International Journal of Research in Marketing 36 (2019) 20–38

Contents lists available at ScienceDirect

IJRM

International Journal of Research in Marketing

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

Full Length Article

Comparing automated text classification methods

Jochen Hartmann *, Juliana Huppertz, Christina Schamp, Mark Heitmann

Marketing & Customer Insight, University of Hamburg, Moorweidenstraße 18, 20148 Hamburg, Germany

ARTICLE INFO

Article history:
First received on August 16, 2017 and
was under review for 5 months
Available online 24 October 2018

Senior Editor: Michael Haenlein

Keywords:
Text classification
Social media
Machine learning
User-generated content
Sentiment analysis
Natural language processing

ABSTRACT

Online social media drive the growth of unstructured text data. Many marketing applications require structuring this data at scales non-accessible to human coding, e.g., to detect communication shifts in sentiment or other researcher-defined content categories. Several methods have been proposed to automatically classify unstructured text. This paper compares the performance of ten such approaches (five lexicon-based, five machine learning algorithms) across 41 social media datasets covering major social media platforms, various sample sizes, and languages. So far, marketing research relies predominantly on support vector machines (SVM) and Linguistic Inquiry and Word Count (LIWC). Across all tasks we study, either random forest (RF) or naive Bayes (NB) performs best in terms of correctly uncovering human intuition. In particular, RF exhibits consistently high performance for three-class sentiment, NB for small samples sizes. SVM never outperform the remaining methods. All lexicon-based approaches, LIWC in particular, perform poorly compared with machine learning. In some applications, accuracies only slightly exceed chance. Since additional considerations of text classification choice are also in favor of NB and RF, our results suggest that marketing research can benefit from considering these alternatives.

1. Introduction

Online social networks, consumer reviews, and user-generated blog content facilitate personal communication between consumers and consumers as well as firms and consumers (Hewett, Rand, Rust, & van Heerde, 2016). This provides marketing research and practice with additional consumer information that can complement traditional market research (Netzer, Feldman, Goldenberg, & Fresko, 2012). Among other things, social media accelerate public opinion building processes. Accordingly, continuously tracking potential communication shifts in terms of sentiment or other predefined categories becomes increasingly important to enable timely responses. Similarly, marketing researchers are increasingly interested in classifying large volumes of unstructured text data to study how sentiment and theoretically meaningful content classes co-evolve with marketing-relevant outcomes (e.g., Berger & Milkman, 2012; Ghose, Ipeirotis, & Li, 2012).

Whenever dictionaries exist, researchers can apply lexicon-based methods such as Linguistic Inquiry and Word Count (LIWC) to relate word choice to content categories of interest (e.g., Cavanaugh, Bettman, & Luce, 2015; Ordenes, Ludwig, Grewal, & Wetzels, 2017). Alternatively, human coding (e.g., in terms of positive vs. negative sentiment) can be used to train supervised machine learning algorithms to automatically classify any additional data (e.g., Hennig-Thurau, Wiertz, & Feldhaus, 2015; Lee, Hosanagar, & Nair, 2018). All of these approaches attempt to structure unstructured text data by assigning categories to individual text documents. This is particularly relevant whenever comprehensive human coding is not feasible due to the amount of data or because immediate classification information is required.

* Corresponding author.
E-mail address: jochen.hartmann@uni-hamburg.de. (J. Hartmann).

https://doi.org/10.1016/j.ijresmar.2018.09.009
0167-8116/

Industry reports estimate the market volume of such automated text analysis to reach 8.8 billion USD by 2022, with annual growth up to 17.2% (Markets & Markets, 2018). However, according to a survey among 3300 C-level executives, the selection of adequate methods for specific application contexts is regarded as one of the main challenges that currently prohibits further machine learning proliferation (McKinsey Global Institute, 2017).

Balancing different objectives when choosing text classification approaches can be complex. Objectives exclusively related to maximizing classification accuracy suggest comprehensively testing all available approaches to identify the best solution for each individual task. On the other hand, taking comparability and scarce resources into account employing the same approach across applications can be reasonable. Choosing between both extremes requires knowledge about the size of the potential accuracy trade-offs and their monetary consequences.

Prior research provides little guidance on these issues. While in marketing, method comparisons of text classification are scarce, several such comparisons exist in computer science. However, these publications have different objectives. Even small improvement increments are of interest, e.g., to understand promising avenues for further developments. For these reasons, method comparisons are often limited to a few new method candidates and reference methods. Accordingly, the empirical evidence on predictive accuracy is scattered across publications with different types of data and implementations. These comparisons also often lack statistical tests of significance and economic relevance or investigations on the practical relevance of the observed performance differences (e.g., Bermingham & Smeaton, 2010; Pang, Lee, & Vaithyanathan, 2002). In terms of data, this literature studies diverse datasets, many of which are only of peripheral interest to marketing, such as classifying political blogs (e.g., Melville, Gryc, & Lawrence, 2009) or medical abstracts from bibliographic databases (e.g., Joachims, 1998). A particularly often mentioned conclusion is the no free lunch theorem, i.e., no single method works best across all applications and each application requires an exhaustive method comparison to find the optimal approach for the task at hand (e.g., Fernández-Delgado, Cernadas, Barro, & Amorim, 2014; Wolpert, 1996).

Social media marketing covers a smaller and likely more homogenous set of text classification problems. In addition, marketing is particularly interested in economic relevance, interpretability of results, and implementational costs. Among other things, empirical observations are relevant for theory building, which necessitates comparable results across studies in terms of similar methodological approaches, parameters, and their interpretations. Moreover, text classification often provides only a few variables as part of more comprehensive econometric models (e.g., Hewett et al., 2016). This makes implementational costs relevant and favors repeated application of well-established approaches. Against this background, it is not clear that the no free lunch theorem advocated in computer science is similarly reasonable advice for marketing.

The particular concern of marketing research with efficiency in application and comparability in terms of results is evidenced by the fact that marketing research has gravitated towards two main classification approaches used across applications: support vector machines (SVM) and LIWC. With very few exceptions (e.g., Netzer, Lemaire, & Herzenstein, 2016), marketing research does not conduct the types of exhaustive method comparisons computer science would suggest. Repeated applications of easily implementable and interpretable methods such as LIWC can appear reasonable considering the aforementioned text classification objectives. However, the trade-offs in terms of accuracy require further research. In particular, it is not clear whether an individual approach exists that performs consistently well and within a reasonable boundary compared with the top performing approaches. It is also not clear under which circumstances which methods are most likely suitable.

This research attempts to fill this gap. We compare the performance of SVM and LIWC to other approaches tested particularly often for text classification outside of marketing. This includes artificial neural networks (ANN), k-nearest neighbors (kNN), naive Bayes (NB), and random forest (RF) as well as four additional lexicon-based methods. We study how well these automated text classification methods represent human intuition across a collection of 41 social media datasets, covering different sample sizes, languages, major social media and ecommerce platforms as well as corporate blogs (e.g., Facebook, Twitter, IMDb, and YouTube).

To the best of our knowledge, only one study is similar in spirit to our investigation. Kübler, Colicev, and Pauwels (2017) also investigate method performance for applied marketing problems. They focus on SVM and LIWC as the prevailing methods in marketing. Their investigation is based on a single social network and studies whether an individual firm should rely on these specific classification methods for sentiment extraction and consumer mindset prediction. Our research complements their work by including additional methods (i.e., ANN, kNN, NB, and RF as well as four additional lexicon-based methods) and tests how SVM and LIWC perform relative to these and which classification method is best suited given the specific tasks.

We have found no other comparative study investigating a similar scope of social media datasets and methods. This allows us to explore a potential middle ground between the no free lunch theorem and treating all classification problems with the same (simple) approach. In particular, we can study the variance of method performance across methods and datasets to understand what drives accuracy and under which conditions which methods perform best. This allows making more informed method choices without requiring full method comparisons for each application.

2. Related research

2.1. The use of automated text classification in marketing research

We identified marketing publications applying automated text classification by searching relevant marketing journals (i.e., JM, JMR, Mrkt. Sci., JCR, IJRM, Mgnt. Sci., JAMS), for papers that mention at least one of the methods we study in their titles, abstracts, or keywords or explicitly state the application of automated text classification. We also conducted a keyword search regarding these methods as well as text classification and screened the websites of the authors we identified. Note, we may still have missed

individual publications since text classification sometimes provides only a single variable of a more comprehensive empirical analysis and is consequently only briefly mentioned in some articles. However, to the best of our knowledge we cover a representative majority of relevant publications (see Web Appendix A for a detailed list).

These studies are typically geared towards substantive contributions with comprehensive method comparisons beyond their scope. Automated text classification is used in marketing research across a wide range of very different research objectives, e.g., to predict defaults based on online loan requests (Netzer et al., 2016), to elicit customer preferences (Huang & Luo, 2016), to optimize search personalization (Yoganarasimhan, 2018), to forecast movie demand (Hennig-Thurau et al., 2015) and customer engagement (Lee et al., 2018), to understand the link between consumer sentiment and business outcomes (Hewett et al., 2016), or to model stock market returns based on review sentiment (Tirunillai & Tellis, 2012). Interestingly and despite these diverse research objectives, >70% of the publications rely on either SVM or LIWC, with LIWC being twice as popular as SVM. While such implicit conventions might benefit comparability of results, more than a third of the studies mention no rationale for their method choice. Many of the remaining ones refer to successful applications in previous publications as a rationale, in particular when applying SVM or LIWC.

Notably, many important factors such as the classification task, data source or average text length vary fundamentally across studies. This strong reliance on such a limited set of methods suggests that computer science research on text classification performance has not been guiding text classification choices. Marketing rather appears to follow the implicit assumption that an individual method is similarly effective across applications. Regarding the dictionary LIWC in particular, it is conceptually not clear whether simple word counts can deal with complex figures of speech, e.g., litotes such as "… *is really not bad*", or differences in meaning across domains (e.g., high product quality being positive, high blood pressure negative). Consequently, classification accuracy must not coincide with popularity of LIWC in marketing research.

Computer science has taken a different direction. Here, lexicon-based approaches are less popular while additional machine learning approaches play a stronger role. We will review this research next.

2.2. Evidence from comparative studies in computer science

Computer science publications collectively cover a more diverse set of classification approaches. Due to their focus on methodological advancements such as algorithm improvements (e.g., Melville et al., 2009; Ye, Chow, Chen, & Zheng, 2009) or feature engineering (e.g., Bermingham & Smeaton, 2010; Neethu & Rajasree, 2013), individual publications compare only subsets of the relevant methodology and focus on a limited set of reference methods. For similar reasons, these publications typically demonstrate the effectiveness of novel approaches for a single or only a few datasets (e.g., Fang & Zhan, 2015).

Despite these limitations, several conclusions emerge from this work (see Web Appendix B for a detailed list): First, recent computer science studies also apply SVM, but in addition rather suggest ANN and NB than LIWC as the best possible option for text classification. Approaches such as RF, which are suggested in this literature, have not been applied in the marketing publications we have been able to identify. Few compare these classifiers to lexicon-based methods, which are frequently used in marketing. In addition, some methods such as ANN, RF or kNN are less often compared with other text classifiers, perhaps because ANN and RF have more recently been introduced for text classification. The empirical evidence based on certain conditions does suggest that ANN and RF can be particularly effective. We include ANN and RF in our analysis also because these methods often achieve superior performances outside text classification.

Second, the results across different studies vary in terms of the top performing method, suggesting that a single method such as SVM is unlikely to work best across application contexts. For example, Annett and Kondrak (2008) find that NB performs better than SVM, whereas Pang et al. (2002) arrive at the opposite conclusion. Although several authors propose that method performance might be dependent on specific properties of the dataset like its length (e.g., Bermingham & Smeaton, 2010) or the sample size (e.g., Dumais, Platt, Heckerman, & Sahami, 1998; Ye et al., 2009), empirical tests of these conjectures are scarce and multivariate estimates across diverse application contexts do not exist.

Third, prior research has focused on either sentiment or content classification tasks. In applied marketing settings, content classes are often identifiable by strong single signal words, while sentiment is most frequently expressed in more subtle and complex ways (e.g., involving irony or sarcasm), which can require a deeper understanding of the social media text. In their paper, Pang et al. (2002) argue that content might be simpler than sentiment classification, although they do not test any content classification tasks themselves. To reveal the potential consequences given these differences, we include both sentiment and content classification tasks in our analysis.

3. Automated text classification methods

3.1. Conceptual overview of text classification

We distinguish between sentiment and content classification, which are both of particular interest for marketing applications. The former involves predicting the emotion of an unlabeled text document such as the following exemplary movie review, drawn from one of our datasets: "*All in all, a great disappointment*". In sentiment classification, the goal of the text classification methods would be to detect the emotion conveyed through this text and to correctly classify it as negative. Lexicon-based or supervised machine learning methods are the two major approaches to accomplish this task (see Appendix A for an overview of text classification problems and approaches).

Supervised machine learning methods learn either sentiment or custom content categories based on manually labeled text data and inductively construct classifiers based on observed patterns without requiring manual coding of classification rules (Dumais et al., 1998). This makes them flexible in understanding grammatical construction specific to certain domains. In comparison, lexicon-based methods (e.g., NRC or LIWC) require expert-crafted dictionaries, consisting of elaborate word lists and associated labels to classify text documents (Mohammad & Turney, 2010; Pennebaker, Boyd, Jordan, & Blackburn, 2015). These are often generic across domains but can be extended by custom word lists. If no suitable dictionary is available, researchers must create their own (e.g., Hansen, Kupfer, & Hennig-Thurau, 2018). As the creation of such dictionaries is cumbersome, lexical methods such as LIWC are most commonly used for two-class sentiment classification because several off-the-shelf dictionaries exist (e.g., Hennig-Thurau et al., 2015; Ordenes et al., 2017). While these methods are quick and easy to employ, they also come with drawbacks. For example, LIWC may struggle to correctly predict the negative sentiment of a post like the previous example, as the individual words "*great*" and "*disappointment*" point in two opposite emotional directions unless such phrases are included in the dictionary. In contrast, machine learning methods can learn that the word pair "*great disappointment*" indicates negative sentiment without the need to curate a dictionary a priori.

In comparison, content classification refers to the task of assigning custom category labels to new text documents, e.g., to automatically detect that YouTube comments such as "*Subscribe to my channel*" have commercial rather than user interest background. Such tasks are potentially easier than the extraction of emotion, which often requires higher context understanding, e.g., irony, playing a larger role (Das & Chen, 2007).

In contrast to all other methods we study, latent Dirichlet allocation (LDA, Blei, Andrew, & Jordan, 2003) is an unsupervised machine learning method originally developed and applied for knowledge discovery purposes (Humphreys & Wang, 2017). In marketing research, LDA is most commonly used for explorative topic modeling or latent topic identification (e.g., Puranam, Narayan, & Kadiyali, 2017; Zhang, Moe, & Schweidel, 2017). Such types of analyses have different objectives and are conceptually and empirically not comparable to the remaining methods in terms of accurately recovering researcher-defined class labels and are therefore beyond the scope of this investigation.

3.2. Algorithmic approaches and characteristics of text classification methods

In total, we test a set of ten text classification methods due to their conceptually different algorithmic approaches, their use and relevance for marketing research, and their proven performance in other disciplines. This includes five machine learning methods, i.e., ANN, kNN, NB, RF, and SVM, as well as five lexicon-based methods, i.e., AFINN (Nielsen, 2011), BING (Hu & Liu, 2004), LIWC (Pennebaker et al., 2015), NRC Emotion Lexicon (Mohammad & Turney, 2010), and Valence Aware Dictionary for Sentiment Reasoning (VADER, Hutto & Gilbert, 2014). While ANN, RF, and SVM are discriminative classifiers, NB is a generative, probabilistic classifier. In contrast, kNN is a non-parametric classifier, belonging to the family of proximity-based algorithms. We explain each of these approaches in more detail next.

ANN are the most highly parametrized method in our comparison. Neurons, which are connected to the input layer, inductively learn patterns from training data to allow predictions on test data (Efron & Hastie, 2016). The simplest form of ANN consists of only one input and output layer (perceptrons). The number of units in the output layer corresponds to each of the possible classes. Current computational capabilities enable the inclusion of multiple hidden layers in between (e.g., LeCun, Bengio, & Hinton, 2015; Sebastiani, 2002). The number of nodes in the hidden layer is linked to the complexity of the classification task (Detienne, Detienne, & Joshi, 2003). As common text classification problems represent linearly separable class structures in high-dimensional space (Aggarwal & Zhai, 2012), single-layer ANN with a non-linear activation function are most frequently applied for text classification (e.g., Moraes, Valiati, & Neto, 2013).

Due to their flexible structure, ANN can be considered particularly versatile, performing well across different classification tasks, which is likely relevant when handling noisy social media data. Moreover, ANN can learn subtle text patterns. This can be important for sentiment problems, where the link between individual word features and the class may be more complex compared with content classification tasks. However, this ability to adapt to even contradictory data and potentially better recognition of higher context tends to negatively affect the computational costs of ANN (Sebastiani, 2002). The more complex the network typology, the higher the computational time both in the training and prediction phase. While RF can be easily parallelized, ANN are more difficult to multi-thread, posing a larger optimization problem. Moreover, given their large number of parameters and complex structure, ANN are difficult to interpret intuitively and require expert knowledge for parameter tuning.

kNN is a lazy learning algorithm with no offline training phase (Yang, 1999). All training documents are stored and computation is deferred to the prediction phase (Sebastiani, 2002). For each test document, kNN ranks the nearest neighbors of the labeled examples from the training set and uses the categories of the highest-ranked neighbors to derive a class assignment. The more near neighbors with the same category, the higher the confidence in that prediction (Yang & Liu, 1999).

Computing the respective distances between all test and training documents makes kNN computationally costly when applied to high-dimensional, sparse text data (Aggarwal & Zhai, 2012), especially if the training set is large (Sebastiani, 2002). Moreover, as a non-parametric method, kNN suffers from the curse of dimensionality (Bellmann, 1961), requiring an exponentially larger number of training examples to generalize well for many features. This makes kNN prone to overfit in-sample and predict poorly out-of-sample. Thus, relative performance of kNN is likely lower for longer texts with many features and, in turn, more favorable relative to all other methods for shorter texts.

NB is one of the simplest probabilistic classifier models (Yang, 1999). The classifier estimates a class-conditional document distribution $P(d|c)$ from the training documents and applies Bayes' rule to estimate $P(c|d)$ for test documents, where the documents are modeled using their terms. To efficiently compute the conditional probabilities, NB assumes all features to be independent. This naïve assumption can provide a reasonable trade-off between performance and computational costs. Domingos and Pazzani (1997) find that NB can also perform well when features are interdependent. In addition, Netzer et al. (2016) argue that the resulting generative model is easy to interpret and explain. Moreover, NB as a generative classifier may be recommended for smaller sample sizes due to its inherent regularization, making it less likely to overfit compared with discriminative classifiers (e.g., Domingos, 2012; Ng & Jordan, 2002). However, NB is not capable of modeling interaction effects among features. Thus, we expect it to perform relatively well for problems with strong individual signal words and straightforward relationships between the text features and the respective classes, e.g., for simple forms of promotion content detection (Yang, Nie, Xu, & Guo, 2006) and two-class sentiment classification exhibiting strong polarity.

RF is an ensemble learning method that grows a multitude of randomized, uncorrelated decision trees (Breiman, 2001). Each decision tree casts a vote for the class of the test example. The most popular class determines the final prediction of the RF classifier. This procedure is called bagging (Breiman, 1996). The larger the number of predictors, the more trees need to be grown for good performance. There are different ways to introduce randomness and decorrelate the individual decision trees, e.g., through random feature selection and randomly chosen data subsets (Breiman, 2001). While individual decision trees are prone to overfitting due to their high flexibility (Domingos, 2012; Sebastiani, 2002), RF overcomes this issue by combining a multitude of decision trees on a heterogeneous randomly drawn subset of variables.

As RF is more robust to noise and outliers (Breiman, 2001), we expect consistently high performance across all social media datasets. Moreover, given their hierarchical structure, RF can learn complex interactions between features, perform automatic feature selection, and model highly non-linear data. This leads us to believe that RF can deal well with both content and more complex sentiment classification, where higher context understanding is required, as signals are subtly embedded in the text and spread across features. Lastly, the training time of RF increases linearly with the number of decision trees in the ensemble. As each tree is grown individually, processing can be easily parallelized. This makes RF scalable and computationally efficient, enabling quick training of classifiers.

SVM are discriminative classifiers, fitting a margin-maximizing hyperplane between classes. They were initially developed as binary linear classifiers (Cortes & Vapnik, 1995), but can be extended to non-linear problems of higher dimensionality through the use of kernels that can accommodate any functional form (Scholkopf & Smola, 2001). Unlike other classifiers with higher capacity to fit the training data, SVM are less likely to overfit and generalize better (Bennett & Campbell, 2000). Following research convention, we study linear classifier kernels since they represent the most common application in text mining (e.g., Boiy, Hens, Deschacht, & Moens, 2007; Pang et al., 2002; Xia, Zong, & Li, 2011). The margin-maximizing hyperplane is determined solely by the support vectors (Sebastiani, 2002). Beyond determining the position of the discriminant plane, the support vectors carry only limited information (Bennett & Campbell, 2000). Computing the parameters of the margin-maximizing hyperplane poses a convex optimization problem (Moraes et al., 2013), a task that can be computationally costly depending on the sample size and number of features.

SVM have been shown to be effective for certain text problems such as news article categorization and sentiment prediction (e.g., Joachims, 1998; Pang et al., 2002), as they can deal well with high dimensionality (Bermingham & Smeaton, 2010; Wu et al., 2008). However, their limited representation may result in a lack of ability to model nuanced patterns in the training data (Domingos, 2012). At the same time, SVM have been argued to be less prone to overfitting (Joachims, 1998). Therefore, we expect SVM to perform similarly to a simple method like NB, but worse than more flexible methods like ANN and RF.

In addition to the supervised machine learning methods, we investigate the performance of five lexicon-based methods for sentiment classification. First, LIWC, counts words belonging to a linguistic category (Pennebaker et al., 2015). For this task, LIWC uses manually created dictionaries that identify words in texts and assigns labels based on word frequencies per document. Typically, simple ratios (e.g., share of words with positive or negative emotion) or count scores (e.g., number of words) are computed based on this. The exemplary user expression, drawn from a movie review dataset "*I loved it, it was really scary*" has a total word count of seven words. Thereof, one is counted as positive, i.e., "*loved*", and one is counted as negative, i.e., "*scary*". In two separate columns LIWC would report $1/7 = 14.3\%$ for both the positive and negative word ratio.

Second, we analyze the performance of NRC, which includes both unigrams and bigrams with a few recent applications in marketing research (e.g., Felbermayr & Nanopoulos, 2016). Third, we test VADER, a dictionary specialized on microblog content, incorporating special characters such as emoticons, emojis, and informal language (Hutto & Gilbert, 2014). Forth, we include AFINN, a dictionary developed by Nielsen (2011), dedicated to analyzing microblog texts and emphasis on acronyms such as "*LOL*" and "*WTF*". Lastly, we analyze BING, a labeled list of opinion adjectives constructed to detect the emotional orientation of customer reviews (Hu & Liu, 2004). The dictionary can cope with misspellings and morphological variations through fuzzy matching.

All dictionaries are simple to employ and will likely provide best results for texts following a stringent train of thought with strong emotion-laden signal words. However, for noisy social media texts with a high degree of informality and netspeak, we expect LIWC to perform relatively poorly compared with dictionaries such as VADER and BING, which are specialized on informal texts and include larger lexica. Moreover, all lexicon-based methods' classification accuracies may suffer from shorter texts, as this reduces the probability of matching words from texts to off-the-shelf dictionaries. Additionally, reviews that point in different emotional directions as the example above pose a challenge for all dictionary methods.

Table 1
Dataset descriptions.

Classification task	Social media type	Source (publicly available at)	ID	Authors	Language	Avg. words/document	Max. sample size[1]	# features	Majority class share	# classes (DV)
Sentiment	Product review titles	Amazon (McAuley et al., 2015 for EN)	AMT	UGC	DE/EN	3/5	3000	161/239	0.50	2 (pos, neg)
	Product reviews	Amazon (McAuley et al., 2015 for EN)	AMR	UGC	DE/EN	92/82	3000	3117/3374	0.50	2 (pos, neg)
	Movie reviews	IMDb (Kotziats, Denil, De Freitas, & Smyth, 2015)	IMD	UGC	EN	15	1000	557	0.50	2 (pos, neg)
	Restaurant reviews	Yelp (Kotziats et al., 2015)	YEL	UGC	EN	11	1000	480	0.50	2 (pos, neg)
	Social network comments	Facebook	FBK	UGC	DE	13	3000	549	0.33	3 (pos, neg, neu)
	Corporate blog comments	Fortune 500 blogs	CBC	UGC & firm	DE	36	2942	1274	0.58	3 (pos, neg, neu)
	Microblog posts	Twitter	TWS	UGC & firm	EN/ES/DE	10/11/9	3000	349/339/330	0.58	3 (pos, neg, neu)
Content	Social network comments	YouTube (Alberto, Lochter, & Almeida, 2015)	YTU	UGC	EN	17	1000	624	0.50	2 (promotion, user communication)
	Text messages	Telecom provider (Almeida, Gómez Hidalgo, & Yamaki, 2011)	SMS	UGC	EN	19	1000	861	0.50	2 (promotion, user communication)
	Movie reviews	Rotton Tom. & IMDb (Pang & Lee, 2004)	ROT	UGC	EN	22	3000	855	0.50	2 (subjective, objective)
	Corporate blog posts	Corporate blogs	CBP	Firm	EN	344	1000	10,170	0.54	3 (high, med, low storytelling score)
	Microblog posts	Twitter	TWC	UGC & firm	EN	10	3000	358	0.55	3 (emotion, information, combination)

Note: # features for N = 1000. 1: The maximum TWS sample sizes for ES and DE are 1000.

Given their methodological diversity, we expect relatively low inter-method correlations in terms of accuracy. This would be in line with the no free lunch theorem (Wolpert, 1996). Having said this, research on other applications beyond text classification suggests ANN and RF to be among the top performing methods given their versatile structures. NB is expected to perform well for smaller sample sizes. For larger samples sizes, we expect better performances of machine learning methods, as more data has substantial impact on the ability to identify the different types of expressions contained in a particular dataset more comprehensively. In contrast, lexicon-based methods by definition do not benefit from additional data.

For content classification, accuracies are likely to be higher than for sentiment classification, as there is a clearer link between the word features and the respective class. Overall, we expect the highest performance for two-class content classification, as this poses the conceptually simplest task. Obviously, as the number of classes increases, classification becomes more challenging. In addition, all methods are likely to suffer from noisy data, e.g., in terms of netspeak. In contrast, data carrying strong signals, e.g., adjectives, are likely to produce better results.

4. Research design and methodology

4.1. Data collection

To understand whether these conceptual differences materialize in applied social media settings, we compare the text classification methods on 41 different social media datasets covering different sample sizes, languages, and platforms. Specifically, we have obtained three sample sizes for nine social media types (500; 1000; 3000) and analyze seven social media types in two sample sizes (see Table 1). A relevant driver of predictive accuracy is the amount of available training data, which requires human coding. We chose these sample sizes for two reasons. First, they pose a reasonable effort in terms of manual annotation. Second, similar ranges have been used in previous comparative studies (for example, Pang et al., 2002 with a dataset of 700 positive and negative reviews). Alternatively, some researchers suggest transfer learning, i.e., using labeled data from other domains (e.g., Kübler et al., 2017). However, this did not produce better results on our applications (see Web Appendix C).

The text classification problems we study represent a large variety of real-world marketing tasks in social media. Specifically, we work with short posts from microblogs (Twitter) and social networks (Facebook), short text messages, extended discussion posts from 14 different Fortune 500 blogs, product reviews and their titles from an online shop (Amazon) as well as restaurant (Yelp) and movie reviews (IMDb, Rotten Tomatoes). Our data contains both firm- and user-generated communication in three different languages, representing both colloquial and formal language.

Review titles (AMT) with an average of three to five words per document contribute the shortest texts, whereas corporate blog posts (CBP) contain the longest texts with an average of 344 words. The number of features per dataset, which is proportional to the number of unique words, varies substantially (i.e., 161 compared with >10,000 for AWT and CBP, respectively). For the sake of consistency and better comparability, we report the number of features for $N = 1000$ for all social media types although maximum sample sizes differ depending on the data source. All datasets but CBP contain user-generated content, representing typical social media application settings (e.g., understanding public sentiment).

The English microblog dataset is used for two classification tasks. Specifically, we manually code sentiment (positive, negative, neutral) and content (emotion, information, combination of both) following previous research in marketing (e.g., Akpinar & Berger, 2017). In addition, we analyze comments from Facebook, which exhibit similar text characteristics. For both, we include the neutral class, which in most real-world problems cannot be neglected (Go, Bhayani, & Huang, 2009). While one might expect higher accuracies for texts from corporate blogs compared with microblog posts, such long posts can be challenging due to high levels of information and low noise density (Bermingham & Smeaton, 2010) and therefore can be difficult to annotate even for experienced human coders (Melville et al., 2009).

From our 12 social media types, seven have been used in prior publications and are made publicly available, including four from the UCI repository and one from Pang and Lee (2004). The English Amazon reviews, i.e. AMT and AMR, have been sampled out of >142 million reviews from 1996 to 2014 and originate from McAuley, Targett, Shi, and Van Den Hengel (2015). To understand whether methods are capable of inferring quality assessments from unstructured texts, we transform star ratings to two classes, combining reviews with less than three stars to a negative class and more than three stars to a positive class and excluding reviews with three stars following prior research (e.g., Moraes et al., 2013). Web Appendix D lists representative text examples for all datasets.

4.2. Preprocessing, document representation, and method specification

Text classification methods are typically applied to preprocessed datasets, since raw text data can contain high levels of noise such as typographical errors as frequently observed on social media (e.g., Aggarwal & Zhai, 2012). The most frequently applied steps include tokenization, case transformation, stop-word removal, term weighting, stemming, and building n-grams (Yang, 1999). The goal is to eliminate all non-informative features, which do not contribute to the underlying text classification task. This not only yields better generalization accuracy but also reduces issues of overfitting (Joachims, 1998). All detailed preprocessing steps are summarized in Web Appendix E.

4.3. Performance assessment

We evaluate the performance of each text classification method based on its ability to develop a similar understanding as a human reader, since many marketing applications intend to capture the information communicated to other professional or non-professional users. As our primary performance measure, we compare prediction accuracies to understand how well each method represents human intuition. The accuracy on dataset i is defined as the sum of all correct predictions on the hold-out test set divided by the sum of all predictions.

To obtain an unbiased estimate of out-of-sample accuracy, we split each dataset into a training set (80% of the data) and a hold-out test set (20% of the data). All accuracies reported in this paper are based on predictions on the hold-out test set (see Ordenes et al., 2018 for a similar approach). Importantly, none of the methods could learn from this data during the training phase, producing unbiased performance estimates for classifier effectiveness (Domingos, 2012). If a method overfitted on the training data, it is expected to generalize poorly on the hold-out test data, producing worse accuracies than other methods with better regularization.

Using five-fold cross-validation, we tune the most important parameters for each method on the training set (see Web Appendix E for further details). This means that each dataset is partitioned into five equal training and validation subsets. The goal of this grid search procedure is to test a reasonable set of parameter values to identify the best model for a given task based on validation accuracy. This approach is computationally more complex than simply using the default parameter values of each method across all classification scenarios. However, parameter tuning is necessary because default values must not be appropriate for all individual applications and this can vary across methods. The parameter values producing the best average accuracy across all five folds are used to fit each model on the entire training set. Lastly, we use the tuned models to make predictions on the hold-out test set and report those accuracies for each dataset.

To obtain standard errors, we run five times repeated ten-fold cross-validation for each tuned model on the entire data, producing 50 accuracies for each method. This allows us to conduct two-tailed t-tests on the mean paired accuracy differences (see Moraes et al., 2013 for a similar approach). Web Appendix E presents the R packages we have used. Web Appendix F describes an exemplary R script containing all steps for running the machine learning methods.

5. Results

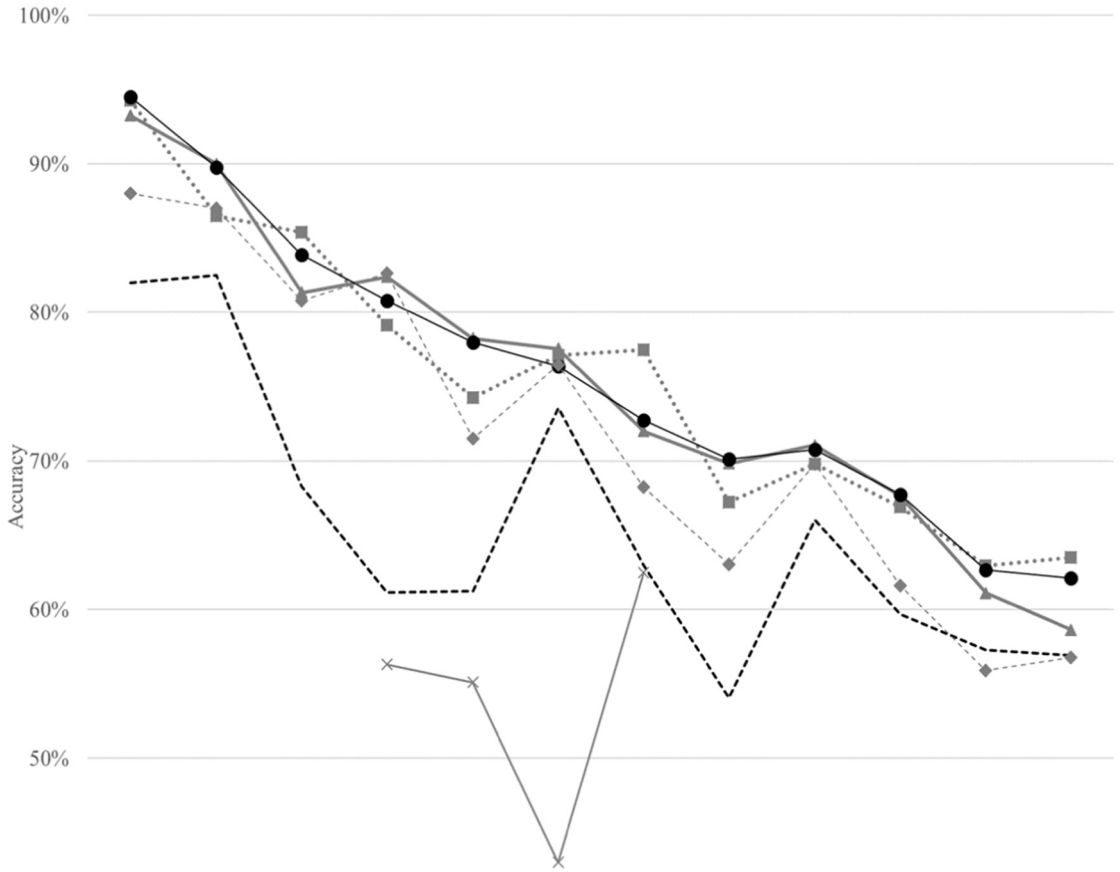
5.1. Comparison of method performances across all datasets and classification tasks

To facilitate interpretation and in the interest of parsimony, we group all similar datasets and compare similar sample size resulting in 12 distinct types of social media text data. Specifically, we aggregate across languages since we do not detect a significant impact of language on classification performance for any method. Fig. 1 summarizes the resulting accuracies for all methods. The performance of the lexicon-based methods is reported for all two-class sentiment problems. Only two out of the five dictionaries perform slightly better than the weakest machine learning algorithm (kNN) and that occurs only in three instances. However, these few instances where one of the dictionary approaches exceeds a machine learning method is due to the poor performance of kNN for these datasets having >15% lower accuracy than the best performing approach. Since none of the dictionaries achieve a performance close to the winning approaches, we summarize these as the average performance in Fig. 1 and provide all details in Appendix B.

Fig. 1 suggests several conclusions. First, there are large differences in maximum accuracies between the easiest task, i.e., promotion detection of short text messages (SMS) at 94.5%, and the most difficult task, i.e., sentiment prediction of user-generated comments to corporate blog posts (CBC) at 63.5%. In line with conventional wisdom, this implies that some dependent variables are easier to predict than others. Second, the performance spread across the five machine learning methods varies across the different data sources. While the difference between the best and worst method for sentiment classification of Amazon review titles (AMT) is only 4 percentage points between ANN and kNN, the difference increases to >21 percentage points between the two methods for Amazon review texts (AMR). Nevertheless, across all different contexts, ANN, NB, and RF consistently achieve the highest performances.

Comparing the absolute average performance across all datasets, the results reveal that the winning methods produce the highest accuracies for two-class content classification. The top three social media types, i.e., SMS, YTU, and ROT, all belong to this kind of classification task. Evidently and due to chance alone, two-class classification has a higher likelihood of correct classifications than three-class classification. In addition, for content classification, individual words tend to be more predictive of the correct class compared with sentiment classification, where the signals are often more subtly embedded in the text. The five lowest accuracies are produced for three-class problems, including both sentiment and content classification, i.e., FBK, CBP, TWS, TWC, and CBC. All but CBP represent user-generated content; among those are two from Twitter and one from Facebook, exhibiting the highest degree of noise. This likely exerts a deteriorating effect on all methods' performances.

Comparing the relative performance across classifiers for significant differences ($p < .05$), RF and NB are among the best performing methods for 11 out 12 social media types (see Fig. 1). This is consistent with conceptual conjectures. Breiman (2001), for example, argues that RF is particularly robust to noise and outliers. This may explain its relatively good performance even on the noisiest text corpora from Twitter, i.e., TWS and TWC. Interestingly, the relatively simple approach of NB is equally



	SMS	YTU	ROT	AMR	YEL	AMT	IMD	FBK	CBP	TWS	TWC	CBC
ANN	93.3	90.0°	81.3	82.4°	78.3°	77.6°	72.0	69.8°	71.0°	67.7°	61.1	58.6
kNN	82.0	82.5	68.3	61.1	61.3	73.6	63.0	54.0	66.0	59.7	57.3	56.9
NB	94.3°	86.5	85.4°	79.1	74.3	77.1°	77.5°	67.2	69.8	66.9°	63.0°	63.5°
RF	94.5°	89.8°	83.9	80.8	78.0°	76.4	72.8	70.1°	70.8°	67.7°	62.7°	62.1
SVM	88.0	87.0	80.8	82.6°	71.5	76.4	68.3	63.0	69.8	61.6	55.9	56.8
DICT				56.3	55.1	43.0	62.5					

Fig. 1. Accuracies of automated text classification in reflecting human intuition across 12 social media types. Note: ° indicate insignificant differences between the best methods ($p > .05$). DICT is the average of five lexicon-based methods, i.e., LIWC, NRC, AFINN, BING, and VADER (see Appendix B for details).

often among the top performing methods as RF (7 out of 12 cases). However, its variance in performance is larger and it can perform poorly relative to other methods in particular when datasets are larger (see also Section 5.2).

SVM are considered less prone to overfitting (Joachims, 1998). However, they may miss important features by reducing all available training data down to the support vectors. Due to their limited representation (Domingos, 2012), they are not as flexible to model feature combinations, which RF can represent through hierarchical trees. Overall, SVM exhibit a significantly lower performance across all social media types compared with the winning methods, except for Amazon reviews (AMR). Here, they rank first – although not significantly different from ANN ($p > .05$), which is similarly versatile as RF and also has little variance in terms of relative performance. However, in the case of ANN, this has higher trade-offs in terms of implementational costs and interpretability, which are in turn similar to SVM.

kNN do not produce competitive accuracies across all tasks except for Amazon review titles (AMT) and comments from corporate blogs (CBC). As a non-parametric method, kNN suffer from the curse of dimensionality (Bellmann, 1961). Specifically, it typically overfits in-sample and, in turn, generalizes and predicts poorly out-of-sample. The number of training examples needed to maintain good generalizability for kNN grow exponentially with the number of features (Domingos, 2012), as it cannot extract a concise summary from the data like the other machine learning methods. In contrast, when there are only few features, kNN may perform reasonably well. AMT are by far the shortest text documents across all datasets with a mean of only four words,

potentially explaining why kNN compare more favorably with the other methods' performances for review titles. CBC is among the most challenging classification tasks, resulting in all methods to perform poorly and only slightly better than random chance. This also results in small performance differences across methods.

The lexicon-based classifiers such as LIWC cannot learn custom classes automatically from training data. Instead, they require expert-crafted dictionaries for such purposes (e.g., Kuhnen & Niessen, 2012). Hence, LIWC, for example, is most often used for sentiment classification in marketing research (e.g., Berger & Milkman, 2012; Hennig-Thurau et al., 2015; Hewett et al., 2016). Following this research, we apply all five dictionaries, i.e., LIWC, NRC, VADER, BING, and AFINN, to all two-class sentiment classification tasks.

On average, the lexical methods lack behind all supervised machine learning methods. For example, LIWC exhibits the highest average accuracy on IMD (61.5%), still worse than the weakest performing kNN. As Appendix B reveals, VADER and BING with the highest number of words perform best. As expected, dictionary size and specialization appear to indeed allow for higher levels of accuracy. Within the dictionary group, LIWC performance is average and never exceeds the best performing dictionary but also never falls below the weakest one.

For short review titles (AMT), the performance of the majority of dictionaries (LIWC, NRC, AFINN) does not exceed random chance. Similarly, VADER and BING also achieve accuracies of only 54.0% and 52.1%, barely above chance alone. Here, the overall average accuracy of all five dictionaries is 43.0%, compared with 77.6% of ANN, the best performing machine learning approach. For the entire review text (AMR), the dictionaries perform almost 13.3 percentage points better compared with AMT but still significantly worse than all other methods. This is intuitive as the probability of finding a positive or negative word from the dictionary within a short title of only three to five words on average is much lower compared with review texts that are on average 82 to 92 words long (see Table 1). LIWC, for example, contains 620 positive and 744 negative words (Pennebaker et al., 2015). For reviews that follow a stringent logic, contain little noise, and carry emotion-laden words, dictionaries may perform well.

However, in many real-world application scenarios the sentiment of a text tends to be conveyed in less obvious, subtler ways, especially when dealing with reviews. For example, the following review clearly conveys a negative emotion and is properly classified by all five machine learning methods. Due to the lack of negative signal words, LIWC, in contrast, cannot infer the correct sentiment: "*I received a completely different product with packaging that say made in Hong Kong! And the banana was only yellow. Not white and yellow like the photo*." Additionally, dictionaries may struggle with phrases such as in "*Damn good steak*". Again, all machine learning methods predict the correct class. LIWC, however, requires manual coding to correctly classify the bigram "*Damn good*" and consequently assigns a score of 33.3% to both the positive ("*good*") and negative ("*Damn*") word count, resulting in an ambiguous classification.

Web Appendix G reports the inter-method correlations in terms of classification accuracy. The values range from 0.35 between ANN and kNN to 0.64 between ANN and RF. These values again suggest that individual methods arrive at different accuracies depending on the task at hand. Given their consistently high levels of flexibility and high overall performance, it is not surprising that ANN and RF exhibit the highest correlation of 0.64. kNN reveal the lowest correlation with all other methods (between 0.35 with ANN to 0.47 with RF). This finding is in line with the fundamentally different learning approach of kNN, i.e., not building a model on the training data like the other methods, but instead comparing all features of each classification document with all features of the training instances.

5.2. Multivariate analysis of the drivers of text classification performance

The previous descriptive findings and statistical tests suggest certain plausible explanations for the differences in accuracies. However, the data we investigate are taken from diverse social media settings, and the different potential explanations for the observed accuracy differences cannot be disentangled (e.g., amount of available training data, languages, text length, number of classes or type of classification task). To investigate whether the conceptually plausible drivers of performance have an impact over and above the remaining factors, we run logistic regression models across all datasets with accuracy of predicting human coding (correct vs. incorrect) as the dependent variable, text and data characteristics as independent variables, and social media types as random intercepts to control for unobserved heterogeneity. To ensure each social media type is represented equally, we randomly sample 300 observations from all hold-out test sets. We include the interaction between the number of classes and the type of classification task since an additional sentiment class (neural sentiment) is conceptually different from an additional content class. We also test interactions between the number of classes and the remaining variables but find no significant effects ($p > .05$). Table 2 reports the odds ratios of this analysis.

These results reveal that the number of classes (three vs. two) and the type of classification task (sentiment vs. content) as well as their interaction exert strong effects on accuracy for all methods. Specifically, in line with conventional wisdom, content classification tasks with three compared with two classes yield lower accuracies (OR = 0.174–0.433, $p < .001$–$p < .05$). Across all methods except kNN, for two classes, sentiment classification is more challenging than content classification (e.g., OR = 0.372–0.481, $p < .001$ - $p < .05$). This is, conceptually plausible since sentiment classification often necessitates higher context understanding (Das & Chen, 2007). For example, reviews can reflect "thwarted expectations" (Pang et al., 2002), containing more negative than positive words but overall conveying a positive sentiment.

However, a strong positive interaction between number of classes and classification task (OR = 2.359–3.618, $p < .05$–$p < .001$) for all methods except kNN suggests that these differences are attenuated for more than two classes. Across all methods, the interaction is highest for RF. This is due to a relatively similar performance for three class-content and three-class sentiment classification but a much better performance for two-class content than two-class sentiment classification. Put differently, the number of classes influences the accuracy of the respective methods asymmetrically: Relative to content classification, sentiment classification suffers less from

Table 2
Random-intercept logistic regression on method performance as a function of task and text characteristics.

	Dependent variable: hold-out accuracy				
	ANN	kNN	NB	RF	SVM
Task characteristics					
# classes (1 = 3 classes, 0 = 2 classes)	0.204***	0.433*	0.244***	0.179***	0.174***
Task (1 = sentiment, 0 = content)	0.481*	0.643	0.469***	0.372***	0.441***
Interaction # classes × Task	2.861*	2.056	2.359**	3.618***	3.010***
Text characteristics					
# words (in 100)	0.904*	0.859***	0.942	0.948	0.994
# characters per word	0.974	1.012	0.924*	0.969	1.011
Language (1 = non-English, 0 = English)	1.090	0.782	0.994	0.967	1.127
Text signals					
Adjectives (e.g., happy, free)	1.013*	1.015**	1.021***	1.010	1.011*
Clout (signaling expertise)	1.004*	1.002	1.005*	1.002	1.003
Text noise					
Netspeak (e.g., lol, 4ever)	0.992	0.991	0.992	0.994	0.994
Nonfluencies (e.g., hmm, uh)	0.989	1.003	0.983	1.011	1.048
Sample size (1 = N > 1000, 0 = N ≤ 1000)	1.259*	1.058	1.033	1.262*	1.225*

Note: N = 3600; *$p < .05$, **$p < .01$, ***$p < .001$ (two-tailed). All effects reported as odds ratios. Text signals and noise variables are operationalized with LIWC.

increasing the number of classes. In other words, the marginal difficulty of adding an additional class is higher for content compared with sentiment classification. This finding may be due to the fact that sentiment in social media is often neutral. In such cases, assignment of binary values, i.e., positive vs. negative, can result in arbitrary conclusions. While an additional class by definition increases classification difficulty, better assignability of three-class sentiment appears to partially compensate this. For content classification, similar effects are less likely and additional classes make classification more complex.

Several additional effects and differences between methods deserve attention. Specifically, ANN and kNN appear to suffer from longer texts ($OR_{ANN} = 0.904$, $p < .05$; $OR_{kNN} = 0.859$, $p < .001$). The average number of words per document is proportional to the total number of features underlying the document-term-matrix. This increased complexity seems to particularly undermine the performance of these two methods. Suffering from too little training data relative to the number of features, kNN are likely to overfit, resulting in poor generalization on the out-of-sample test data. Recall kNN perform relatively poorly for longer Amazon review texts (AMR) compared with short Amazon review titles (AMT), as can be seen from Fig. 1. In terms of average word length, we find no significant effects, except for NB, where longer words seem to slightly hurt performance ($OR_{NB} = 0.924$, $p < .05$).

We control for specific text characteristics that may send particularly strong signals or induce noise. Regarding the former, all methods but RF benefit from the presence of adjectives, which are likely to contain information facilitating classification tasks. Given the strong performance of RF across all classification tasks, the absence of this effect may also indicate that RF is less dependent on such strong signal words, but instead can detect and interpret also more subtle features or feature combinations (Humphreys & Wang, 2017). Clout, a variable signaling expertise and confidence of the sender, exhibits a small positive effect for ANN and NB ($OR_{ANN} = 1.004$, $p < .05$; $OR_{NB} = 1.005$, $p < .05$). This is intuitive as a low score represents a more tentative and insecure writing style (Pennebaker et al., 2015), likely to be correlated with less precise and structured communication, which in turn may be more difficult to classify correctly. In contrast, netspeak (e.g., *4ever*, *lol*, and *b4*) and nonfluencies (e.g., *hm*, *umm*, and *uh*) do not show significant effects. For netspeak, however, all coefficients are negative, which directionally shows that noisier social media data may lead to worse performance.

We also control for language and sample size when estimating these models. According to these results, none of the assessed methods are sensitive to the language of the text corpora (all $p > .05$), suggesting that text classification performance is comparable across Germanic (e.g., German) and Romanic languages (e.g., Spanish) as well as English. Regarding sample size, three methods exhibit a significant learning curve. Among those, ANN and RF benefit the most ($OR_{ANN} = 1.259$, $p < .05$; $OR_{RF} = 1.262$, $p < .05$) and SVM the least from adding additional training data ($OR_{SVM} = 1.225$, $p < .05$). In contrast, NB and kNN appear not to significantly improve predictive performance for larger sample sizes or conversely will perform relatively better for smaller datasets ($p > .05$).

5.3. Post-estimation analysis

Based on the logistic regression results, we perform a more detailed post-estimation analysis based on the two most critical drivers of method performance: the number of classes and the type of classification task. Fig. 2 summarizes average accuracies across all 41 datasets per method.

Overall, the depicted patterns illustrate how the number of classes moderates the effect of classification task on classification accuracy. As indicated earlier, for the two-class scenarios, classifying sentiment tends to be more challenging than classifying content for all methods. For classification problems with three classes, the differences in performance between content and sentiment classification are less pronounced for all methods but kNN. Regarding ANN, RF, and NB, the differences between three-class sentiment and content are not statistically significant ($p > .05$). Conversely and from the perspective of the number of classes, means

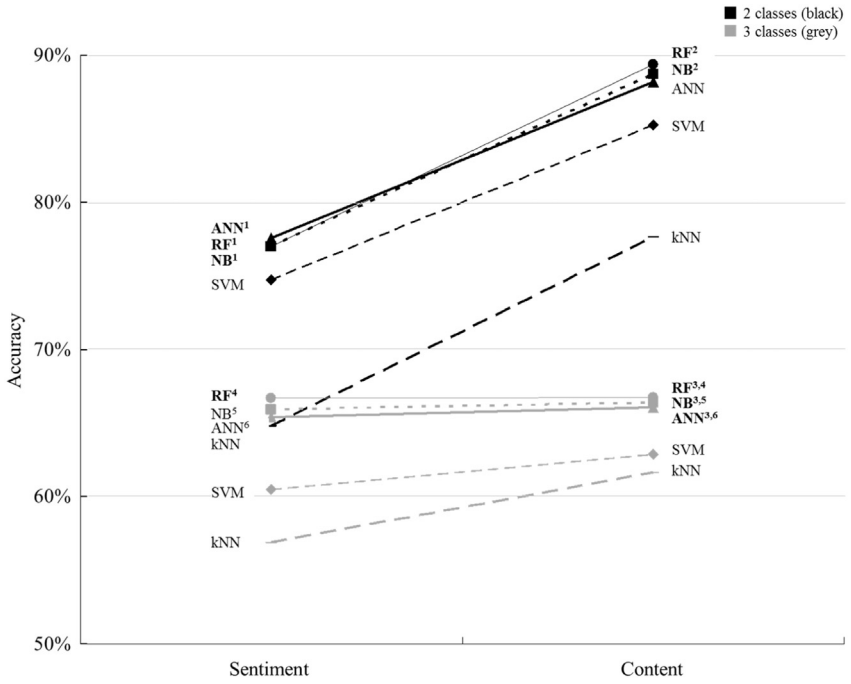


Fig. 2. Accuracies for different number of classes (three vs. two) and classification tasks (sentiment vs. content). Note: Superscripts indicate statistically insignificant differences ($p > .05$).

drop by >20 percentage points across all methods when increasing the number of classes for content classification (from 85.8% to 64.7%). However, they drop by only slightly >10 percentage points for sentiment classification (from 74.2% to 63.1%). Since additional classes contain additional information and social media communication is not exclusively negative or positive, three-class sentiment classification appears a more reasonable choice than two-class sentiment. This provides a more nuanced sentiment understanding while still allowing for high levels of classification accuracy. For example, ANN results in 77.5% classification accuracy for a 3000 documents dataset for three sentiment classes.

Most important for method selection, there are noteworthy performance differences of the methods across the four scenarios that inherently reflect the functioning of the respective learning methods. Overall, RF shows robust high performance across all four contexts. This is consistent with findings from other classification domains and suggests that RF is a particularly versatile approach also for social media text classification (e.g., Caruana & Niculescu-Mizil, 2006; Fernández-Delgado et al., 2014).

NB, on the other hand, lacking the ability to interact features and making the naïve assumption that all features are independent, performs unexpectedly well for a large variety of classification tasks. For the two-class condition of content classification, its performance is not significantly different from that of RF at $p < .05$. Also for two-class sentiment and three-class content classification it compares favorably to more flexible methods like ANN and RF. For all three tasks, texts tend to contain strong signals indicating the correct class, especially for content classification. However, also for the binary sentiment classification problems, NB performs surprisingly well, perhaps because binary sentiment classification requires less complex feature combinations than higher-order classification tasks. In contrast, when adding the neutral class, more nuanced context understanding may be required to distinguish neutral comments from positive and negative comments. For this task, RF performs significantly better than all remaining methods on average. ANN exhibit a similar performance to NB, except for two-class content where they perform slightly worse.

Recall SVM is the most frequently used supervised machine learning approach in marketing and text classification. However, SVM ranks second to last across all four classification scenarios and performs particularly poorly when the number of classes increases. While SVM appear to work reasonably well for some specific problems such as classification of journalistic and medical articles, which typically contain stringent communication and low degrees of informality (e.g., Joachims, 1998), they may miss important text signals relative to other methods by reducing all training data down to the support vectors. This can reduce their ability to capture more subtle information, e.g., when multiple features need to be interacted to produce the correct prediction. In line with this, SVM lacks behind by the largest difference for the three-class sentiment condition (60.5% for SVM vs. 66.6% for RF), where slight text nuances may determine whether a microblog post is positive, negative, or neutral.

6. Predictions based on actual consumer behavior and economic consequences of suboptimal method choices

So far, we have compared the predictive performance of ten text classification methods against the benchmark of human judgment. This approach mirrors the current focus of marketing research, which mainly uses text classification techniques

to explore the sentiment of different social media activities (e.g., Hennig-Thurau et al., 2015; Homburg, Ehm, & Artz, 2015; Tirunillai & Tellis, 2012) or classifies communication content based on theoretically meaningful categories (e.g., Ghose et al., 2012; Ordenes et al., 2018). In marketing practice, machine learning techniques are employed for cost reduction purposes as well as revenue growth (e.g., reducing the costs of market research or by providing more tailored user experiences). We investigate whether our previous results generalize to such settings. This also allows us to understand the economic consequences of suboptimal method choices. Specifically, we study three exemplary application scenarios: (1) cost reductions by automating customer service classification tasks for an online travel agency, (2) impact of social media communication on online demand and (3) website visits. These three content classification tasks vary in their number of classes, i.e., binary classes for application scenarios 2 and 3 vs. five classes for application scenario 1. As we work with custom classes, lexical classifiers are not applicable to these settings. Fig. 3 presents the methods' classification accuracies for all three scenarios. Overall, the relative performs mimics our previous results, i.e., across all scenarios either RF or NB perform best.

6.1. Application scenario 1: cost reduction for an online travel agency

The first application evaluates how textual classification can enable faster and more efficient processing of customer requests. Our data covers a sample of 3000 incoming customer emails to an online travel agency that offers leisure activities, tours, as well as tickets for local sights and attractions to tourists. Their customer service department of 150 service representatives is structured in five teams mirroring typical customer queries (i.e., questions about activities and the booking process, questions about an existing booking, cancellations, booking amendments, and complaints). Currently, the platform receives an average of 17,000 emails per week, and due to their fast organic growth over the last two years, emails are still opened unguided by any available service representative who skims through the query, before forwarding the customer inquiry to the responsible service team. We analyze how well the machine learning methods perform on this content classification task and observe substantial performance differences between the methods (see Fig. 3). Consistent with our previous findings on multi-class content classification, RF outperforms all other methods with an accuracy of 70.7%, surpassing kNN as the worst method by about 20 percentage points. Note, all methods surpass random chance of 20% to a considerable degree.

Considering the 17,000 customer emails the platform receives each week and assuming an average 30 s for the right allocation of the query, a work week of 40 h, and 46 work weeks per year, the yearly number of hours for manual inquiry handling equals 6517 (or an equivalent of >3.5 full-time service representatives). For a gross salary of €43,000, this amounts to a theoretical maximum of €152,292 in cost savings for a classifier with perfect predictions. Applied to the accuracies we observe, the best performing method (i.e., RF with 70.7%) can save >3300 h of yearly classification work compared with the random-chance baseline of 20%, equaling an overall annual cost reduction of €77,273. Compared with kNN, the higher precision of RF materializes in over €30,000 savings per year, i.e., about €1500 per percentage point increase in accuracy. Hence, suboptimal method choice can result in relevant cost consequences even for small- to medium-sized companies.

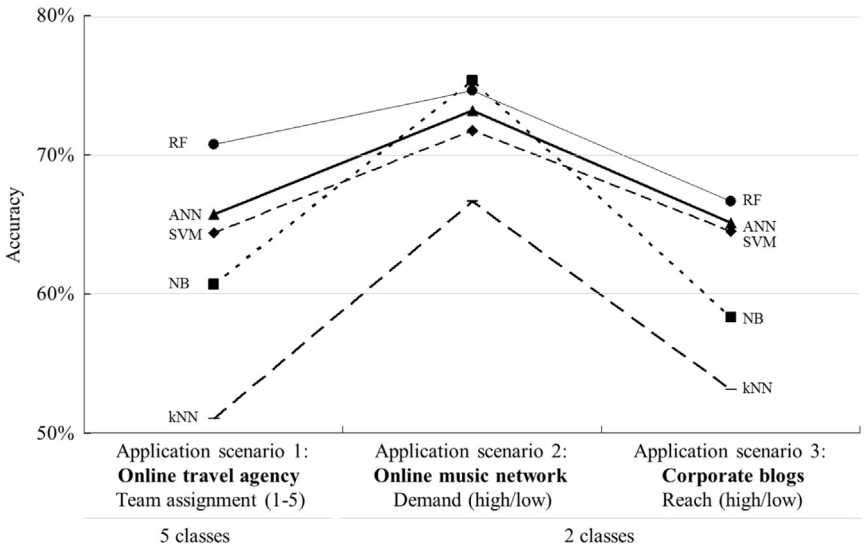


Fig. 3. Accuracies across three field application scenarios. Note: RF exhibits significantly better accuracies than all other methods at $p < .05$, except for application scenario 2 where difference between NB and RF is not significant.

6.2. Application scenario 2: demand prediction in an online music network

Many firms are interested in predicting the demand for their products. Social media data have been frequently used in marketing research as a basis to make such predictions through "social listening" (e.g., Barasch & Berger, 2014; Schweidel & Moe, 2014). The underlying assumption is that people say what they think and that this is closely related to their future behavior (e.g., Hewett et al., 2016; Kannan & Li, 2017).

To investigate the consequences of suboptimal method choice, we study a social network geared towards newcomer musicians and music fans. Music fans who seek entertainment and information visit the profiles of artists, can connect with them, and download their songs. The artists, in turn, use the platform to establish a fan base and audience for their songs, to promote upcoming concerts and new song releases as well as to create electronic word-of-mouth. Moreover, they can engage with their fans and increase visibility by writing public comments on their fans' profile pages.

The number of song downloads of an artist in the subsequent period serves as the primary marketing outcome variable of interest. Individual song demand follows a highly skewed distribution with a significant long-tail of weakly demanded songs. We median split the number of monthly song downloads to differentiate between successful and less successful songs and study whether the monthly communication in the artist social network is predictive of success in the subsequent month. Our data contains 691 communication texts for 441 music artists. Note that the communication text is Swiss-German, a colloquial German dialect, which sometimes lacks orthographic conventions, suggesting increased classification difficulties.

Despite this, all methods clearly perform better than random chance in predicting song success of an artist based on social media communication (see Fig. 3). NB and RF are the winning methods with around 75% accuracy and kNN with the worst performance of 66.7%. This mirrors our findings on two-class content classification (Fig. 2).

Based on the mean downloads for the high and low class as well as the respective method accuracies, we can quantify the forecast error for each method (further details in Web Appendix H). For this two-class example, two types of errors are possible. Specifically, the classifiers can assign low-class labels to the above median class (false positive, FP) and vice versa (false negative, FN), using the above median class as the reference category. We assume equally high costs of over- and underestimating future demand and therefore study the absolute deviation from the true means (false positive and false negatives contribute equally).

Recall, the distribution of song downloads is highly skewed, resulting in a ratio of slightly below eleven between the means of above- and below-median class, i.e., 3094 and 33,770.[1] This corresponds to a potential forecast error (difference between both classes) of 30,676 downloads or a percentage point difference in predictive accuracies in a forecast error of 307. Assuming text classification accuracy is comparable across music networks and further assuming an average price of €0.99 for a song on typical platforms such as iTunes, the differences in predictive accuracy between the best method (i.e., NB with 75.4%) and the worst method (i.e., kNN with 66.7%) can result in a significant forecast error of economic demand of 2669 song downloads per month, or €31,705 in annual revenues. Note, this is likely a conservative estimate based on the aforementioned download amounts. Higher average downloads would result in higher consequences.

6.3. Application scenario 3: forecast of corporate blog reach

Many companies attempt to operate proprietary social media such as corporate blogs with the objective of reaching as large an audience as possible. We analyze the communication data of 14 Fortune 500 blogs to evaluate the impact of corporate social media activity on future reach. Specifically, we analyze how the content of firm-generated blog posts drives the number of future visitors, returning to subsequent posts. As in the previous case, we median split the number of unique visitors of a given post, resulting in a balanced dataset with two classes. The high- and low-reach posts attract mean visits of 672 and 32, respectively. Again, we assume both types of forecast errors are equally costly (see Application scenario 2: demand prediction in an online music network), i.e., total potential forecast error of erroneous classification is 640 visitors.

According to Fig. 3, the average absolute accuracy is slightly lower compared with the previous scenario, presumably because of the weaker relationship between the blog content and the dependent variable. Nevertheless, RF and ANN achieve accuracies above 65% (i.e., 66.7% and 65.1%, respectively), whereas again kNN performs worst and only slightly above random chance (53.1%). This suggests that firms can obtain meaningful assessments of likely post impact using automated text analysis and thereby optimize their activities across employees.

Following the same computation as in application scenario two, we quantify the forecast error to translate the performance differences into economic consequences. Assuming 150 blog posts annually per company, choosing RF instead of kNN for predicting future reach reduces the forecast error by >13,000 visitors. Also, assuming a conservative value of €2 per visit due the benefits of earned media and organic traffic, i.e., reduced marketing costs (e.g., search engine advertising), suboptimal classifier choice translates to an economic impact of more than €26,000 comparing the best to the worst method.

[1] Note, the means are scaled by a constant factor, as we are not permitted to publish absolute demand information.

7. Discussion

Given the constantly growing stream of social media data, automatic classification of unstructured text data into sentiment classes or other theoretically or practically relevant researcher-defined content categories is likely to continue attracting attention. Research in marketing has gravitated towards SVM and lexicon-based methods such as LIWC for text classification. Marketing research is typically interested in how communication content appears to and affects human readers. Moreover, when choosing a text classification method, trade-offs between interpretability of results, economic relevance of differences, and implementational costs are of particular concern.

The results of our analysis both confirm the conjecture of computer science that no single method performs equally well across all application settings and also suggests simple heuristics for making practically meaningful method choices without requiring extensive method comparisons for each application. In particular, RF – which is underrepresented in text classification research both within and outside of marketing so far – is versatile and performs well across most application contexts, especially for three-class sentiment classification, which is a relevant application for marketing research (Fig. 2). Despite its conceptual simplicity, NB has also provided high accuracies in recovering human intuition. In our regression, NB and kNN are the only methods where smaller sample sizes do not result in reductions of performance (Table 2). To illustrate the implications of this, consider application scenario 2, which contained the least amount of training data and resulted in the best performance of NB out of all methods, also slightly better than RF (Fig. 3). Similarly, NB significantly outperforms all other methods in classifying the two-class sentiment of movie reviews (IMD), where the maximum sample size available to us is limited to 1000 observations. Focusing on NB and RF for text classification would have identified the best approach for all three practical applications as well as for 11 out 12 social media types.

AMR is the only exception to this. It is the only example where SVM is among the best performing methods, together with ANN. However, ANN is clearly more versatile and among the best performing methods for 7 out of 12 social media types. When implementational costs and interpretability are of lower consideration, ANN is therefore an additional promising candidate marketing research may wish to consider. Recall, according to the multivariate regression ANN and RF benefit most from larger sample sizes. Therefore, ANN appears particularly relevant for large datasets.

Overall, our findings are in contrast to emphasis of lexical classifiers in marketing research. These perform consistently and considerably worse compared with the best algorithms in recovering human intuition. All five dictionaries show also strongly inferior accuracies compared with similarly simple and intuitive approaches such as NB.

Moreover, our results confirm social media marketing applications also require context specific method choices to find optimal solutions. However, focusing on NB and RF appears as a reasonable trade-off between the objectives of interpretability, implementational costs, and economic relevance. If the former objectives are a smaller concern relative to accuracy, more exhaustive method comparison can be of interest. A few recent marketing publications such as Netzer et al. (2016) have followed this approach. For their datasets, they arrive at similar conclusions as Fig. 2 suggests, i.e., they find that NB performs best for two-class content classification.

Another intuitively appealing solution also followed by Netzer et al. (2016) is applying method ensembles, which entails training multiple classifiers. Majority votes, i.e., choosing the class the majority of methods predicts, are a simple way of accomplishing this (e.g., Xia et al., 2011). Despite their conceptual appeal, such approaches are both particularly complex and time-consuming to estimate as well as challenging to interpret and implement. In particular, they require parameter tuning for each individual method, selecting appropriate methods to include, and additionally choosing an appropriate form of aggregation. This is a combinational search problem in discrete space that is many times more complex than optimizing parameters for a single method. Even

Table 3
Objectives and characteristics of individual methods of automated text classification.

	ANN	kNN	NB	RF	SVM	DICT
Performance • Accuracy • Versatility	High, esp. for large sample sizes	Low, esp. for growing number of features (curse of dimensions)	High, esp. for small sample sizes and content classification	High, esp. for all three-class sentiment problems	Medium, relatively low for three-class problems	Low, esp. for high degree of noise, e.g., colloquial language
Implementational costs • Expert knowledge • Computational requirements	High, esp. for complex network typology (e.g., number of neurons, hidden layers)	Medium, costs are deferred to slow prediction phase (lazy learner)	Low due to naïve assumption of independent features	Low as conceptually accessible, and easy parallelization of individual decision trees	Medium to high for set-up, but training difficult to scale and parallelize	Low costs for initialization, no training time when using off-the-shelf dictionaries
Interpretability • Comprehensibility • Comparability	Low due to being highly parameterized, difficult to tune and interpret parameters	High due to low number of parameters, but difficult to interpret in high-dimensional space	High due to low number of parameters (i.e., Laplace factor) and accessible interpretation of conditional probabilities	High due to few core parameters, intuitive for individual trees, intuitive feature importance	Medium due to few core parameters to tune but support vectors conceptually difficult to interpret	High due to intuitive word counts, no parameters to tune

if parallel processing is used for each method, computation time is at least as slow as the slowest method in the ensemble and also takes more training time to identify appropriate ensemble level choices. Similarly, interpretation is many times more challenging than the most complex method of the ensemble. We have experimented with majority vote ensembles of the approaches we have covered but did not achieve better results in the applications we study. This is in line with prior research on text classification both within and outside of marketing, which also relies on a single approach and applies ensembles only in few exceptions (e.g., Lee et al., 2018; Neethu & Rajasree, 2013). Considering all marketing objectives, reliance on RF, NB and potentially ANN is likely to result in an acceptable trade-off between efficiency and predictive performance.

In terms of computational costs of individual methods, researchers and practitioners may wish to consider how efficiently the individual methods can be implemented and parallelized. Whereas ANN and SVM involve more complex optimizations, higher sample size and a larger amount word features drives computation time. In contrast, RF can be easily parallelized (Breiman, 2001). Although a few parallel implementations of SVM exist, SVM typically face scalability issues both in terms of memory and processing time required (e.g., Chang et al., 2008). In our applications, RF trains on average about four times faster than SVM across all our datasets. For our real-world classification problem from the online travel agency, SVM trains >30 times slower than RF and even slower than ANN. These differences can quickly amount to hours if not days in training time in actual application settings. Given the overall poor performance of SVM, longer computational times appear to not necessarily generate better results. Consequently, approaches such as NB or RF do not require many trade-offs and are appealing both in terms training time and predictive performance.

In addition to this, costs associated with interpretation and communication of results as well as the number of parameters associated with tuning drive application costs. In that respect, the conditional probabilities of NB further favor its use and allow for an intuitive interpretation and explanation. In contrast, SVM and ANN can be considered "black box" methods due to their complex structures, making interpretations costlier (see Table 3 for a summary of performance, implementational costs, and interpretability consequences).

There are of course limitations to this research. First, while we have analyzed an exhaustive set of the major social media platforms and real-world problems, results may differ for other types of text sources, languages, and classification objectives. Second, given the large number of analyses we run, we focus on the most important parameters for tuning. This follows the few prior method comparisons who have considered tuning (e.g., Joachims, 1998). Still, an even more extensive optimization may produce different results. Third, we also follow prior comparisons by applying standard procedures in terms of preprocessing and document representation. There are many ways of extending this to the specific task at hand, which can improve performance. Overall, our results can be viewed as a lower performance boundary which further emphasizes the potential of automated text classification. In addition, there are algorithms and dictionaries beyond the ones we study. However, we believe we cover a representative set of both machine learning and lexical methods.

Moreover, commercial alternatives have recently appeared that marketing researchers and practitioners can use to generate insights through automated natural language processing of unstructured texts, e.g., from Microsoft, Amazon or Google. This research has been limited to the types of methods applied in prior publications. Conceptually, these commercial solutions require fewer or no training examples and less technical expertise (e.g., to implement and calibrate the machine learning methods). However, they are also less specialized for the task and text domain at hand. Consequently, they may not be optimally suited for the specific questions marketing researchers may pose, e.g., when dealing with special types of texts or custom classes. For example, the Google Cloud Natural Language API can classify a generic set of several hundred content classes, which will often not match more specific interests of marketing research. However, the field is developing rapidly and we strongly encourage readers to monitor the development of such commercial services.

In many ways the results of this research represent a middle ground between marketing research and computer science. While the latter emphasizes exhaustive method comparisons for each application, the former applied marketing perspective must also consider application efficiency, standardization, and comparability. In extension to the computer science literature, our results suggest that inferior method choices can indeed result in important economic consequences. At the same time and with very few exceptions, performance differences between RF, ANN, and NB are small and likely subordinate to most research applications. According to our results choosing between RF and NB based on sentiment vs. content classification as well as the number of classes appears a reasonable trade-off between efficiency, comparability, and accuracy. We hope these findings make sound automated text classification more approachable to marketing researchers and encourage future research to integrate social media communication as a standard component in econometric marketing models.

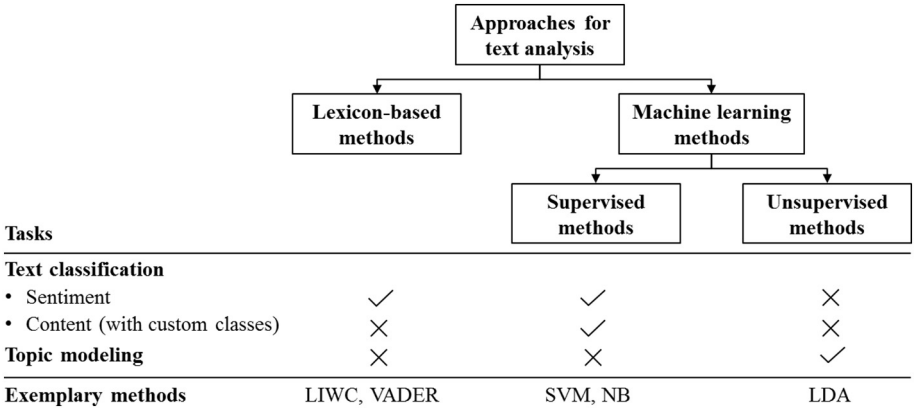
Funding

This work was funded by the German Research Foundation (DFG) research unit 1452, "How Social Media is Changing Marketing", HE 6703/1-2.

Acknowledgements

The authors thank Chris Biemann, Brett Lantz, Julius Nagel, Robin Katzenstein, Christian Siebert, and Ann-Kristin Kupfer for their valuable feedback and suggestions.

Appendix A. Approaches for automated text analysis



Appendix B. Dictionary characteristics of five lexicon-based methods and classification accuracies

		Dictionary characteristics			Classification accuracies (in %)			
Method	Author(s)	Type	Positive words	Negative words	AMR	YEL	AMT	IMD
AFINN	Nielsen (2011)	Strength	878	1,598	56.8	51.8	46.1	62.0
BING	Hu & Liu (2004)	Polarity	2,006	4,783	58.2	62.3	52.1	68.2
LIWC	Pennebaker et al. (2015)	Polarity	620	744	54.0	53.0	39.2	61.5
NRC	Mohammad & Turney (2010)	Polarity	1,070	814	48.1	45.0	23.8	53.0
VADER	Hutto & Gilbert (2014)	Strength	3,344	4,173	64.4	63.5	54.0	67.8

Note: We test VADER in Python and AFINN, BING, and NRC as implemented in the *syuzhet* package in R. For NRC and LIWC we evaluate all Amazon datasets in both English and German. To make the results of all dictionaries comparable, we convert the sentiment strength scales of AFINN and VADER to binary polarities.

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijresmar.2018.09.009.

References

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. *Mining text data* (pp. 163–222). Boston, MA: Springer.
Akpinar, E., & Berger, J. (2017). Valuable virality. *Journal of Marketing Research*, *54*(2), 318–330.
Alberto, T. C., Lochter, J. V., & Almeida, T. A. (2015). Comment spam filtering on YouTube, Proceedings of the 14th IEEE international conference on machine learning and applications.
Almeida, T. A., Gómez Hidalgo, J. M., & Yamaki, A. (2011). Contributions to the study of SMS spam filtering: New collection and results. Proceedings of the 2011 ACM symposium on document engineering.
Annett, M., & Kondrak, G. (2008). A comparison of sentiment analysis techniques: Polarizing movie blogs. Conference of the Canadian Society for Computational Studies of IntelligenceBerlin, Heidelberg: Springer.
Barasch, A., & Berger, J. (2014). Broadcasting and narrowcasting: How audience size affects what people share. *Journal of Marketing Research*, *51*(3), 286–299.
Bellmann, R. E. (1961). Adaptive control processes: A guided tour. *Princeton University Press*. Princeton: NJ.
Bennett, K. P., & Campbell, C. (2000). Support vector machines: Hype or hallelujah? *ACM SIGKDD explorations newsletter. 2(2).* (pp. 1–13).
Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192–205.
Bermingham, A., & Smeaton, A. F. (2010). Classifying sentiment in microblogs: Is brevity an advantage? Proceedings of the 19th ACM international conference on information and knowledge management.
Blei, D. M., Andrew, N. G., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(1), 993–1002.
Boiy, E., Hens, P., Deschacht, K., & Moens, M. F. (2007). Automatic sentiment analysis in online text. *ELPUB* (pp. 349–360).
Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.
Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.
Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. Proceedings of the 23rd international conference on machine learning.
Cavanaugh, L. A., Bettman, J. R., & Luce, M. F. (2015). Feeling love and doing more for distant others: Specific positive emotions differentially affect prosocial consumption. *Journal of Marketing Research*, *52*(5), 657–673.
Chang, E. Y., Zhu, K., Wang, H., Bai, H., Li, J., Qiu, Z., & Cui, H. (2008). Parallelizing support vector machines on distributed computers. *Advances in Neural Information Processing Systems*, 257–264.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.
Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, *53*(9), 1375–1388.
Detienne, K. B., Detienne, D. H., & Joshi, S. A. (2003). Neural networks as statistical tools for business researchers. *Organizational Research Methods*, *6*(2), 236–265.
Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.
Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*(2), 103–130.
Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. Proceedings of the seventh international conference on information and knowledge management.
Efron, B., & Hastie, T. (2016). *Computer age statistical inference.* Cambridge, MA: Cambridge University Press.
Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, *2*(5), 1–14.
Felbermayr, A., & Nanopoulos, A. (2016). The role of emotions for the perceived usefulness in online customer reviews. *Journal of Interactive Marketing*, *36*, 60–76.
Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*(1), 3133–3181.
Ghose, A., Ipeirotis, P. G., & Li, B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science*, *31*(3), 493–520.
Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report. 1(12)*. (pp. 1–6). Stanford.
Hansen, N., Kupfer, A. K., & Hennig-Thurau, T. (2018). Brand crisis in the digital age: The short- and long-term effects of social media firestorms on consumers and brands. *International Journal of Research in Marketing*, 1–51 forthcoming https://www.sciencedirect.com/science/article/abs/pii/S0167811618300351.
Hennig-Thurau, T., Wiertz, C., & Feldhaus, F. (2015). Does twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. *Journal of the Academy of Marketing Science*, *43*(3), 375–394.
Hewett, K., Rand, W., Rust, R. T., & van Heerde, H. J. (2016). Brand buzz in the echoverse. *Journal of Marketing*, *80*(3), 1–24.
Homburg, C., Ehm, L., & Artz, M. (2015). Measuring and managing consumer sentiment in an online community environment. *Journal of Marketing Research*, *52*(5), 629–641.
Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining.
Huang, D., & Luo, L. (2016). Consumer preference elicitation of complex products using fuzzy support vector machine active learning. *Marketing Science*, *35*(3), 445–464.
Humphreys, A., & Wang, R. J. -H. (2017). Automated text analysis for consumer research. *Journal of Consumer Research*, *44*(6), 1274–1306.
Hutto, E., & Gilbert, C. J. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media textEighth international conference on weblogs and social media (ICWSM-14).
Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine Learning*, *ECML-98*, 137–142.
Kannan, P. K., & Li, H. A. (2017). Digital marketing: A framework, review and research agenda. *International Journal of Research in Marketing*, *34*(1), 22–45.
Kotziats, D., Denil, M., De Freitas, N., & Smyth, P. (2015). From group to individual labels using deep features. *KDD* (pp. 1–10).
Kübler, R. V., Colicev, A., & Pauwels, K. (2017). Social media's mindset: When to use which sentiment extraction tool? *Marketing Science Institute working paper series*, *17(122)*. (pp. 1–99).
Kuhnen, C. M., & Niessen, A. (2012). Public opinion and executive compensation. *Management Science*, *58*(7), 1249–1272.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.
Lee, D., Hosanagar, K., & Nair, H. (2018). Advertising content and consumer engagement on social media: Evidence from Facebook. *Management Science*, 1–27 (forthcoming).
Markets and Markets (2018). *Text analytics market by component.* September 21, 2018, accessed from https://www.marketsandmarkets.com/PressReleases/text-analytics.asp.
McAuley, J., Targett, C., Shi, Q., & Van Den Hengel, A. (2015). Image-based recommendations on styles and substitutes. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval.
McKinsey Global Institute (2017). *Artificial intelligence. The next digital frontier?* , 1–80.
Melville, P., Gryc, W., & Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining.
Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text.
Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, *40*(2), 621–633.
Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. *Computing, communications and networking technologies (ICCCNT)* (pp. 1–5).
Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, *31*(3), 521–543.
Netzer, O., Lemaire, A., & Herzenstein, M. (2016). *When words sweat: Identifying signals for loan default in the text of loan applications.* (Working Paper).
Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems*, 841–848.
Nielsen, F. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. Proceedings of the ESWC2011 workshop on 'making sense of microposts': Big things come in small packages.
Ordenes, F. V., Grewal, D., Ludwig, S., Ruyter, K. D., Mahr, D., Wetzels, M., & Kopalle, P. (2018). Cutting through content clutter: How speech and image acts drive consumer sharing of social media brand messages. *Journal of Consumer Research*, 1–65 (forthcoming).
Ordenes, F. V., Ludwig, S., Grewal, D., & Wetzels, M. (2017). Unveiling what is written in the stars: Analyzing explicit, implicit, and discourse patterns of sentiment in social media. *Journal of Consumer Research*, *43*(6), 875–894.
Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proceedings of the 42nd annual meeting on association for computational linguistics.
Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on empirical methods in natural language processing*10*. (pp. 79–86).
Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015.* Austin, TX: University of Texas at Austin.
Puranam, D., Narayan, V., & Kadiyali, V. (2017). The effect of calorie posting regulation on consumer opinion: A flexible latent Dirichlet allocation model with informative priors. *Marketing Science*, *36*(5), 726–746.
Scholkopf, B., & Smola, A. J. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond.* Cambridge, MA: MIT Press.
Schweidel, D. A., & Moe, W. W. (2014). Listening in on social media: A joint model of sentiment and venue format choice. *Journal of Marketing Research*, *51*(4), 387–402.
Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47.
Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, *31*(2), 198–215.
Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, *8*(7), 1341–1390.
Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37.
Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, *181*(6), 1138–1152.
Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, *1*(1–2), 69–90.
Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

Yang, Z., Nie, X., Xu, W., & Guo, J. (2006). An approach to spam detection by naive Bayes ensemble based on decision induction. Intelligent systems design and applicationsISDA'06. Sixth international conference*2*. (pp. 861–866).
Ye, J., Chow, J. H., Chen, J., & Zheng, Z. (2009). Stochastic gradient boosted distributed decision trees. Proceedings of the 18th ACM conference on information and knowledge management.
Yoganarasimhan, H. (2018). Search personalization using machine learning. *Management Science*, 1–52 (forthcoming).
Zhang, Y., Moe, W. W., & Schweidel, D. A. (2017). Modeling the role of message content and influencers in social media rebroadcasting. *International Journal of Research in Marketing*, *34*(1), 100–119.