (2008)

Lasswell, H., Namenwirth, J.: The Lasswell Value Dictionary. Yale University Press, New Haven (1968)

Laver, M., Benoit, K., Garry, J.: Extracting policy positions from political texts using words as data. Am. Polit. Sci. Rev. **97**, 311–331 (2003)

Leopold, E., Kindermann, J.: Text categorization with support vector machines. How to represent texts in input space? Mach. Learn. **46**(1–3), 423–444 (2002)

Li, J., Ezeife, C.: Cleaning web pages for effective web content mining. In: Database and Expert Systems Applications. Springer, Berlin, pp. 560–571 (2006)

Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA (1999)

Monroe, B.L., Schrodt, P.A.: Introduction to the special issue: The statistical analysis of political text. Polit. Anal. **16**(4), 351–355 (2008)

Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations Trends Inf. Retrieval **2**(1–2), 1–135 (2008)

Pennings, P., Keman, H.: Towards a new methodology of estimating party policy positions. Qual. Quant. **36**(1), 55–79 (2002)

Popping, R.: Computer-assisted Text Analysis. Sage, Thousand Oaks, CA (2000)

Porter, M.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)

Potter, W.J., Levine-Donnerstein, D.: Rethinking validity and reliability in content analysis. J. Appl. Commun. Res.H **27**(3), 258–284 (1999)

Purpura, S., Hillard, D.: Automated classification of congressional legislation. In: Proceedings of the 2006 International Conference on Digital Government Research, pp. 219–225 (2006)

Reidsma, D., Carletta, J.: Reliability measurement without limits. Comput. Ling. **34**(3), 319–326 (2008)

Roberts, C.: Introduction. In: Roberts C (ed) Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts, pp. 1–8. Lawrence Erlbaum Associates, Mahwah, NJ (1997)

Roberts, C.W.: A conceptual framework for quantitative text analysis. Qual. Quant. **34**(3), 259–274 (2000)

Schrodt, P.: Automated Production of High-Volume, Near-Real-Time Political Event Data. Paper presented at the 2010 APSA Conference (2010)

Schrodt, P., Davis, S., Weddle, J.: Political science: KEDS—a program for the machine coding of event data. Soc. Sci. Comput. Rev. **12**(4), 561 (1994)

Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1), 1–47 (2002)

Sheng, V., Provost, F., Ipeirotis, P.: Get another label? Improving data quality and data mining using multiple, noisy labelers. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 614–622. ACM, Las Vegas, NV (2008)

Siefkes, C., Assis, F., Chhabra, S., Yerazunis, W.S.: Combining winnow and orthogonal sparse bigrams for incremental spam filtering. In: Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, pp. 410–421 (2004)

Stone, P., Dunphy, D., Smith, M., Ogilvie, D.: The General Inquirer: A Computer Approach to Content Analysis. The MIT Press, Cambridge, MA (1966)

van Atteveldt, W.: Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content. BookSurge Publishers, Charleston, SC (2008)

van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, N.: Semantic network analysis: A two-step approach for flexible, reusable, and combinable content analysis. Paper presented at the 2010 ICA conference, Singapore (2010)

van Cuilenburg, J.J., Kleinnijenhuis, J., de Ridder, J.A.: Artificial intelligence and content analysis. Qual. Quant. **22**(1), 65–97 (1988)

Weare, C., Lin, W.: Content analysis of the World Wide Web: Opportunities and challenges. Soc. Sci. Comput. Rev. **18**(3), 272–292 (2000)