Social media research: The application of supervised machine learning in organizational communication research

van Zoonen, W.; van der Meer, T.G.L.A.

DOI
10.1016/j.chb.2016.05.028

Publication date
2016

Document Version
Final published version

Published in
Computers in Human Behavior

Computers in Human Behavior 63 (2016) 132–141

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Social media research: The application of supervised machine learning in organizational communication research.

Ward van Zoonen*, Toni, G.L.A. van der Meer

The Amsterdam School of Communication Research ASCoR, University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 25 November 2015
Received in revised form
10 May 2016
Accepted 12 May 2016
Available online 20 May 2016

Keywords:
Twitter
Supervised machine learning
Communication research
Content analysis

ABSTRACT

Despite the online availability of data, analysis of this information in academic research is arduous. This article explores the application of supervised machine learning (SML) to overcome challenges associated with online data analysis. In SML classifiers are used to categorize and code binary data. Based on a case study of Dutch employees' work-related tweets, this paper compares the coding performance of three classifiers, Linear Support Vector Machine, Naïve Bayes, and logistic regression. The performance of these classifiers is assessed by examining accuracy, precision, recall, the area under the precision-recall curve, and Krippendorf's Alpha. These indices are obtained by comparing the coding decisions of the classifier to manual coding decisions. The findings indicate that the Linear Support Vector Machine and Naïve Bayes classifiers outperform the logistic regression classifier. This study also compared the performance of these classifiers based on stratified random samples and random samples of training data. The findings indicate that in smaller training sets stratified random training samples perform better than random training samples, in large training sets ($n = 4000$) random samples yield better results. Finally, the Linear Support Vector Machine classifier was trained with 4000 tweets and subsequently used to categorize 578,581 tweets obtained from 430 employees.

1. Introduction

Social media use in organizations is evolving at an unprecedented rate (Treem & Leonardi, 2012). Social technologies may enable employees to communicate more effectively, widen the scope of their work and boost their performance (e.g., Ollier-Malaterre, Rothbard, & Berg, 2013). Likewise, social media can play a crucial role in organizations' post-crises communication (Schultz, Utz, & Goritz, 2011; van Zoonen & van der Meer, 2015), as well as in their efforts to influence evaluations of corporate reputation (Helm, 2011) and to engage in stakeholder dialogues (Lovejoy, Waters, & Saxton, 2012). Some scholars argue that social media are reshaping the nature of the workplace and of work itself (Bucher, Fieseler, & Suphan, 2013), whereas others suggest social media are the new hybrid element in the promotion mix (Mangold & Faulds, 2009). Hence, it is of no surprise that social media is top of the agenda for many practitioners and scholars (Kaplan & Haenlein, 2010).

Organizational communication scholars frequently rely on content analysis research for a wide range of empirical questions. Specifically so, the advent of social media propelled the use of content analysis in organizational communication research as it provides a wealth of well-documented information (e.g., Gallaugher & Ransbotham, 2010; Ki & Nekmat, 2014; McCorkindale, 2010; Rybalko & Seltzer, 2010; van Zoonen, Verhoeven, & Vliegenthart, 2016). The use of content analysis is particularly appealing to researchers because examination of narrative texts - such as social media interactions and blogs - allow the unobtrusive study of organization-public interactions that are otherwise difficult to obtain (Duriau, Reger, & Pfarrer, 2007).

The growth of social media use in the workplace has provided an abundance of data that reflects new media activities and artifacts relevant to organizational communication research (Lewis, Zamith, & Hermida, 2013). In spite of the alluring promise of online data availability, analysis of this information remains arduous. Online data abundance often proves to be 'fool's gold' as: a) researchers often struggle to 'trap' these large data streams, b) large-scale content analyses tend to be unfeasible due to the time-

* Corresponding author. University of Amsterdam, The Amsterdam School of Communication Research *ASCoR*, Nieuwe achtergracht 166, 1018 VW Amsterdam, The Netherlands.

E-mail address: w.vanzoonen@uva.nl (W. van Zoonen).

http://dx.doi.org/10.1016/j.chb.2016.05.028
0747-5632/© 2016 Elsevier Ltd. All rights reserved.

consuming and costly realities of human coding methods, and c) computer-assisted content analysis methods are insufficiently institutionalized in organizational communication research (Lewis et al., 2013). In order to alleviate these concerns this study aims to explore the application of an advanced computer-assisted content analysis approach – i.e., supervised machine learning (SML: Grimmer & Stewart, 2013 Hillard, Purpura, & Wilkerson, 2008; Rusell & Norvig, 2002) – that could benefit a variety of research streams in organizational communication.

This study applies a method that is based on SML requiring that a computer learns how to automatically predict content-analytical variables in the corpus of data from a set of human-coded training documents. Notably, in the past SML procedures have predominantly been applied to larger texts such as newspaper articles or speech transcripts (Burscher, Odijk, Vliegenthart, de Rijke, & de Vreese, 2014; Hillard et al., 2008; Scharkow, 2013). Significantly, this study explores the performance of SML to *short* social media messages - i.e., tweets, which are typically limited to 140 characters. Introducing the use of SML for social media content minimizes both the efforts and investments required for content analysis of big data, and helps address substantial issues currently faced by organizational communication scholars and practitioners. The central aim of this study is to answer the research question: To what extent can SML be applied to the content analysis of social media messages?

This study explores the application of SML and aims to advance organization studies by providing a method with which to overcome the limitations associated with both human coding procedures and computer coding procedures. If SML can be used to code social media content a vast array of online information relevant to organizational research awaits empirical analysis.

Moreover, this study explores what type of classifier yields the most reliable coding decisions. The performance of three different classifiers is examined: Linear Support Vector Machine, Naïve Bayes, and standard logistic regression. The algorithms are used to identify the extent to which tweets are related to employees' work. The performance of these classifiers in terms of categorizing social media content is assessed by evaluating key performance indices – i.e., accuracy (AC), recall (RC), precision (PC), the area under the precision-recall curve (AUC), and Krippendorf's Alpha (KA). Additionally the SML procedure is applied to a dataset containing tweets of employees from various organizations. Thereby, this study demonstrates the use of SML in the content analysis of employees' social media messages. In the next paragraph a discussion of the application of content analysis in organizational communication research is provided, followed by an analysis of the case study used to test SML in the coding of social media content. The case study builds on the work by van Zoonen et al., 2016.

2. Content analysis in organizational communication

For organizational communication scholars, central to the value of content analysis is the assumption that content analysis of text and speech provides a replicable methodology to access deep individual or collective structures, such as values, intentions, attitudes, and cognitions (Huff, 1990). That is, ensuring content analysis is applicable to a broad range of organizational phenomena. Content analysis in the field of organizational communication research has included crisis communication (An & Gower, 2009), corporate social responsibility (Campopiano & De Massis, 2015; Cho & Hong, 2009), organization-public communication of non-profit organizations (Lovejoy & Saxton, 2012; Waters & Jamal, 2011; Waters, 2007; Waters, Burnett, Lamm, & Lucas, 2009), strategic management (Short & Palmer, 2008), and employees' work-related use of personal social media accounts (van Zoonen et al., 2016).

The current era of *Big Data* is both as enticing as it is vexing for communication scholars, as it offers a vast array of information on human and organizational communication, yet the technique and time needed to extract and analyze this data can prove unreasonably longwinded. Nonetheless, the quantity of information available has attracted a wide range of content analysis work in the field of organizational communication. In this thriving branch of research, studies predominantly focus on one specific issue (e.g., Humphreys, Gill, Krishnamurthy, & Newbury, 2013; Small, 2011), event (e.g., Reinhardt, Ebner, Beham, & Costa, 2009) organization or context (e.g., Chew & Eysenbach, 2010). The rationale behind this is a pragmatic one - to reduce the data size and data depth, to make either human coding or dictionary approaches more feasible.

In content analysis research, scholars use a technique to systematically, objectively and quantitatively describe manifest communication content (Berelson, 1952; Lewis et al., 2013). Several approaches to content analysis of social media data have been adopted in communication research. When investigating social media content, scholars often rely on human coding with indicator questions or dictionary-based computer-aided coding.

The most widely used approach in content analysis is human coding, which uses indicator questions often formulated in codebooks. It is often applauded for its systematic rigor and sensitivity towards the subtleties in human language. However, whilst being a reliable method, it is also a highly resource-intensive process. Furthermore, in times where researchers must no longer choose between data size and data depth (since data is abundantly available online) human coding procedures present fundamental challenges to content analysis (Lewis et al., 2013). Researchers are increasingly confronted with 'too much data', forcing them to resort to one of three following tactics: 1) collecting smaller samples by using less sources, reducing the timeframe, or narrowing the context of the study, 2) using random or stratified sampling methods to reduce data size (Riffe, Lacy, & Fico, 2005) or 3) allocating more financial resources (if available) to increase the number of coders hired to carry out the work (Holsti, 1969). Computational methods could offer a solution to some of the sampling and coding limitations of human coding procedures (Lewis et al., 2013).

There are two types of deductive automated coding procedures, dictionary based coding and SML. Deductive automated coding relies on a priori defined categories, in contrast to inductive automated content analysis where categories are automatically derived from the data. One of the most widely used computational coding procedures employed in communication science is dictionary-based computer-aided coding. In such cases, character strings and rules for their combination are defined a priori to code text units into content categories (Krippendorff, 2004). For example, in sentiment analysis of political texts, words such as 'respect,' 'vindication,' and 'cheerfulness' were identified as indicative for positive sentiment, whereas 'insolence,' 'malevolence,' and 'painfulness' indicated negative sentiment (Young & Soroka, 2012). Although this permits the analysis of very large datasets, several drawbacks of dictionary-based approaches cannot be overlooked. For instance, the applicability of a dictionary-approach is limited, as the phenomenon under study needs to be singular; a precondition that is often violated in organizational communication research, as research questions often span across organizational boundaries and require group comparisons and between-subject designs. Consider the following example - in a representative sample of the workforce the use of employees' personal Twitter accounts for work is analyzed. However, in order to code employees' tweets one cannot rely on a dictionary approach, as employees work in a variety of industries and jobs. Thus, creating an a priori set of words and combinations to determine work-related use becomes unattainable, if sufficient sensitivity to contextual elements is to be maintained when employees' tweets span across industries, organizations, departments and jobs.

An additional factor to consider is the particular rigidity of dictionary-approaches, which sacrifices nuanced meanings in texts and speech, or as Simon puts it: 'The chief disadvantage is that the computer is simply unable to understand human language in all its richness, complexity, and subtlety as can a human coder' (2001, p. 87). To overcome these aforementioned limitations associated with both methods, researchers suggest a hybrid approach, such as SML, combining manual and computational approaches within a single analysis, each complimenting the other (Lewis et al., 2013; Sjovaag, Moe, & Stavelin, 2012).

2.1. Supervised machine learning

SML is increasingly used to analyze (social) media content (Hillard et al., 2008; Burscher, Vliegenthart, & de Vreese, 2015). Essentially, in SML the computer attempts to replicate the coding decisions of human coders. In this way, the strengths of human coding are preserved – alongside its systematic rigor and contextual sensitivity towards latent meaning and the subtleties of human language - whilst the capacity and accuracy of computational methods is correspondingly maximized (Lewis et al., 2013). The goal is to automatically code large numbers of text documents (e.g., tweets) into previously defined content categories (Durant & Smith, 2007). Thus, a set of documents already pre-coded for the content categories is a key precondition of SML. This set is the input, or training set for the SML procedure. SML generally involves three steps. Initially, the text documents in the training set are converted to be made accessible for computational analysis. Each document is represented as a vector of quantifiable text elements (e.g., word counts), called 'features'. Secondly, feature vectors of all documents in the training set, together with the documents' content labels (e.g., presence of organization name), are used to train a classifier to automatically code the content categories. In doing so, a supervised machine-learning algorithm statistically analyses document features 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 (Burscher et al., 2014). For a detailed introduction to SML see Russell and Norvig (2002) or Grimmer and Stewart (2013).

SML offers several advantages over manual coding procedures or dictionary-based approaches. Primarily, SML enables researchers to expand the scope of their analysis. Adopting SML to organizational communication research allows a more nuanced, conditional and comparative research, which is relevant as information becomes increasingly available online. Additionally, once the classifier is trained to code the content categories it can be applied to very large datasets. This is both efficient and effective as the rules used for automated coding are based on a statistical analysis of human-coded data. Moreover, since manually coded data is available, the accuracy of computer coding decisions can be assessed systematically. Thirdly, SML permits the analysis of entire populations of texts, thereby reducing the risks associated with sampling errors and problems related to statistical accuracy (Burscher et al., 2014).

3. Case study: employees' work-related Twitter use

In order to examine the adequacy of SML it was applied to the context of organizational communication, specifically employees' use of social media for work. In particular, this context presents researchers with familiar challenges of both manual and computational methods. The quantity of employee tweets makes manually coding the corpus near to impossible, whilst, the varied work contexts of employees from different organizations and industries presents challenges that are hard to overcome in dictionary-based approaches. The context of the empirical example will briefly be outlined in the following, and the results of the SML method will subsequently be presented.

3.1. Social media use in organizations

Social media impact the nature of the workplace and of work itself (Bucher et al., 2013). As social media offer new ways of interacting with stakeholders and co-workers, its integration in the workplace is evolving at a rapid pace (Treem & Leonardi, 2012). The immediate and dialogic nature of these media has proven pervasive in today's organizations, and because these media offer a well-documented dialogue between organizations, employees and stakeholders, it presents new and exciting possibilities to answer a variety of empirical questions related to the field of organizational communication. Currently, research is directing a significant proportion of attention to employees' use of social media (e.g., Leftheriotis & Giannakos, 2014; van Zoonen et al., 2016).

Employees' social media use has been linked to increased performance and enhanced horizontal and vertical communication in organizations, yet has correspondingly also posed challenges in the form of difficult navigation between different life domains, and increased stress levels owing to social media use. Moreover, social media use during working hours is often regarded as a time-wasting or even as risky behavior (Landers & Callan, 2014). Conversely, others suggest that employees can prove to be credible and authentic communicators of brand promises (e.g., Dreher, 2014; van Zoonen & van der Meer, 2015). However, crucially none of these studies include a content analysis of employees' social media content. In fact, most studies which focus employees use of social media are conceptual (e.g., Dreher, 2014; Ollier-Malaterre et al., 2013) or rely on self-report measures (e.g., Leftheriotis & Giannakos, 2014; Marwick & Boyd, 2010) As a result, unsurprisingly several scholars have noted that there is a pronounced need for more quantitative research on the topics related to employees' social media use (El Ouirdi, El Ouirdi, Segers, & Hendrickx, 2015) and claim that social technology use in organizations is outpacing our empirical understanding on the subject matter (Treem & Leonardi, 2012).

3.2. Case study and sample

To explore the application of SML in organizational communication research, this study examines employees' work-related use of Twitter. Employees are increasingly using their personal social media accounts to discuss their work (van Zoonen et al., 2016). Based on van Zoonen and colleagues' manual content analysis of tweets, this study builds a classifier to detect whether additional tweets can be coded as work-related private. In their study the authors were able to acquire 578,803 tweets from 433 employees. van Zoonen and colleagues derived a stratified sample of 38,124 tweets because the manual analysis of the full sample would be too time consuming and costly. With the application of SML this case study demonstrates that researcher no longer need to choose between data size and data depth, as classifiers can be trained to code potentially limitless datasets. In this case study the results of the manual coding procedure are compared to the SML procedure.

The sample consists of 578,803 tweets all of which are in the Dutch language. These tweets were sent by 433 employees, who work an average of 39.62 (*SD* = 10.07) hours per week in an organization with at least 30 employees. Of these employees, 64.3% are male and the average age of the employees is 42.05 years (*SD* = 11.33). 28.1% of those sampled have an academic degree.

Together these employees have 186,139 followers and follow 162,410 other Twitter users. Thus, on average they have 495.20 (*SD* = 1225.32) followers, while they follow 426.23 (*SD* = 771.65) other Twitter users, and have sent on average 4125.09 tweets (*SD* = 9807.68).

The employees worked for different organizations in the following sectors: government/public administration (16.6%), education/science (12.1%), health care (11.7%), business services (11.2%), trade/commercial services (7.4%), industry (6.5%), and financial services (5.1%). In total, 34.6% hold a managerial position within their organization.

3.3. Analysis

In this case, SML is applied to code whether tweets are work-related or personal. Hence, the analysis performed here is aimed at categorizing employees' tweets using SML classifiers. In doing so, the applicability of SML techniques in the context of social media and work-related communication is examined. If the SML classifiers are able to correctly predict the work-related tweets, large amounts of data (collected by other researchers) can be easily coded using the computer. Furthermore, this study aims to provide insights into how much manually coded training data is needed to build a well preforming classifier. The training data is a sub-sample of the full dataset. The training data was extracted from the full dataset based on a random sample and based on a stratified random sample. The performance of the SML classifier for both these sampling techniques was compared, thereby providing direction for the adoption of SML techniques in the analysis of social media content in organizational communication research. The performance of each classifier was assessed through examining the accuracy, precision, recall, under the precision-recall curve and Krippendorf's Alpha. Finally, the trained classifier with the highest scores on the performance indices is used to code the entire dataset and, additionally, it is tested if the results based on the stratified sample from the study by van Zoonen et al. (2016) hold when all tweets are coded using SML.

Prior to conducting the automated analysis it is required that the texts are simplified and transformed into quantitative data: the pre-processing step is necessary to simplify the texts' vocabulary. Using stemming, the words are reduced to their base form in order to decrease the complexity of words that refer to the same concept. Moreover, punctuation and capitalization are removed. Second, several words are deleted from the analysis, including function words that do not convey meaning – these are removed using a stop words list. Additionally, words that appear in more than 90 percent or in less than 1 percent of the texts are removed in order to disregard very common and uncommon words. Thirdly, term frequency-inverse document frequency (TF-IDF) weighting is applied. In this instance, a word is assigned the number of times it occurs in the document (TF), weighted by its frequency of text in the dataset containing the word (IDF). The TF-IDF increases proportionally to the number of times a word occurs in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words occur more frequently in general. Finally, the texts are transformed into a document-word matrix that serves as the input for the SML.

3.4. Classifiers

To test the SML approach, a classifier needs to be trained to predict the indicator questions that distinguish the work-related category. This classifier was trained on a subset of the entire data set. To test how well the trained classifier predicts the category, the learning algorithm trained on the held-in dataset was used to obtain predictions for the held-out dataset. A specific SML algorithm must be selected. Arguably, different classifiers perform better, depending on the category and data distribution. Hence, the performance of different classifiers was compared. This study used the Scikit-Learn machine learning toolkit (cf. Pedregosa et al., 2011).[1] Scikit-Learn is a general-purpose machine-learning module for the Python programming language. This Python module integrates a wide range of state-of-the-art machine learning algorithms and provides a library of the Python programming language applied by users to tackle supervised and unsupervised problems. As Scikit-learn exposes a wide variety of machine learning algorithms, this enables easy comparison of methods for a given application. The algorithms, implemented in a high-level language, are used as building blocks for case specific approaches e.g., work-related media use.

There are several different classifiers that can be applied in an SML approach. (1) Support Vectors Machines are universal learners based on the Structural Risk Minimization principle taken from computational learning theory (Joachims, 1998). In their basic form, Support Vectors Machines learn linear threshold functions. Nevertheless, with the use of an appropriate kernel function, they can be used to learn polynomial classifiers, radial basic function networks, and three-layer sigmoid neural nets. (2) Naïve Bayes is based on applying Bayes' theorem with the assumption of independence between each pair of features (Zhang, 2004). Despite this simplified assumption, research and real-world situations have documented that this classifier works quite well. (3) Logistic regression is a linear model for classification, also known as logit regression, maximum-entropy classification, or the log-linear classifier. In logistic regression models, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function (Yu, Huang, & Lin, 2011). (4) Kernel ridge regression combines Ridge Regression (linear least squares) with a kernel trick. It learns a linear function in the space induced by the respective kernel and the data (Murphy, 2012). (5) Stochastic Gradient Decent Classifier is a simple yet very efficient approach to discriminative learning of linear classifiers under convex loss functions. It is especially useful when dealing with a very large number of samples and features. The classifiers Linear Support Vector Machine, Naïve Bayes, and standard logistic regression are compared in this study, as these three approaches performed considerably better than other classifiers such as Kernel Ridge regression and Stochastic Gradient Decent Classifier (Pedregosa et al., 2011).[2]

3.5. Evaluation performance

In order to evaluate the coding performance of the SML approach this study relies on several performance measures. The performance indices are obtained by comparing the coding decisions of the classifier to the coding decisions of the human coders. The accuracy (AC) refers to the percentage of agreement between the manual coding and the computer-based classifications. Furthermore, two measures that are based on the understanding and measure of relevance are used. Precision (PC) is reported, also known as positive predictive value, referring to the fraction of retrieved instances that are revealed. Recall (RC), also called sensitivity, is the fraction of relevant instances that are retrieved. The area under the precision-recall curve (AUC) is reported. AUC is a

[1] See Pedregoas et al., 2011 and the open source website (http://scikit-learn.org) for more information on the use of the Scikit-Learn machine toolkit.

[2] The trained classifiers used in this study and the dataset are available from the first author upon request.

commonly used performance measure for binary classification that takes into account that one already has a 50 percent chance to correctly predict the value (Sokolava & Laplame, 2009). In general, AUC measures how well an SML approach discriminates between the presence and absence of the work-related category. The key advantage of this evaluation method is that it is less sensitive to unbalanced datasets. Finally, the commonly applied inter-coder agreement statistic Krippendorff's Alpha (KA) is reported.

As with most reliability statistics in content analysis (Lombard, Snyder-Duch, & Bracken, 2002) performance-measures with a score of 1.0 imply a perfect model. These principles also apply to SML, meaning that the performance of the classifiers used in this study is evaluated by their relative score on each of the indices presented above. The reliability of the coding procedure is higher when values on these indices approach 1.0.

4. Results

4.1. Classification of work-related tweets

In Table 1 presents the performance indices (AC, PC, RC, AUC, and KA) of the Linear Support Vector Machine, Naïve Bayes, and standard logistic regression classifiers for the work-related category. Half of the manual coded dataset was randomly selected as a training dataset and the other half as the test data. The results are compared to a random baseline model. In short, this random model predicts whether a tweet should be coded as work-related or otherwise, based on the prevalence of work-related tweets in the training set.

The performance results presented in Table 1 are promising. Linear Support Vector Machine, Naïve Bayes, and logistic regression classifier indicate high coding performance. Hence, the findings suggest SML is indeed a suitable approach for automatically coding tweets on whether they are work-related or personal. For the Linear Support Vector Machine, Naïve Bayes, and logistic regression all the performance measurements outperform the random baseline model. This finding indicates that the SML model improved the prediction of correct categorization as compared to random predication based on prevalence in the training set.

When comparing the classifiers it can be observed that Linear Support Vector Machine and Naïve Bayes approaches, which perform almost equally well, surpass the logistic regression on all performance measurements. The Linear Support Vector Machine and Naïve Bayes outperform the logistic regression in terms of classification accuracy (AC), precision (PC), recall (RC), area under the precision-recall curve (AUC), and Krippendorff's Alpha (KA). It can therefore be concluded that the Linear Support Vector Machine and Naïve Bayes are more effective SML approaches than the logistic regression model in predicting the work-related variable in tweets.

4.2. Testing the classifiers with random sampling

In the following paragraphs this study explores the amount of manually coded training data that is needed to build a well performing classifier. In order to assess the relationship between the number of training tweets and the performance of SML classification, the classifiers are repeatedly trained for categorizing the tweets on being work related or personal while gradually increasing the number of tweets in the training set. Thereby, this study examined whether the performance of the classifiers increased with higher numbers of tweets in the training set. Two approaches were tested: (1) Random sampling and (2) stratified random sampling.

Table 1
Classification performance of work-related categorization.

	AC	PC	RC	AUC	KA
Random baseline[a]	0.57	0.59	0.54	0.50	–
Linear Support Vector Machine classifier	0.81	0.84	0.71	0.73	0.65
Naïve Bayes classifier	0.82	0.85	0.67	0.73	0.60
Logistic regression classifier	0.80	0.80	0.59	0.68	0.53

[a] As the random baseline model is based on chance KA is not calculated.

For the random sampling approach, a random sample of N = 1000 was used for all manual-coded tweets (N = 38,124) that were held out as a fixed set for testing. The remaining tweets were used to draw several random samples of different sizes for the training sets. In total seven training sets were constructed with the following number of tweets: 100, 200, 500, 1000, 2000, 3000, and 4000 (Burscher et al., 2014). For the Linear Support Vector Machine, Naïve Bayes, and logistic regression the measurement scores are shown in Table 2 and plotted in Fig. 1a–c. The results show that the performance indices are closer to 1.0 when the number of training tweets is higher (see Table 2). This means that larger samples of training data yield more reliable classification results as indicated by the performance indices, AC, PC, RC, AUC, and KA. When expanding the number of training tweets, the performance measurements increase more slowly for the logistic regression than for the Linear Support Vector Machine and Naïve Bayes when expanding the number training tweets. Especially from 200 to 500 one can observe a steep increase in performance. Furthermore, at 500 tweets the AUC value of 0.50 is passed indicating sufficient performance in comparison to a random baseline model.

4.3. Testing the classifiers with stratified sampling

The results of the random sampling procedure are now compared to the stratified random sampling procedure. For the stratified random sampling approach the Twitter account names were used as strata. This provided insights into whether selecting a fixed amount of tweets from each Twitter account improved prediction performance compared to drawing a random sample from the entire dataset of tweets. This is assumed since the context of work-related is not singular. As such the classifier has more specific information per user in the stratified random sampling technique. An equal percentage ratio of number of tweets for the training and the test datasets, as in the random sampling approach, was applied. Again, seven training sets were constructed with the following number of tweets per Twitter account: 1, 2, 5, 10, 20, 30, and 40. For the training dataset a stratified random sample of 10 tweets per account was held out. To make sure that the number of tweets evenly increased for every step when drawing a larger stratified sample, only the accounts with more than 40 coded tweets were selected for analysis. This selection criterion resulted in the selection of 374 Twitter accounts of the total number of 433. Hence, the training set consisted of 3740 tweets – i.e., ten tweets for each of the 374 accounts. For the Linear Support Vector Machine, Naïve Bayes, and logistic regression the measurement scores are shown in Table 3 and plotted in Fig. 2a–c. Similar to the results of the random sampling approach, the findings show that the number of training tweets led to increased performance for all classifiers, where Linear Support Vector Machine and Naïve Bayes outperformed the logistic regression. Comparing the performance of the classifiers for the stratified and random sampling, it can be concluded that the stratified sampling is a more successful approach in smaller samples. Especially on the performance indices of RC, AUC, and KA, the stratified approach seems to show a more steep increase on the performance indices. Notably, at the highest level of training data random sampling seems to perform better than stratified random

Table 2
Relationship classification performance and number of training tweets, random sampling approach.

		100	200	500	1000	2000	3000	4000
Linear support vector machine classifier	AC	0.63	0.65	0.70	0.73	0.80	0.84	0.91
	PC	0.45	0.48	0.59	0.62	0.76	0.80	0.90
	RC	0.38	0.43	0.51	0.59	0.71	0.79	0.86
	AUC	0.41	0.45	0.59	0.61	0.69	0.76	0.85
	KA	0.09	0.10	0.39	0.41	0.54	0.65	0.79
Naïve Bayes classifier	AC	0.63	0.65	0.71	0.75	0.82	0.86	0.91
	PC	0.42	0.46	0.62	0.68	0.81	0.86	0.92
	RC	0.27	0.33	0.47	0.49	0.61	0.69	0.79
	AUC	0.33	0.38	0.60	0.62	0.69	0.77	0.84
	KA	0.08	0.13	0.39	0.40	0.56	0.67	0.78
Logistic regression classifier	AC	0.66	0.67	0.71	0.74	0.79	0.85	0.89
	PC	0.48	0.51	0.63	0.70	0.78	0.89	0.93
	RC	0.04	0.22	0.35	0.39	0.53	0.64	0.73
	AUC	0.08	0.31	0.51	0.55	0.62	0.74	0.82
	KA	0.01	0.09	0.21	0.32	0.48	0.64	0.74

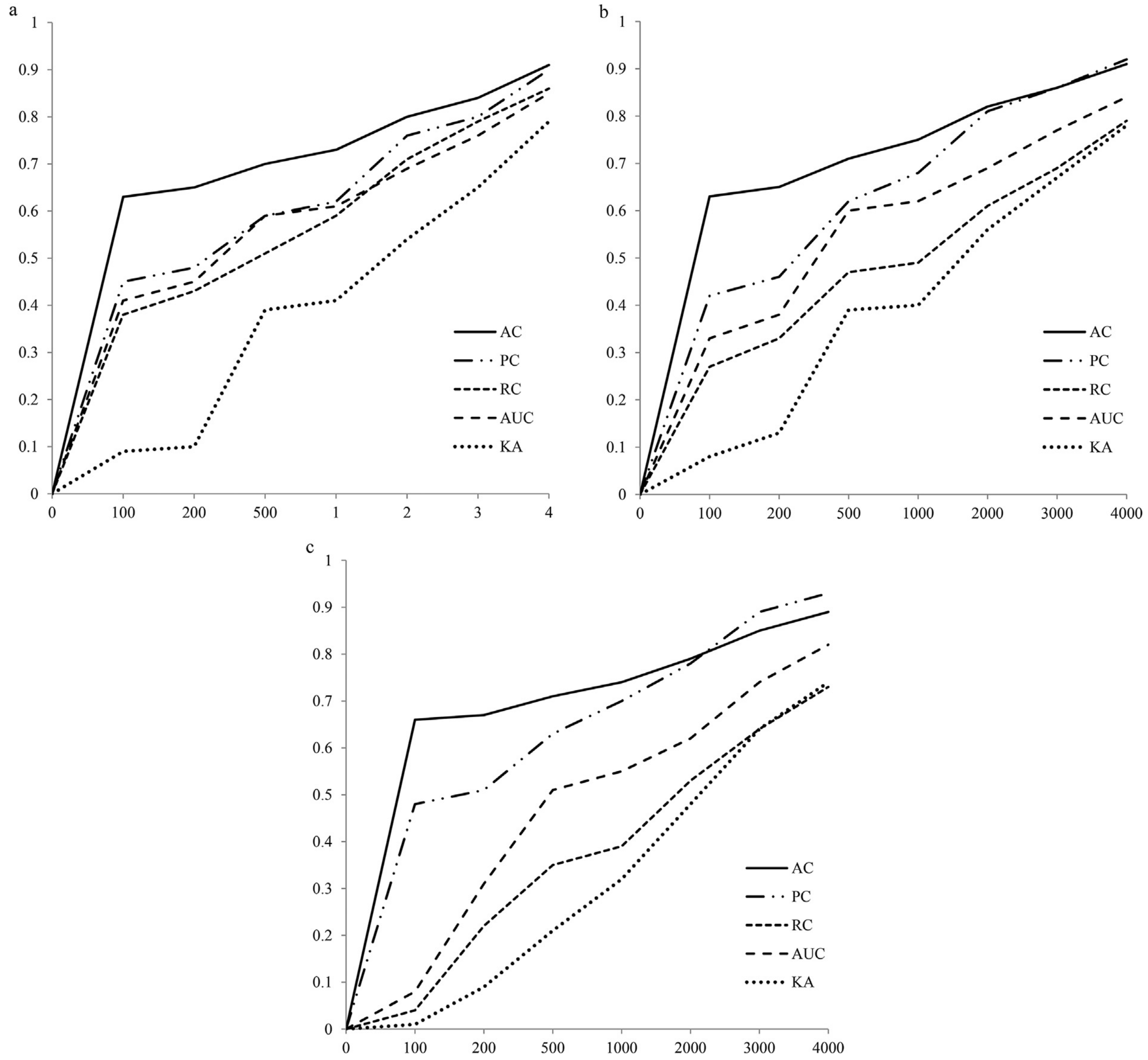

Fig. 1. a. Linear Support Vector Machine classifier, relationship classification performance and number of training tweets, random sampling approach. b. Naïve Bayes classifier, relationship classification performance and number of training tweets, random sampling approach. c. Logistic regression classifier, relationship classification performance and number of training tweets, random sampling approach.

Table 3
Relationship classification performance and number of training tweets, stratified random sampling approach.

		1	2	5	10	20	30	40
Linear support vector machine classifier	AC	0.66	0.67	0.71	0.76	0.80	0.83	0.86
	PC	0.54	0.59	0.66	0.70	0.79	0.82	0.83
	RC	0.41	0.66	0.67	0.76	0.80	0.83	0.87
	AUC	0.46	0.57	0.63	0.67	0.73	0.80	0.82
	KA	0.17	0.30	0.48	0.57	0.68	0.68	0.73
Naïve Bayes classifier	AC	0.66	0.68	0.74	0.78	0.83	0.86	0.87
	PC	0.58	0.59	0.67	0.74	0.81	0.85	0.85
	RC	0.27	0.44	0.56	0.60	0.70	0.75	0.79
	AUC	0.37	0.50	0.61	0.67	0.75	0.80	0.82
	KA	0.14	0.27	0.46	0.55	0.64	0.69	0.72
Logistic regression classifier	AC	0.65	0.67	0.72	0.76	0.81	0.85	0.88
	PC	0.60	0.61	0.67	0.75	0.81	0.86	0.89
	RC	0.15	0.32	0.45	0.51	0.61	0.69	0.76
	AUC	0.24	0.42	0.54	0.60	0.70	0.77	0.82
	KA	0.01	0.19	0.38	0.47	0.58	0.65	0.72

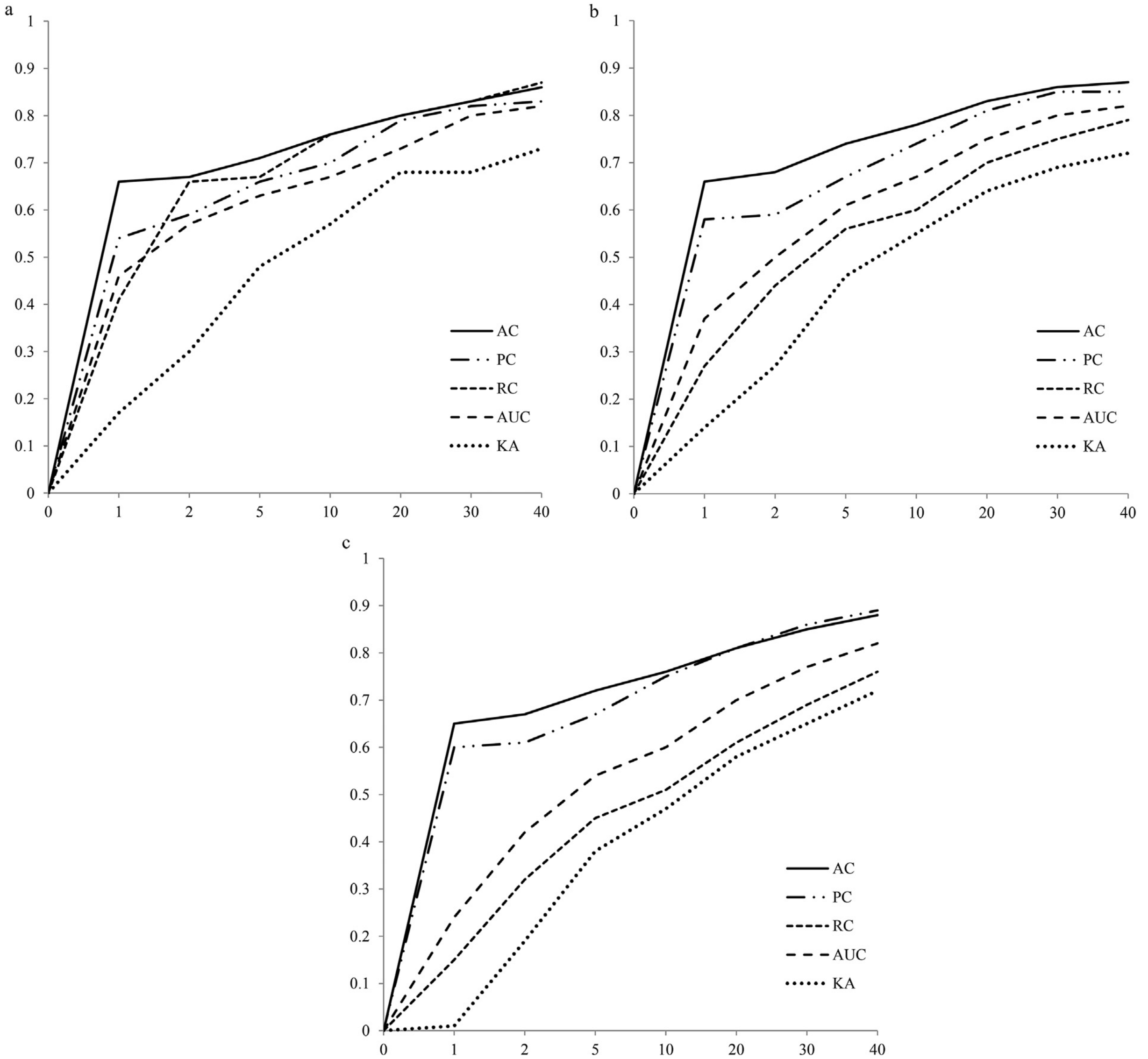

Fig. 2. a. Linear Support Vector Machine classifier, relationship classification performance and number of training tweets, stratified random sampling approach. b. Naïve Bayes classifier, relationship classification performance and number of training tweets, stratified random sampling approach. c. Logistic regression classifier, relationship classification performance and number of training tweets, stratified random sampling approach.

sampling. This can be concluded based on higher values of the AC, PC, AUC, and KA.

4.4. Application of SML

The trained with the best overall performance is now applied to the entire dataset containing 578,803 tweets. The results are compared to the findings reported by van Zoonen and colleagues (2016) based on their manual analysis of a sub-sample of 38,124 tweets. As demonstrated in Table 2 the Linear Support Vector Machine shows the best overall performance on a training-set of 4000 tweets. In effect, this means that around 4000 tweets need to be manually coded to train a well-performing classifier.

The Linear Support Vector Machine classifier was used to code 578,803 tweets. Based on this machine learning procedure, a total of 31.8% (N = 184,060) of the tweets was categorized as work-related (M = 0.318 $S. D.$ = 0.464). Notably, van Zoonen and colleagues (2016) found that the prevalence of work-related tweets was 36.15% (N = 13,783) the analyzed dataset. The accuracy, which is the percent agreement between the human coding and the computer coding is 0.91, in turn the Krippendorf's Alpha is 0.79 (see Table 2), which both indicate satisfactory coder reliability (Lombard et al., 2002).

Furthermore, van Zoonen et al. (2016) identified the interactive features of the tweets and compared whether these features were represented differently in work-related tweets compared to personal tweets. In their sample 69.9% of the tweets included @-mentions, 35.5% hashtags, 33.0% were retweets and 48.5% of the tweets included hyperlinks. The findings of the total sample (N = 578,803) mimic these results as 67.6% of the tweets included @-mentions, 24.2% included hashtags, 24.5% were retweets and 39.7% were tweets that included hyperlinks. Similar to the study by van Zoonen et al. (2016) these features were identified using regular expression, hence there is no measurement error.

Finally, the MANOVA analysis was replicated with interactive features as dependent variables and the work-related content category as independent variable. The findings substantiate the findings by van Zoonen et al. (2016). All variables were binary in which 0 means absent and 1 present. Retweets were more profound in work-related tweets ($M = 0.392$, $S.D. = 0.488$) than in non-work related tweets ($M = 0.179$, $S. D. = 0.038$; $F(1, 578{,}803) = 292, 34$; $p < 0.001$). Similarly, hashtags were more often used is work-related tweets ($M = 0.415$, $S.D. = 0.493$) than in non-work related tweets ($M = 0.311$, $S. D. = 0.462$; $F(1, 578{,}803) = 292, 34$; $p < 0.001$). The same pattern arises when examining the use of hyperlinks ($M = 0.617$, $S.D. = 0.486$), which is higher than in non-work related tweets ($M = 0.299$, $S. D. = 0.458$; F (1, 578,803) = 292, 34; $p < 0.001$). Finally, with respect to @mention the results show that these are less profound in work related tweets ($M = 0.667$, $S.D. = 0.471$) than in non-work related tweets ($M = 0.686$, $S. D. = 0.464$; $F(1, 578{,}803) = 292, 34$; $p < 0.001$). These findings are similar to the findings presented by van Zoonen et al. (see Table 4), demonstrating the value of using SML procedures in analyzing social media content.

5. Discussion

The use of automated content analysis, specifically SML, is insufficiently institutionalized in organizational communication research. This study explored the application of SML to the coding of social media content, specifically tweets. Social media utterances are part and parcel to organizational communication, and SML can advance future research by enabling large-scale content analyses. The findings of SML classifiers in coding social media content are promising. At this point the findings suggest SML can be applied to code social media content in diverse contexts e.g., across a variety jobs and organizations. The findings suggest that SML can overcome specific difficulties associated with dictionary-based coding and manual coding. The application demonstrates that SML coding procedures can be used to code extensive datasets based on a relatively small manually coded training set. This case study shows that 4000 manually coded tweets are sufficient to code as much as 578,803 tweets with satisfactory reliability.

Several findings should be highlighted here. First, the results show that an SML classifier can be trained to automatically code social media content. Three different classifiers Linear Support Vector Machine, Naïve Bayes, and standard logistic regression were compared. The results show that the Linear Support Vector Machine and Naïve Bayes classifier outperform the logistic regression classifier as these yield higher levels of accuracy, precision, recall, precision-recall curve and krippendorff's Alpha overall.

Second, this study shows that the performance of the classifier gradually increases when the training set becomes larger. An important factor in the performance of SML is the overall presence of a category (Burscher et al., 2014). When studying a category that regularly occurs within the data, a smaller training set might be sufficient to build the classifier. When studying an uncommon category, active sampling of positive examples might help to keep manual coding efforts manageable (Burscher et al., 2014; Tong & Koller, 2000).

Third, the results indicate that the Linear Support Vector Machine and Naïve Bayes classifiers perform better based on a stratified random sample of training data. However, as the size of the training samples increase the advantage of stratified sampling over random sampling disappears. In fact, with a training dataset of 4000 tweets the classifiers trained with the random sampling procedure outperforms the classifiers trained with stratified sampling.

Finally, our findings show that the Linear Support Vector Machine reaches satisfactory performance levels when the randomly selected training set contains at least 4000 tweets. For the stratified random sample 20 tweets per user are needed to reach a satisfactory Krippendorf's Alpha (Lombard et al., 2002). The Linear Support Vector Machine appears to yield the highest reliability statistics. This means that the Linear Support Vector Machine can be used to code a potentially limitless dataset of tweets based on a training set of 4000 manually coded tweets. These findings aid the large-scale automated content analysis of social media content in a variety of contexts. Notably, SML adequately replicates the coding decisions

Table 4
Comparison of methods based on MANOVA results.

Interactive features	Original data N = 38,124		Full data N = 578,803	
	ΔMean	Between sub eff.	ΔMean	Between sub eff.
Retweets	0.140	F(1, 38,124) = 292.34 p < 0.000	0.213	F(1, 578,803) = 31.97 p < 0.000
@-mentions	−0.021	F(1, 38,124) = 7.12 p = 0.008	−0.019	F(1, 578,803) = 33.63 p < 0.000
Hashtags	0.079	F(1, 38,124) = 85.05 p < 0.000	0.104	F(1, 578,803) = 42.31 p < 0.000
Hyperlinks	0.102	F(1, 38,124) = 134.87 p < 0.000	0.317	F(1, 578,803) = 160.85 p < 0.000

of human coders and thereby advances computer assisted coding procedures while reducing the human coder investment. These types of big data methodologies do not represent a panacea or a substitute for carefully designed surveys, experiments, and content analyses based on sampling. Instead they represent a complement, an additional resource for better and complete understanding a fast-changing electronic public sphere (Neumann, Guggenheim, Mo Jang, & Young Bae, 2014, p. 210). The SML procedure outlined in this paper permits the analysis of big data in the context social media and organizational communication research.

However, as SML, just like most automated content analysis, also relies on the *bag of words* approach there is room for improvement. The *bag of words* approach uses word frequencies as features of text to inform the analysis and disregard word order (e.g., Hellsten, Dawson, & Leydesdorff, 2010; Miller, 1997). Despite the fact that this approach is widely applied and commonly found to infer substantively interesting features of text (Hopkins & King, 2010), one could critically argue that it fails to provide an accurate account to actually process texts because simply looking at word (co)occurrences considerably reduces the amount of information. Hence, researchers, predominantly in the field of computer linguistics, consistently look for new ways to improve automated tools to content analyze text. Certain other upcoming approaches that are capable of taking more elements of language into account might help to improve SML. Techniques such as part of speech tagging and named entity recognition are examples of tools to move beyond the *bag of word* approach. These techniques take the syntactic structure of sentences into account. For example, Van Atteveldt, Kleinnijenhuis, and Ruigrok (2008) conducted a model to automatically disentangle the syntactical function of the elements of a sentence as well as code semantic relationships. Whilst there will always be room to improve automated analysis in terms of using algorithms in order to understand texts, the SML approach as addressed and applied in this exploratory paper seems to serve its purpose at this point.

SML advances research in the field of organizational communication in today's socially mediated society, as it opens up new possibilities for answering pressing empirical questions in the field of organizational communication. Several scholars suggested that current social media practices in organizations are currently outpacing our empirical understanding (Treem & Leonardi, 2012). Whereas others note that there is an imminent need for more quantitative content analytical research on social media use in the workplace (El Ouirdi et al., 2015). This study contributes to methodological knowledge that aids research on a variety of empirical questions. Social media utterances are part and parcel to organizational communication and SML can advance future research by enabling large-scale content analyses.

5.1. Limitation and future research directions

An important limitation that needs to be acknowledged is that the quality of training data is always subject to the quality of the human coding procedure. If tweets with similar features have different labels, it becomes more difficult for the SML algorithm to estimate a model that can differentiate between the classes.

This study explored the applicability of SML in organizational communication, based on a case study of employees' work related Twitter. Future studies might examine the performance of SML in coding different types of social media content in a variety of topics. For instance, SML can be applied in identifying stakeholders' tone of voice in respect to organizational issues.

This study shows that a trained classifier can be applied to the automatic coding of tweets. It can be anticipated that such an approach is also applicable to other social media content (e.g., Facebook posts) and topics (e.g., sentiment or emotions) as there are many conceptual similarities with the coding of work-related tweets. Future research could build on the SML procedure described in this study as it opens up possibilities for large-scale content analysis. Similarly, the procedure eases the ability to content analyze social media messages allowing researcher to adopt multi-method approaches in their research – for instance by linking survey data and social media content of the same respondents.

References

An, S. K., & Gower, K. K. (2009). How do the news media frame crises? A content analysis of crisis news coverage. *Public Relations Review, 35*(2), 107–112.

Berelson, B. (1952). *Content analysis in communication research*. New York, NY: APA PsycNET.

Bucher, E., Fieseler, C., & Suphan, A. (2013). The stress potential of social media in the workplace. *Information, Communication & Society, 16*(10), 1639–1667.

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

Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015). Using Supervised Machine Learning to Code Policy Issues Can Classifiers Generalize across Contexts? *The ANNALS of the American Academy of Political and Social Science, 659*(1), 122–131.

Campopiano, G., & De Massis, A. (2015). Corporate social responsibility reporting: a content analysis in family and non-family firms. *Journal of Business Ethics, 129*(3), 511–534.

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PloS One, 5*(11), e14118. http://dx.doi.org/10.1371/journal.pone.0014118.

Cho, S., & Hong, Y. (2009). Netizens' evaluations of corporate social responsibility: content analysis of CSR news stories and online readers' comments. *Public relations review, 35*(2), 147–149.

Dreher, S. (2014). Social media and the world of work. *Corporate Communications: An International Journal, 19*(4), 344–356. http://dx.doi.org/10.1108/CCIJ-10-2013-0087.

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 *Advances in web mining and web usage analysis* (pp. 187–206). Berlin Heidelberg: Springer.

Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies: research themes, data sources, and methodological refinements. *Organizational Research Methods, 10*(1), 5–34.

El Ouirdi, A., El Ouirdi, M., Segers, J., & Hendrickx, E. (2015). Employees' use of social media technologies: a methodological and thematic review. *Behaviour and Information Technology, 34*(5), 454–464.

Gallaugher, J., & Ransbotham, S. (2010). Social media and customer dialog management at Starbucks. *MIS Quarterly Executive, 9*(4), 197–212.

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.

Hellsten, I., Dawson, J., & Leydesdorff, L. (2010). Implicit media frames: automated analysis of public debate on artificial sweeteners. *Public Understanding of Science, 19*, 590–608.

Helm, S. (2011). Employees' awareness of their impact on corporate reputation. *Journal of Business Research, 64*(7), 657–663.

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.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Addison-Wesley.

Hopkins, D., & King, G. (2010). Extracting systematic social science meaning from text. *American Journal of Political Science, 54*(1), 229–247.

Huff, A. S. (1990). *Mapping strategic thought*. John Wiley & Sons.

Humphreys, L., Gill, P., Krishnamurthy, B., & Newbury, E. (2013). Historicizing new media: a content analysis of Twitter. *Journal of Communication, 63*, 413–431.

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features* (pp. 137–142). Berlin Heidelberg: Springer.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of Social Media. *Business horizons, 53*(1), 59–68.

Ki, E. J., & Nekmat, E. (2014). Situational crisis communication and interactivity: usage and effectiveness of Facebook for crisis management by Fortune 500 companies. *Computers in Human Behavior, 35*, 140–147.

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research, 30*(3), 411–433.

Landers, R. N., & Callan, R. C. (2014). Validation of the beneficial and harmful work-related social media behavioral taxonomies: development of the work-related social media questionnaire. *Social Science Computer Review, 32*(5), 628–646.

Leftheriotis, I., & Giannakos, M. N. (2014). Using social media for work: losing your time or improving your work? *Computers in Human Behavior, 31*, 134–142. http://dx.doi.org/10.1016/j.chb.2013.10.016.

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: a hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media, 57*(1), 34–52.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: assessment and reporting of intercoder reliability. *Human Communication Research, 28*, 587–604.

Lovejoy, K., & Saxton, G. D. (2012). Information, community, and action: how nonprofit organizations use social media*. *Journal of Computer-Mediated Communication, 17*(3), 337–353.

Lovejoy, K., Waters, R. D., & Saxton, G. D. (2012). Engaging stakeholders through Twitter: how nonprofit organizations are getting more out of 140 characters or less. *Public Relations Review, 38*(2), 313–318. http://dx.doi.org/10.1016/j.pubrev.2012.01.005.

Mangold, W. G., & Faulds, D. J. (2009). Social media: the new hybrid element of the promotion mix. *Business Horizons, 52*(4), 357–365.

Marwick, A. E., & Boyd, D. (2010). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society, 13*(1), 114–133. http://dx.doi.org/10.1177/1461444810365313.

McCorkindale, T. (2010). Can you see the writing on my wall? a content analysis of the fortune 50's Facebook social networking sites. *Public Relations Journal, 4*(3), 1–13.

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

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. London, United Kingdom: MIT press.

Neuman, R., Guggenheim, L., Mo Jang, S., & Bae, S. Y. (2014). The dynamics of public attention: agenda-setting theory meets big data. *Journal of Communication, 64*(2), 193–214.

Ollier-Malaterre, A., Rothbard, N. P., & Berg, J. M. (2013). When worlds collide in cyberspace: how boundary work in online social networks impacts professional relationships. *Academy of Management Review, 38*(4), 645–669. http://dx.doi.org/10.5465/amr.2011.0235.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825–2830.

Reinhardt, W., Ebner, M., Beham, G., & Costa, C. (2009). How people are using Twitter during conferences. creativity and innovation competencies on the web. In *Proceedings of the 5th EduMedia* (pp. 145–156).

Riffe, D., Lacy, S. R., & Fico, F. G. (2005). *Analyzing media messages: using quantitative content analysis in research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Russell, S. J., & Norvig, P. (2002). *Artificial intelligence: a modern approach* (International Edition).

Rybalko, S., & Seltzer, T. (2010). Dialogic communication in 140 characters or less: how Fortune 500 companies engage stakeholders using Twitter. *Public Relations Review, 36*(4), 336–341.

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

Schultz, F., Utz, S., & Göritz, A. (2011). Is the medium the message? perceptions of and reactions to crisis communication via Twitter, blogs and traditional media. *Public Relations Review, 37*(1), 20–27.

Short, J. C., & Palmer, T. B. (2008). The application of DICTION to content analysis research in strategic management. *Organizational Research Methods, 11*(4), 727–752.

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

Sjøvaag, H., Moe, H., & Stavelin, E. (2012). Public service news on the web: a large-scale content analysis of the Norwegian broadcasting corporation's online news. *Journalism Studies, 13*(1), 90–106.

Small, T. A. (2011). What the hashtag? a content analysis of Canadian politics on Twitter. *Information, Communication & Society, 14*(6), 872–895.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437.

Tong, S., & Koller, D. (2000, December). Active learning for parameter estimation in Bayesian networks. In *NIPS* (vol. 13, pp. 647–653).

Treem, J. W., & Leonardi, P. M. (2012). Social media use in organization: exploring the affordances of visibility editability persistence and association. *Communication Yearbook, 36*, 143–189.

Van Atteveldt, W., Kleinnijenhuis, J., & Ruigrok, N. (2008). Parsing, semantic networks, and political authority using syntactic analysis to extract semanticrelations from Dutch newspaper articles. *Political Analysis, 16*, 428–446.

van Zoonen, W., & van der Meer, T. (2015). The importance of source and credibility perception in times of crisis: crisis communication in a socially mediated era. *Journal of Public Relations Research*, 371–388.

van Zoonen, W., Verhoeven, J. W., & Vliegenthart, R. (2016). How employees use Twitter to talk about work: a typology of work-related tweets. *Computers in Human Behavior, 55*, 329–339.

Waters, R. D. (2007). Nonprofit organizations' use of the internet: a content analysis of communication trends on the internet sites of the philanthropy 400. *Nonprofit Management and Leadership, 18*(1), 59–76.

Waters, R. D., Burnett, E., Lamm, A., & Lucas, J. (2009). Engaging stakeholders through social networking: how nonprofit organizations are using Facebook. *Public Relations Review, 35*(2), 102–106.

Waters, R. D., & Jamal, J. Y. (2011). Tweet, tweet, tweet: a content analysis of nonprofit organizations' Twitter updates. *Public Relations Review, 37*(3), 321–324. http://dx.doi.org/10.1016/j.pubrev.2011.03.002.

Young, L., & Soroka, S. (2012). Affective news: the automated coding of sentiment in political texts. *Political Communication, 29*(2), 205–231.

Yu, H. F., Huang, F. L., & Lin, C. J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning, 85*(1–2), 41–75.

Zhang, H. (2004). The optimality of Naıve Bayes. In *Proceedings. 17th international FLAIRS conference, Florida, USA*.