International Journal of Research in Marketing 40 (2023) 75–87

Contents lists available at ScienceDirect

International Journal of Research in Marketing

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

More than a Feeling: Accuracy and Application of Sentiment Analysis

Jochen Hartmann [a,1], Mark Heitmann [b,1], Christian Siebert [b,*,1], Christina Schamp [c]

[a] University of Groningen, Faculty of Economics and Business, Nettelbosje 2, 9747 AE Groningen, the Netherlands
[b] Universität Hamburg, Moorweidenstrasse 18, 20148 Hamburg, Germany
[c] Institute for Digital Marketing and Behavioral Insights, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

ARTICLE INFO

Article history:
First received on July 02, 2021 was under review for 4½ months
Available online 20 June 2022

Area Editor: Michael Haenlein
Accepting Editor: David Schweidel

Keywords:
Sentiment Analysis
Meta-Analysis
Natural Language Processing
Machine Learning
Transfer Learning
Deep Contextual Language Models
Text Mining

ABSTRACT

Sentiment is fundamental to human communication. Countless marketing applications mine opinions from social media communication, news articles, customer feedback, or corporate communication. Various sentiment analysis methods are available and new ones have recently been proposed. Lexicons can relate individual words and expressions to sentiment scores. In contrast, machine learning methods are more complex to interpret, but promise higher accuracy, i.e., fewer false classifications. We propose an empirical framework and quantify these trade-offs for different types of research questions, data characteristics, and analytical resources to enable informed method decisions contingent on the application context. Based on a meta-analysis of 272 datasets and 12 million sentiment-labeled text documents, we find that the recently proposed transfer learning models indeed perform best, but can perform worse than popular leaderboard benchmarks suggest. We quantify the accuracy-interpretability trade-off, showing that, compared to widely established lexicons, transfer learning models on average classify more than 20 percentage points more documents correctly. To form realistic performance expectations, additional context variables, most importantly the desired number of sentiment classes and the text length, should be taken into account. We provide a pre-trained sentiment analysis model (called SiEBERT) with open-source scripts that can be applied as easily as an off-the-shelf lexicon.

Method comparisons have a long tradition in marketing research and have established standards for comparable empirical research (e.g., Andrews, Ainslie, & Currim, 2002; Bremer, Heitmann, & Schreiner, 2016; Reinartz, Haenlein, & Henseler, 2009). An area with unabated interest is classifying large volumes of unstructured text data – e.g., social media posts, consumer reviews, and firm-generated content – to study how sentiment relates to relevant marketing outcomes (e.g., Chakraborty, Kim, & Sudhir, 2022; Hartmann, Heitmann, Schamp, & Netzer, 2021; Netzer, Lemaire, & Herzenstein, 2019; Ordabayeva, Nailya, Cavanaugh, & Dahl, in press). For instance, the valence of text is related to dynamic feedback loops between different sources of brand communication (Hewett, Rand, Rust, & van Heerde, 2016), is a relevant lead indicator for customer lifetime value (Meire, Hewett, Ballings, Kumar, & Van den Poel, 2019), impacts how consumers respond to word

* Corresponding author.
E-mail addresses: j.p.hartmann@rug.nl (J. Hartmann), mark.heitmann@uni-hamburg.de (M. Heitmann), christian.siebert@uni-hamburg.de (C. Siebert), christina.schamp@wu.ac.at (C. Schamp).
[1] listed alphabetically to reflect overall equal contributions.

https://doi.org/10.1016/j.ijresmar.2022.05.005
0167-8116/

of mouth (Lafreniere, Moore, & Fisher, in press), helps to predict stock market returns (Tirunillai & Seshadri, 2012), and is useful for social media firestorm detection (Herhausen, Ludwig, Grewal, Wulf, & Schoegel, 2019).

Unlocking the potential of sentiment analysis necessitates accurate text classification as positive, neutral, or negative. When too many text documents are assigned to the wrong category, substantive misinterpretations (Jaidka et al., 2020) and lower explanatory power in terms of R-squared are likely (Kübler, Colicev, & Pauwels, 2020), which can translate into substantial business loss (Hartmann, Huppertz, Schamp, & Heitmann, 2019). This makes accuracy, i.e., the share of correct sentiment predictions out of all predictions, also known as *hit rate*, a critical concern for sentiment research.

Hartmann et al. (2019) were among the first to conduct a systematic comparison of the accuracy of sentiment analysis methods for marketing applications. They find large accuracy differences between individual methods but also large differences between individual applications, suggesting comparisons on a single dataset or a handful of applications are hardly informative. On average, popular dictionaries like Linguistic Inquiry and Word Count (LIWC, Pennebaker et al., 2015) and Valence Aware Dictionary and sEntiment Reasoner (VADER, Hutto & Gilbert, 2014) performed poorly compared to traditional machine learning methods such as naïve Bayes or random forests. Dictionaries assign sentiment to words or recurring expressions based on theoretical top-down considerations (i.e., hypothesized language patterns, Humphreys, 2019). This makes it straightforward to trace back which expressions occur how often and how much they contribute to overall sentiment scores. Machine learning methods, on the other hand, promise more correct sentiment assignments by automatically finding text classification rules based on a training data sample with human sentiment coding. This latter approach is associated with complex classifications rules, higher-order interactions, and non-linear patterns that make it challenging to follow how individual expressions relate to sentiment scores. These conceptual differences force researchers into a trade-off between the benefits of interpretability associated with lexicons and the higher levels of accuracy associated with machine learning. Informed methodological decisions require knowledge of the size of this trade-off. However, no large-scale assessment of its size and variance across methods and applications is available.

In addition, natural language processing has made a technological leap towards large-scale *transfer learning* in recent years (Ruder, Peters, Swayamdipta, & Wolf, 2019). In contrast to the traditional machine learning studied by Hartmann et al. (2019), these more recent models can transfer statistical language understanding in terms of generic expressions and word relationships from related domains and tasks to a given sentiment dataset. Due to their architecture, these methods can learn fundamental text representations that are useful across domains (Manning, Clark, Hewitt, Khandelwal, & Levy, 2020). These transfer learning models are predominantly based on the language model architecture of BERT (Bidirectional Encoder Representations from Transformers, Devlin et al., 2019), which builds on the seminal work by Vaswani et al. (2017). They have been picked up in recent work by Alantari, Currim, Deng, & Singh (2022), who study a few review datasets (mostly from Amazon). Questions have been raised whether such a limited view suffices to reach valid conclusions, because sentiment analysis performance varies strongly across applications (Hutson, 2020). Therefore, little practical guidance exists about the actual benefits of transfer learning compared to traditional machine learning methods and lexicons. The lack of adequate benchmarks forces researchers to rely on gut feeling or to simply follow the lead of prior research when choosing a sentiment analysis method.

This research addresses these shortcomings by considering about 70 times more sentiment datasets than prior work (Hartmann et al., 2019). Specifically, we conduct a comprehensive meta-analytical assessment, covering 217 publications, more than 1,100 experimental results, nearly 12 million sentiment-labeled text documents, and 272 unique datasets. Based on this data, we can study and explain variation in method performance across datasets to compute realistic performance benchmarks for relevant application scenarios based on different types of research questions, data characteristics, and available analytical resources. We find that state-of-the-art transfer learning models that are most difficult to interpret achieve the fewest misclassifications. We quantify the size of the accuracy-interpretability trade-off compared to more interpretable lexicons. Specifically, we find that transfer learning achieves classification accuracies that are, on average, 20 percentage points higher than interpretable lexicons. Switching from two- to three-class sentiment classification reduces accuracy by 12.7 percentage points, document- versus sentence-level sentiment analysis is associated with a 5 percentage point increase, and increasing the sample size benefits traditional machine learning methods more than lexicons and transfer learning models.

We translate these findings into relevant sentiment analysis scenarios and construct a decision tree on which methods are most appropriate for which contexts. To further assist the adoption of more recent methods, we provide open-source code and a fine-tuned language model that is readily available to automatically construct reproducible sentiment labels for any English text data (https://github.com/j-hartmann/siebert). The model, called SiEBERT (prefix for "Sentiment in English"), can also be fine-tuned to further improve accuracy for novel applications.

Empirical Framework

When faced with a sentiment classification problem, understanding expected method performance is crucial to form realistic performance expectations and make informed method choices. For this purpose, a marketing researcher may run own experiments or draw on existing comparative method assessments. However, resource limitations make exhaustive testing of all conceivable alternatives impractical for typical substantive research applications. The latter approach is left with computer science comparisons that typically test methods on one or a few lighthouse illustrations and provide little information

about likely performance for actual applied tasks. For example, in this literature support vector machines have been reported as a winning method in one context at 98.70% accuracy (Ghosal, Tirthankar, Das, Bhattacharjee, & Saprativa, 2015) and an inferior choice in a different context at 46.20% (Winkler et al., 2015). Similarly, reports on performance of artificial neural networks range from 40.29% (Liu, Bi, & Fan, 2017) to 99.63% (Ravi et al., 2015). Given the inconclusiveness of the literature and the challenges associated with exhaustive method assessments, researchers often seek common ground and apply whatever happens to be popular in marketing research, often failing to take potential drawbacks and context contingencies into account (Hartmann et al., 2019).

Better informed choices require a method comparison that adequately controls for the relevant differences between the individual application scenarios and thereby approximates a like-for-like comparison. We therefore investigate the dimensions that differ between current accuracy reports to reconcile the variations in results. Fig. 1 presents the resulting empirical framework, which identifies four observable drivers along the stages of sentiment research. We can observe and code all of these drivers for each accuracy report, allowing us to make empirically sound inferences how each driver impacts attainable accuracy. We can then employ the estimated coefficients to predict the accuracy researchers can expect for each method and each application scenario keeping all else constant. This allows us to understand the size of the accuracy-interpretability trade-off researchers need to make when choosing between more interpretable lexicons and the various, more accurate machine learning methods. Aside from the implications for theory building (i.e., the degree of interpretability of the different methods), we expect and find that the formulation of the research question (i.e., binary versus three-class sentiment), the characteristics of the available data (i.e., document versus sentence level), and the scale of the analytical procedure (i.e., training data size) matter in terms of accuracy.

Intuitively, some classification problems are more difficult to solve than others. Context matters. Hence, the accuracy of sentiment classification is closely tied to the research question a scholar seeks to address. While we cannot assess all conceivable data characteristics that the research question implies, we can empirically assess the performance impact of a researcher's decision to pursue binary versus multi-class sentiment analysis. Plenty of examples exist that highlight the practical relevance of this dimension. For example, the analysis of firestorms is primarily interested in negative (versus other) sentiment (e.g., Hansen, Kupfer, & Hennig-Thurau, 2018; Herhausen et al., 2019). In fact, the negative class may even be positively related to business outcomes (Ordabayeva et al., in press). Other research questions, in contrast, require more nuance, i.e., a positive, negative, and neutral class. For example, Tang, Fang, & Wang (2014) identify substantial differences between reviews with mixed-neutral sentiment, containing an equal amount of positive and negative claims, and indifferent-neutral reviews, which feature neither positive nor negative claims. Despite this conceptual appeal of more classes, chance alone suggests more false classifications for more nuanced coding. However, the size of the accuracy trade-off is an open question that we address in this research.

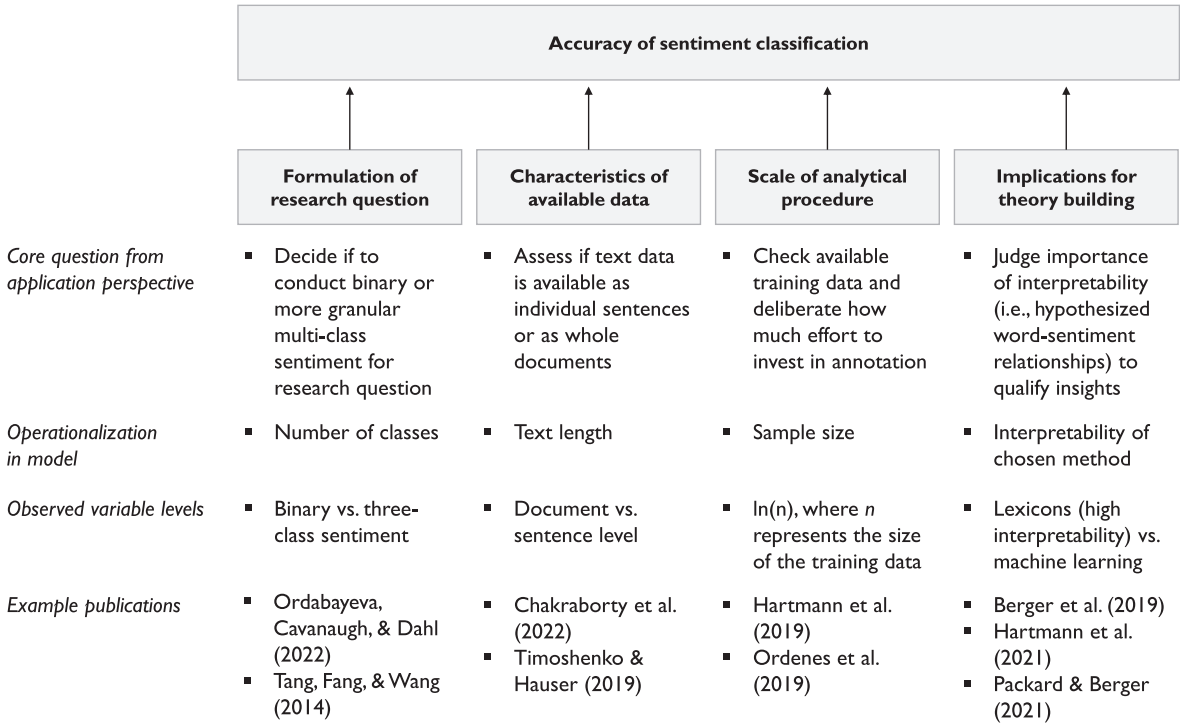

Fig. 1. Empirical Framework with Main Drivers of Sentiment Analysis Accuracy.

Text data also varies in length, and length in turn is related to automated coding capabilities. Intuitively, text length determines the richness and density of information that can be conveyed. A document, consisting of multiple sentences, is likely to contain more sentiment signals than a single sentence. In line with this, Hartmann et al. (2019) find that automatically classifying brief Amazon review titles results in more errors compared to richer full-length reviews. While full documents may be desirable, not all applications can build on them. For example, researchers may have only the titles or short summaries of product reviews available. Certain applications also involve coding sentiment for individual sentences, e.g., to identify pros and cons in a longer product review (Timoshenko & Hauser, 2019; Chakraborty et al., 2022). To investigate the consequences of these different types of text data, we code all accuracy reports as either sentence or document level and determine how these data characteristics relate to attainable accuracy.

The application context also defines the costs of annotated training data and thereby the scale of the analytical procedure. In some situations, assessing the sentiment of a couple of thousand social media posts can be accomplished at reasonable costs using Amazon Mechanical Turk or similar services. In other situations, expert knowledge is needed to obtain appropriate labels. Consider, for example, specialized financial reports that necessitate deep domain expertise to judge whether a disclosure is good, bad, or neutral news. The various different settings result in training data ranging between a few hundred (Hartmann et al., 2019) to several thousand annotated observations (Ordenes et al., 2019). Clearly, more training data likely facilitates better training and better classification performance with fewer errors. However, it is not clear how much training data is needed per method, in particular since some methods can transfer knowledge from related tasks and may need less training data than others. We will estimate the average return on labeled data for each sentiment analysis method, so researchers can make informed decisions about how much data to label and assess whether a lack of training data may be driving lower than expected performance.

Method choice also determines whether researchers can interpret which words or expressions contribute to sentiment scores. By construction, lexicons are interpretable, since each word or phrase is assigned a sentiment score and the (weighted) sum constitutes the total sentiment of a sentence or document (Humphreys, 2019). However, performance hinges on the quality of the lexicon for the task at hand and lexicons seldom generalize well across contexts (Berger et al., 2020). Human language is riddled with ambiguity and the same word and expression can have very different meaning to different readers. Reflecting such nuances makes models more accurate but also more complex to understand and interpret. Compared to highly parameterized machine learning methods, a rule-based linear combination of pre-defined words is unlikely to fully acknowledge all relevant linguistic non-linearities. However, lexicons' radical simplification can help researchers better understand the "why", i.e., the psychological drivers behind human communication, which can contribute to building and extending theory on why sentiment has certain effects (Berger et al., 2020). Our large-scale assessment enables us to substantiate and quantify the accuracy-interpretability trade-off for different application scenarios.

In addition to the main drivers of sentiment analysis accuracy, we meta-analytically control for additional variables (e.g., text language), publication-relevant factors (publication type and year), and any remaining unobserved dataset heterogeneity. To the best of our knowledge, this represents the first large-scale assessment of sentiment analysis accuracy that adequately accounts for differences in datasets and application contexts.

Sentiment Analysis Methods

We identify three conceptually distinct groups of methods that have been applied by marketing research: (1) lexicons, (2) traditional machine learning, and (3) transfer learning. These groups reflect two major methodological transitions: first, from labor-intensive, hand-crafted lexicons to automatic machine learning methods, typically trained on high-dimensional, sparse bag-of-words features (e.g., Hartmann et al., 2019; Netzer et al., 2019; Ordenes, Ludwig, De Ruyter, Grewal, & Wetzels, 2017); and second, learning low-dimensional, dense embeddings from texts through artificial neural networks using pre-training based on large-scale open-domain text (e.g., Hartmann et al., 2021; Puranam, Kadiyali, & Narayan, 2021).

Fig. 2 provides an overview of these three main method types. We use this overview to explain the differences between the methods following the example application of sentiment analysis for social media research.

Sentiment lexicons are particularly accessible and popular in applied research. Each word or phrase in a (standard) dictionary is assigned a sentiment orientation (i.e., positive or negative) or a continuous sentiment score. Lexicons require no labeled training data. In simple versions, the software solely counts the number of word occurrences in the lexicon and classifies each document according to the relative frequency of positive or negative words. This makes results interpretable as they are based on hypothesized language patterns and their association with the desired construct (e.g., sentiment). Off-the-shelf dictionaries exist (e.g., LIWC, VADER; see 1*a* in Fig. 2), but custom dictionaries can also be constructed when less common or task-specific vocabulary plays a relevant role (e.g., for firestorm detection (Hansen et al., 2018); see 1*b* in Fig. 2).

In contrast to lexicons, traditional machine learning methods learn how to assign sentiment labels to new observations based on annotated training data. This procedure is known as supervised learning, as the learning process is supervised by the task-specific labels, i.e., sentiment classes, of the data (see 2 in Fig. 2). Traditional machine learning methods can automatically capture dataset-specific vocabulary and sentiment associations based on manually labelled data in a bottom-up manner on how each word relates to the sentiment coding of the training data (Humphreys, 2019). The models are trained from scratch for each new application (e.g., Hartmann et al., 2019), i.e., what is learned for one application is lost for the next and seldom leveraged to other domains. Hence, marketing scholars seeking to employ traditional machine learning methods

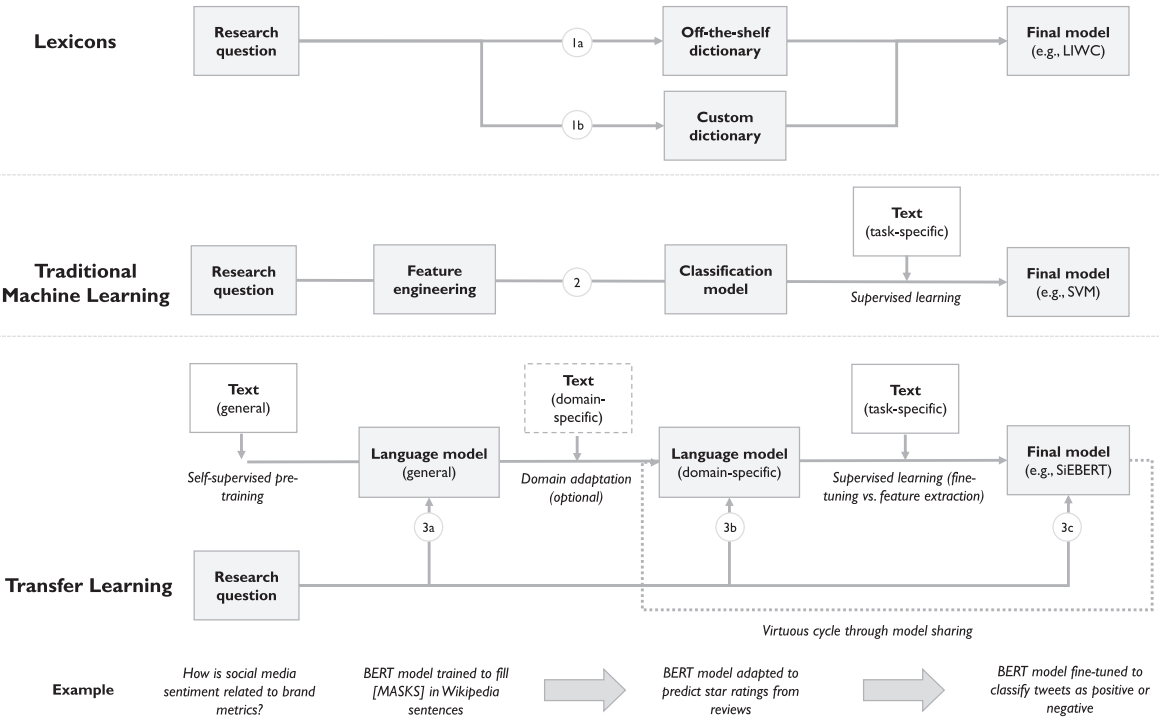

Fig. 2. Main Approaches for Sentiment Analysis: Lexicons, Traditional Machine Learning, and Transfer Learning.

need to obtain annotated training data (e.g., tweets labeled as positive or negative) to train their model and run an automated classification of the full data. Frequently used methods include support vector machines (e.g., Homburg, Ehm, & Artz, 2015; Ordenes et al., 2017), naïve Bayes (e.g., Netzer et al., 2019), and random forests (e.g., Hartmann et al., 2019; Matalon, Magdaci, Almozlino, & Yamin, 2021). Typically, these methods are trained on texts represented as bag-of-words features, reflecting the presence or frequency of words (or n-grams) in a high-dimensional, sparse document-term matrix. This process of finding appropriate representations of the text is known as *feature engineering* (see Fig. 2) and includes numerous pre-processing decisions (e.g., stopword removal, stemming, and frequency thresholds) that require extensive exploration and experience. Another limitation of bag-of-words features is that each unique word or n-gram contributes to the number of parameters with more dramatic increases for more complex models, which can undermine attainable accuracy.

This problem is addressed with word embeddings. Popular algorithms include GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013). These algorithms generate representations of words as dense vectors from large unlabeled text corpora. In addition to dimensionality reduction, the resulting multidimensional vectors have the appealing property of encoding semantic relationships between words, such that the vector representation $vec(horrible) - vec(bad) + vec(good)$ closely aligns with the $vec(wonderful)$. These low-dimensional word embeddings are mostly used in conjunction with artificial neural networks that learn the relationship between the vector representations and sentiment labels (e.g., Timoshenko & Hauser, 2019).

While traditional embedding approaches can simplify the classification task by analyzing aggregate embedding dimensions instead of individual words, the way these embeddings are constructed can be problematic. Specifically, traditional static embeddings assign each word to the same embedding in the exact same way, no matter where it appears, i.e., implicitly assuming the same word meaning irrespective of context (Tunstall, von Werra, & Leandro, 2022). For example, no distinction would be made between the word "fly" as a noun, a verb, or even as an colloquial adjective. A human reader, however, would understand these differences based on the text context.

Context-dependent embeddings address this limitation. The idea is to embed each word based on its surrounding context using what is called self-attention (Vaswani et al., 2017). Training can be done with a language modeling objective, i.e., by predicting words in a masked text based on word context (Liu et al., 2019). Since this approach (known as *self-supervised* learning, Manning et al., 2020) does not require manual data annotations, extremely large text corpora can be taken into account to understand all conceivable word contexts, e.g., 33 billion sub-word tokens for the XLNet model (Yang et al., 2019). These models are used as general-purpose language models (see 3*a* in Fig. 2), that can be fine-tuned to the sentiment analysis task based on labeled training data. The resulting sentiment model essentially combines a general language understanding with a sentiment understanding from training data, i.e., it leverages many times more information than all other

methods, which are confined to the training data. This pre-training strategy is similar in spirit to transfer learning in computer vision, where pre-training neural networks on large data sets (e.g., ImageNet) enables better performance when fine-tuning these initialized networks on the data and problem of interest.

Returning the social media example, marketing scholars could decide to start with a general-purpose language model themselves (3*a*) and fine-tune it with task-specific training data to better reflect the types of expressions in the text of interest. Alternatively, researchers can start with a domain-adapted model that is already trained for sentiment analysis (see 3*b* in Fig. 2). SiEBERT, the language model we provide together with this paper, represents such a domain-adapted model that has been further trained for sentiment analysis based on all publicly available two-class sentiment datasets that we could identify in this research. It can be used as a starting point for different sentiment-related tasks by fine-tuning it for a particular text domain (3*b*). Such training is typically more efficient compared to using a general-purpose language model (3*a*), i.e., requires less training data and has lower computational demands to attain similar levels of accuracy. For example, when fine-tuning SiEBERT on a new three-class sentiment dataset (sentiment140) with as little as 400 training examples, it reaches a classification accuracy of 68.37% after only 1 epoch of training. A general-purpose model such as RoBERTa reaches only 29.59% accuracy with this limited training (see: https://github.com/j-hartmann/siebert). Sharing and improving fine-tuned models helps other researchers attain better results and can result in a virtuous cycle of classifying more and more nuances in more and more complicated texts. If no fine-tuning is desired or needed, researchers can also use a pre-trained language model such as SiEBERT in an off-the-shelf manner without any training, which makes sentiment classifications readily available and transfer learning as easy as applying an off-the-shelf lexicon (see 3c).

Data Description & Model-Free Results

To assess if the large variation in accuracy per method is limited to only a few anecdotal cases, we draw on a large-scale dataset of comparative studies. Specifically, we have systematically identified 217 publications that compare sentiment analysis methods with sufficient reporting for our purposes. Since these comparative evaluations test 1.25 datasets on average, we can take 272 unique datasets and over 12 million sentiment-labeled text documents into account (see Web Appendix for details). In total, we draw on more than 1,100 experimental results.

If the large variation in accuracy were indeed limited only to the aforementioned anecdotal cases, marketing scholars could simply sample selected papers to inform their method choice. However, as Fig. 3 shows, accuracy values per method, depicted by grey dots, are scattered across the entire value range. Since most applications test only a few methods, this analysis alone does not permit a meaningful like-for-like comparison of the different methods for relevant application scenarios. On average, artificial neural networks emerge as the best-performing method (85.75%), while k-nearest neighbors perform worst (67.31%). Lexicons emerge as the second-worst performing method (70.29%). This is intuitive as neither of the two methods have any learning capacity. The five other machine learning methods range in between (support vector machine: 81.94%, logistic regression: 80.00%, random forest: 79.25%, naïve Bayes: 79.12%, decision tree: 70.43%). Overall, variance in observed accuracy is high both within and across methods, i.e., picking a few datasets (or publications) at random is hardly informative to understand actual method performance.

The above observations suggest that generalizations on method performance necessitate multivariate control of the contextual characteristics of individual applications. Unfortunately, no similar number of publications comparing more recent deep contextual language models exists. Standardized benchmarks such as SST-2 (Socher et al., 2013) make classification accuracy comparable, but are hardly informative, as the lighthouse performances for these few benchmark datasets seldom generalize to real-world applications (Hutson, 2020).

We therefore obtain a like-for-like comparison between transfer learning and the other methods by testing all the publicly available datasets from our meta-analytic sample (covering close to 40% of experiments and more than 400,000 labeled observations). We use these datasets to test four popular and conceptually diverse language models: ULMFiT (Howard, Jeremy, Ruder, & Sebastian, 2018), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019). They represent the main model architectures (long short-term memory- versus transformer-based, Vaswani et al., 2017), different pre-training objectives (next-word prediction, masked, or permutation language modeling), and different pre-training dataset sizes, ranging from 180 MB for ULMFiT to nearly 160 GB for RoBERTa.

According to Fig. 4, the overall median accuracy across all datasets is 88.64%. This is significantly below the values of 95 to 97% suggested by the popular GLUE benchmark (on the simple, standardized two-class SST-2 dataset) for these methods. Across all datasets, RoBERTa achieves the highest average accuracy (92.0% for two-class tasks, 79.1% for three-class tasks) and also the highest accuracy for 12 out of 19 datasets, suggesting larger models pre-trained on more text data can capture more relevant nuances of language that are also relevant for sentiment classification. At the same time, RoBERTa's performance differences relative to XLNet are mostly small. Both methods perform better on average than ULMFiT, which also exhibits the highest variance in performance across applications.

Across all methods, we find a spread of more than 10 percentage points in accuracy even across the two-class datasets. Apparently, accuracy of the more advanced methods is similarly context-dependent to that of the methods prevalent in applied research so far, i.e., traditional machine learning and lexicons. Furthermore, we find average declines in performance between 11.8 (RoBERTa) and 16.4 (ULMFiT) percentage points when switching from two- to three-class sentiment analysis. This suggests that state-of-the-art leaderboard values can be out of reach for many practical applications with less standard-

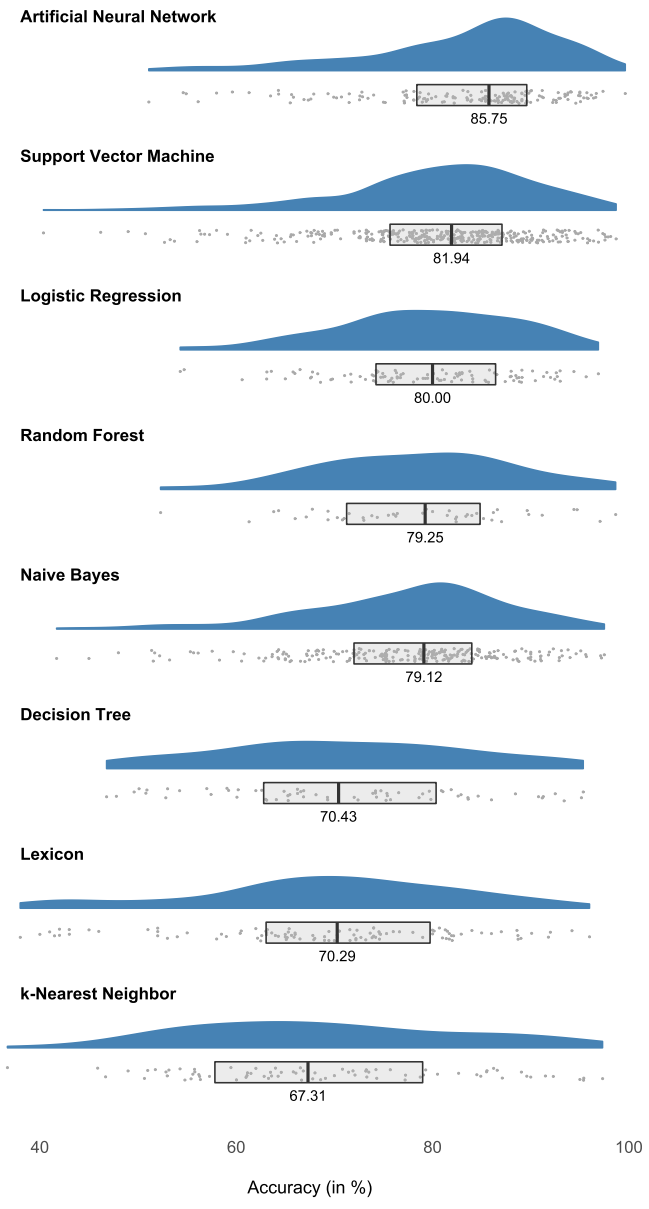

Fig. 3. Distribution of Sentiment Analysis Accuracy by Method.
Note: Each grey dot represents a reported accuracy level per method. N = 1,169 experimental results.

ized datasets, especially when three (or more) sentiment classes are of substantive interest. Taken together, considering both attainable accuracy and robustness, RoBERTa appears a particularly promising candidate model for sentiment analysis.

Fig. 4 does not suggest a clear link between accuracy and sample size in terms of manually labeled sentiment scores. We observe high levels of accuracy for both large and small datasets. To investigate this in more detail, we perform experiments with smaller sample sizes (i.e., $1,000, 5,000, 10,000$, and $50,000$) gathered from the four largest datasets (see Web Appendix for details). The average accuracy for both RoBERTa and XLNet drops by only about 1.5 percentage points, even for the smallest sample size. The two language models with comparatively smaller text corpora for pre-training (ULMFiT and BERT) are up to 5 percentage points less accurate when applied to small samples, suggesting pre-training is particularly beneficial when labeled training data is scarce and that transfer learning has diminishing marginal returns in terms of classification accuracy.

To understand what to expect for individual application scenarios, we next consolidate our findings in multiple regressions and benchmark the different methods controlling for the main drivers we identified in Fig. 1 as well the other control variables mentioned earlier.

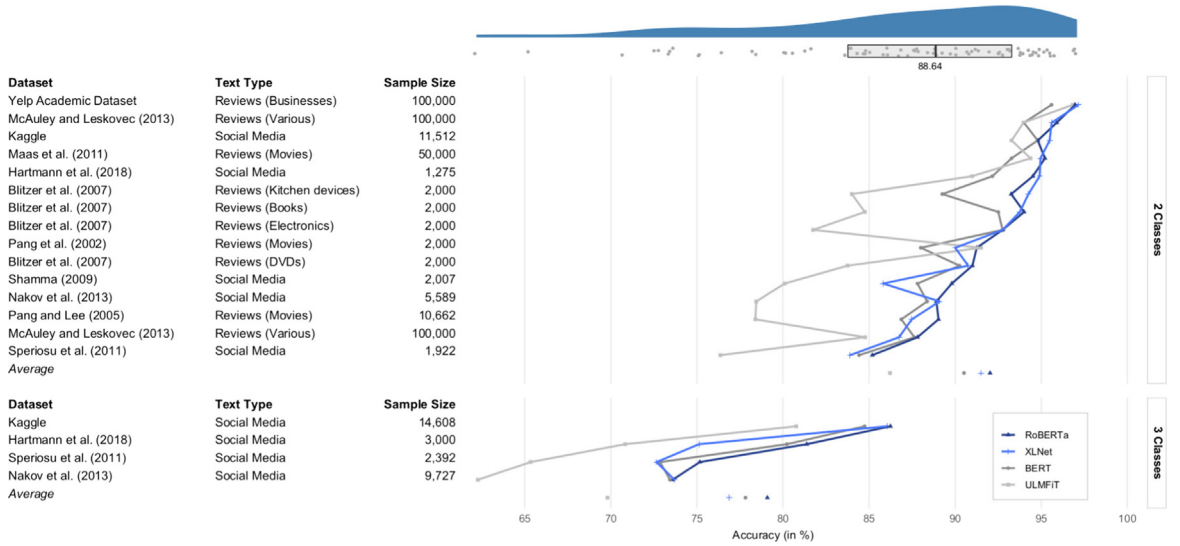

Fig. 4. Classification Accuracy of Transfer Learning Models.
Note: Boxplot shows inter-quartile range and median. Label refers to median. Sorted by best accuracy across language model architectures. Please refer to Table 2 in the Web Appendix for details.

Consolidated Results

Table 1 reports the results of a multi-level regression based on our empirical framework. Note that traditional machine learning methods have very similar accuracy and grouping them reduces R-squared only marginally. We therefore group these methods, allowing us to test relevant interactions between the main method types (lexicons, traditional machine learning, transfer learning) and contextual factors. In the interest of interpretation, we estimate a linear probability model (Wooldridge, 2019) with sentiment analysis accuracy as the dependent variable (results based on logit-transformed accuracy are similar).

Model (1) includes only the three main method types (lexicons as the reference, traditional machine learning, and transfer learning). Given we estimate a linear probability model, the coefficients can be intuitively interpreted as percentage point differences of classification accuracy. On average, lexicons (reference category) classify only 69.23% of texts correctly, which is well below the conventional 80% threshold that is often referred to (Berger et al., 2020). Transfer learning emerges as the best performing method type, outperforming lexicons and traditional machine learning by 20.22 ($p < .01$) and 10.77 ($p < .01$) percentage points, respectively. On average, the better intrepretability of lexicons therefore costs more than 20 percentage points of additional false classifications relative to transfer learning. This likely represents a relevant trade-off that should be considered in a conscious choice for each application.

Note, however, that Model (1) is limited to method choice and explains relatively little variance in the data (7% of the variation in sentiment analysis accuracy). This suggests further dataset and implementation factors need to be considered to better understand in which contexts how much accuracy can be attained. Model (2) therefore adds the main drivers of sentiment analysis accuracy from our empirical framework as well as additional control variables to our regression models. This improves R-squared to 63%. In terms of the main drivers of sentiment analysis accuracy, we make several noteworthy observations. First, we find that more classes, i.e., switching from binary to three-class sentiment analysis, make classification more challenging. Specifically, three-class sentiment analysis yields a predicted accuracy that is 12.72 ($p < .01$) percentage points lower compared to two-class sentiment analysis (i.e., the difference of about what would be expected due to chance alone). Second, document-level analysis is indeed better suited for automated sentiment classification than brief sentences, with 5.03 percentage points ($p < .05$) more false classifications for sentence-level sentiment coding. Third, more sentiment-labeled training data improves accuracy. Specifically, the natural logarithm (ln) of each additional observation improves classification performance by .79 percentage points ($p < .01$), i.e., moving from 100 to 10,000 labeled text documents improves accuracy by 3.6 percentage points on average.

Note that individual methods might be affected differently by these factors. We therefore test for relevant interactions and find the impact of both text length and sample size differs significantly across methods (Model 3). Specifically, the effect of document- versus sentence-level texts is attenuated for traditional machine learning methods ($-4.43, p = .053$). This is likely due to the fact that the variety of words (i.e., terms) and consequently document-term-matrix is often larger for a document- than a sentence-level analysis. Keeping training data size constant, more parameters mean fewer observations for each, which can result in less accurate parameter estimates that offset some of the advantages of having more text. Lexicons have no such issue. Intuitively, they benefit most from document-level sentiment analysis, as longer texts contain more

Table 1
Multi-Level Regression Results.

Dependent variable: Accuracy	(1)	(2)	(3)
Method Choice			
Lexicons (Constant)	69.23***	63.56***	61.80***
	(1.09)	(2.06)	(2.32)
Traditional Machine Learning	9.45***	9.96***	13.94***
	(1.14)	(.93)	(2.08)
Transfer Learning	20.22***	21.42***	23.35***
	(2.80)	(2.21)	(6.72)
Main Drivers			
Number of Classes: 3 (Ref: 2)		−12.72***	−12.72***
		(1.03)	(1.03)
Text Length: Document (Ref: Sentence)		5.03***	7.45***
		(1.62)	(2.22)
Sample Size (ln)		.79***	-.57
		(.23)	(.49)
Interactions			
Traditional Machine Learning × Text Level: Document			−4.43*
			(2.29)
Transfer Learning × Text Level: Document			−1.67
			(6.75)
Traditional Machine Learning × Sample Size (ln)			1.70***
			(.52)
Transfer Learning × Sample Size (ln)			.79
			(1.49)
Control Variables: Dataset			
Text Source: Review (Ref: Other)		2.64**	2.32**
		(1.07)	(1.07)
Dataset Balance: Imbalanced (Ref: Balanced)		1.22	1.12
		(.92)	(.92)
Language: English (Ref: Other)		.34	.09
		(1.03)	(1.03)
Control Variables: Publication			
Hyperparameter Tuning: Yes (Ref: No)		3.20***	3.06***
		(.76)	(.77)
Number of Experiments (ln)		2.69***	2.85***
		(.35)	(.36)
Publication Year		.37**	.29*
		(.15)	(.15)
Journal Publication (Ref: Other)		.72	.83
		(.76)	(.76)
Citations per Year (ln)		.67**	.52*
		(.32)	(.32)
Observations	1,188	1,188	1,188
R^2	.07	.63	.63

Note. *p < .1; **p < .05; ***p < .01.
Standard errors in parentheses. All numeric variables are mean-centered.

sentiment signals, which the lexicon can pick up on ($7.45, p < .01$). Transfer learning models that represent text as contextual, dense embeddings can deal better with increasing vocabulary sizes. Unlike traditional machine learning, they do not differ significantly from lexicons in this regard. This means that the advantages of transfer learning over traditional machine learning are even larger for longer texts, which traditional machine learning cannot leverage as well.

In terms of the available training data, we also observe nuanced differences between the three main method types. The performance of lexicons is not significantly affected by the amount of training data ($-.71, p = .24$), which is intuitive as – unlike traditional machine learning and transfer learning – no bottom-up learning process from the data takes place. Interestingly, transfer learning does not significantly benefit from additional training data either, presumably due to the pre-training on large text corpora ($.79, p = .60$). In contrast, traditional machine learning benefits significantly from more training data, with 5.2 percentage points improvements in accuracy when moving from 100 to $10,000$ labeled documents ($1.13, p < .01$). As no transfer learning takes place, the methods must rely exclusively on the information contained in the annotated training data. Clearly, in this situation, more data pays off more compared to a method that can benefit from transfer learning.

Results also suggest there is little improvement in accuracy over time when controlling for method choice (less than a third of a percentage point per year). Put differently, incremental method improvements seem to have played a less important role for sentiment analysis progress than the new method architectures that have become available (from lexicons to traditional machine learning to transfer learning).

The remaining significant control variables are all in the expected directions. Conducting hyperparameter tuning improves classification accuracy by 3.06 percentage points ($p < .01$). Similarly, running more experiments is positively associated with classification accuracy ($2.85, p < .01$). In contrast to Hartmann et al. (2019) and Alantari et al. (2022), the number and diversity of the datasets included in our meta-analytic sample allows us to meaningfully compare different types of data. We find that sentiment analysis of product reviews such as from Amazon is more accurate than sentiment analysis of social media data, which tend to be noisier and less structured ($2.32, p < .05$). This also suggests that benchmark values based on more structured review data can be misleading for other types of applications as accuracy values may be inflated. Taken together, our model allows for meaningful a priori performance expectations – but this requires variance in datasets to take dataset characteristics into account. We will compute predicted values for relevant scenarios next.

Discussion

Sentiment analysis is likely the most prominent use case for natural language processing and text classification, drawing much attention by both scholars and practitioners for a wide variety of applications (Hirschberg & Manning, 2015; Berger et al., 2020; Wang et al., 2022; Sukhwal et al., 2022). To date, marketing research has predominantly relied on dictionary-based tools to extract sentiment from text data. These have clear advantages in terms of interpretability, but this comes at the cost of accuracy and we have quantified this trade-off in this research. These differences in accuracy can translate to lower explanatory power (Kübler, Colicev, & Pauwels, 2020) and erroneous substantive conclusions (Jaidka et al., 2020). Moreover, Hartmann et al. (2019) find that suboptimal method choice can have substantial economic consequences even for small- to medium-sized companies, i.e., the exact performance differences matter when weighing relative performance.

This paper provided a rigorous and comprehensive assessment of the available sentiment analysis methods. Specifically, we compared intrepretable lexicons with traditional machine learning, and more recent transfer learning models that have lower levels of interpretability. The field of natural language processing is dynamic and progresses rapidly. Despite this, our empirical findings suggest leaps of performance are mainly driven by new model architectures, with incremental improvements of existing methods playing a subordinate role. Since fundamentally new architectures emerge less often, it is possible to see through the clutter of more frequent minor advancements and identify several long-term trends. This is particularly relevant for applied researchers who are not in the position to comprehensively test the various, sometimes difficult-to-implement methods, but require readily implementable standards to produce comparable research findings. When applying sentiment analysis, it may be tempting to follow in the footsteps of prior publications. However, when technological progress is rapid, such choices can be dysfunctional. Moreover, consulting selected computer science papers or standardized benchmarks may not serve applied researchers well, as accuracy levels largely depend on contextual factors, rendering a like-for-like comparison across papers impossible.

To remedy this problem, we conducted a large-scale assessment of sentiment analysis accuracy, meta-analytically analyzing more than $1,100$ experimental results based on 272 distinct datasets. In addition, we compared four state-of-the-art language models on 19 distinct sentiment analysis tasks from different application contexts. Overall, transfer learning using state-of-the-art language models emerged as the winning method type in terms of sentiment analysis accuracy, on average outperforming lexicons and traditional machine learning methods by more than 20 and 10 percentage points, respectively.

However, these average values do not permit reasonable performance expectations for individual applications as performances also varies by sample size, text length, and the amount of sentiment classes. We therefore summarize our findings in a decision tree (Fig. 5) that serves as a guide for marketing scholars on what to expect from each sentiment analysis method given the application context. The decision tree builds on our empirical framework and – at its leaves – presents predicted classification accuracies based on our consolidated multi-level regression analyses. Specifically, we run a post-estimation analysis, considering both the interactions with sample size and text type (document versus sentence), keeping all continuous control variables at their mean and categorical variables at their reference category. This allows us to reflect the context-contingent performance of the three method alternatives.

When running sentiment analysis, researchers must decide ex ante if they seek to apply an interpretable or non-interpretable method. This decision should consider the impact on accuracy to consciously trade off both dimensions. Lexicons, as top-down methods that identify hypothesized patterns in language and link them to specific words, are inherently interpretable, i.e., each classification can be traced back to the combination of words that led to it. On the other hand, the inherent limitations of dealing with misspellings and noise, short texts, and large vocabularies result in much lower classification accuracy than all other method classes. This accuracy-interpretability trade-off needs to be quantified. We find that entering the interpretability path (right branch in the decision tree) with lexicons, accuracy values can be as low as 50% and reach a maximum of only 72%, all below the conventional 80% threshold (indicated by yellow color; red reflects accuracy values of 60% and lower). Despite that, lexicons remain among the most popular method choice in applied research across disciplines. However, according to Fig. 5, standard lexicons are barely acceptable in terms of accuracy for three class sentiment coding and sentence-level analysis in particular.

Whenever interpretability is of less concern, transfer learning and traditional machine learning are the best performing methods. For sentence-level data, both traditional machine learning and transfer learning can produce acceptable results, in particular for two-class sentiment. For data on a document level, transfer learning has the strongest advantage and is clearly and significantly better than traditional machine learning with above 80% accuracy in all application scenarios and top

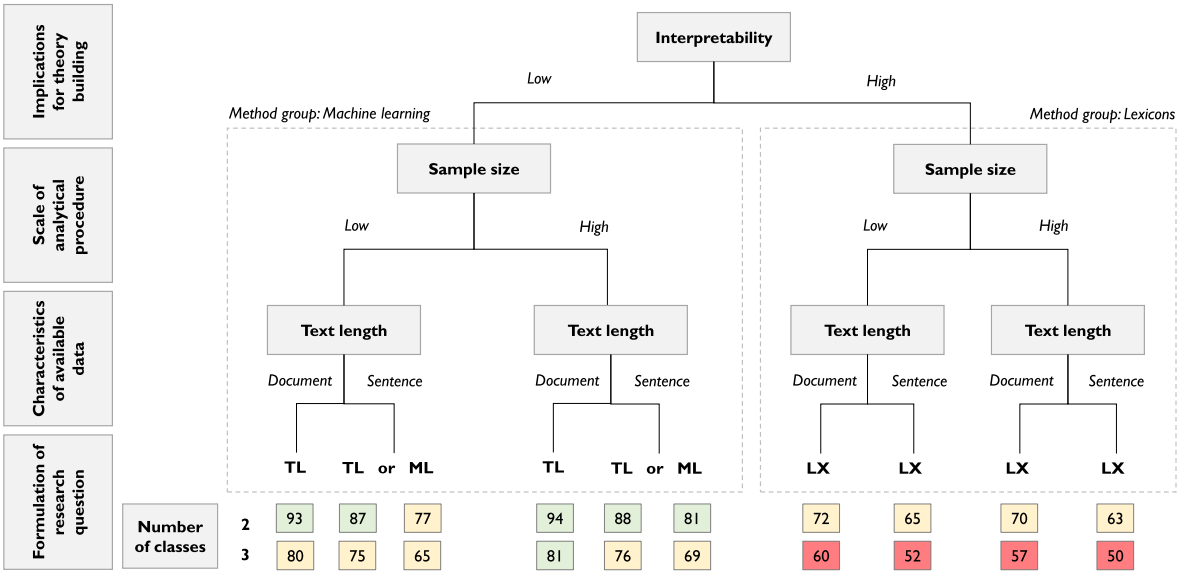

Fig. 5. Decision Tree Summarizing Consolidated Results.

Note: TL = Transfer Learning, ML = Traditional Machine Learning, LX = Lexicons. Boxes at the leaves of the tree contain predicted accuracy levels, contingent on method choice and the main drivers. Predictions are based on Model (3) from Table 1. Values are color-coded for ease of interpretation: green if > 80% accuracy, yellow if > 60% and <= 80%, red if <= 60%. For sentence-level tasks with low interpretability, the 90% confidence intervals of TL and ML overlap. Hence, the decision tree reports the predicted performance levels for both. For post-estimation results, small and large sample size are set at 1,000 and 50,000, respectively.

performances with hit rates above 90% for two-class sentiment. Our recommendation is therefore to apply transfer learning to all document-level analyses when interpretability is of little concern. Traditional machine learning can be considered for sentence-level analysis, in particular when binary classification that needs no detailed interpretation is of interest. When interpretability and theory building are more important, lexicons are the best choice, but researchers should be mindful of low performance for three-class sentiment on the sentence level in particular (e.g., when classifying short review titles or tweets as positive, negative, and neutral).

To illustrate these conclusions, consider two example scenarios. First, consider a research group which seeks to analyze the sentiment of an open-ended text field of a survey to validate the scale-based measures and to extract additional qualitative insights (as suggested by Berger et al., 2020). The research group's core priorities are ease of use and interpretability, i.e., an intuitive sentiment-word mapping based on theoretical considerations (Humphreys, 2019). Due to the guided nature of the questions in the survey (versus unsolicited social media posts) as well as the higher motivation of the respondents who are monetarily incentivized to participate, the texts are expected to be relatively structured, well written, and feature relatively few misspellings and little informal language. Additionally, the costs of incorrectly classifying texts are relatively low, making accuracy of lower concern relative to other real-world applications. In such a scenario, lexicons are the preferred choice for this research group. Traditional machine learning, on the other hand, will often result in a suboptimal compromise in between lexicons that are easier to interpret and transfer learning that promises more accurate classification. When choosing an available off-the-shelf lexicon, we recommend assessing performance based on manually annotated data. Given the variance in accuracy and the high level of context sensitivity, independent manual annotation is an effort well spent to understand actual lexicon performance.

Next, consider a collaboration between marketing researchers and practitioners, seeking to classify the sentiment and severity of customer complaints (similar to the case study described in Hartmann et al. (2019)). The objective is to build an automatic tool that can efficiently classify incoming complaints and assigns them to different teams. The complaints are prioritized according to their severity. A wrong assignment has economic consequences as it (a) costs time to forward it to the correct team, and (b) a slow handling of very negative complaints may increase customer churn. Hence, obtaining the highest possible accuracy is a top priority. Additionally, the classifier needs to be able to deal well with noisy, unstructured text beyond obvious terms that a standard lexicon would contain. Consider the example: "Your service isnt good!", which SiEBERT correctly classifies as negative, but LIWC-22 as positive (positive tone: 25, negative tone: 0). Since interpretations of underlying classification rules are of little interest and false classifications are costly, a method should be chosen that is most likely to maximize accuracy. This is best accomplished with the advances of transfer learning by fine-tuning an own method (either domain pre-trained or general purpose). According to Fig. 4, RoBERTa and XLNet are the most promising candidates.

Taken together, most applications will likely get the best mix of benefits with the extremes of either simple lexicon-based word counts (due to their high interpretability) or fine-tuning a large-scale pre-trained model (due to their high classification accuracy). If fine-tuning is not possible or desired, state-of-the-art language models are available for application in an off-the-shelf manner that is comparable to what marketing researchers are used to from popular off-the-shelf lexicons such as VADER.

With transfer learning, additional benefits of model sharing become apparent. Specifically, feedback loops are now possible where researchers benefit from prior researchers' work by building on and further fine-tuning domain-specific models (see Fig. 2). This will ultimately allow better off-the-shelf language models and enable researchers to utilize the same coding principles across datasets and publications without making trade-offs in terms of accuracy. Moreover, fine-tuning existing models comes at much lower computational (and environmental) costs compared to training models from scratch. Ongoing efforts in model interpretability are likely to further enhance the appeal of transfer learning. These considerations suggest the overall relative advantages will further shift from lexicons to transfer learning.

While new method innovations are likely to emerge (e.g., multi-modal methods for transfer learning, i.e., leveraging visual information to improve performance on tasks involving textual data, or zero- and few-shot learning, lowering the need for sentiment-labeled training data even further), the general differentiation between lexicons, traditional machine learning, and transfer learning and their strengths and weaknesses in our framework is likely to sustain. It is safe to assume that leveraging knowledge outside the training data by fine-tuning a pre-trained model will continue to outperform models confined to only the training data both in terms of effectiveness and efficiency. To facilitate the use of language models, we provide a readily available pre-trained language model based on the data of this research, called SiEBERT. SiEBERT can be applied like a lexicon in an off-the-shelf manner. It can also be fine-tuned for a variety of specific sentiment-related tasks, e.g., to classify more granular emotions expressed in reviews (anger, joy, disgust) or to detect hate speech on social media. We caution that commercial services with similar technologies are often a black box to researchers and are also modified from time to time, which can undermine replication efforts. We hope the results of this research and the scripts we provide facilitate the use of open-source models that can be shared within the academic community and allow for informed decisions of when and how to meaningfully conduct sentiment analysis.

Funding

This work was funded by the German Research Foundation (DFG) research unit 1452, HE 6703/1–2, and by research grant number 460037581. Jochen Hartmann is grateful for the grant "Challenging the Boundaries of Natural Language Processing" from the Claussen-Simon Foundation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijresmar.2022.05.005.

References

Alantari, Huwail J., Currim, Imran S., Deng, Yiting, & Singh, Sameer (2022). An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. *International Journal of Research in Marketing, 39*(1), 1–19.

Andrews, Rick L., Ainslie, Andrew, & Currim, Imran S. (2002). An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. *Journal of Marketing Research, 39*(4), 479–487.

Berger, Jonah, Humphreys, Ashlee, Ludwig, Stephan, Moe, Wendy W., Netzer, Oded, & Schweidel, David A. (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing, 84*(1), 1–25.

Bremer, Lucas, Heitmann, Mark, & Schreiner, Thomas (2016). When and how to infer heuristic consideration set rules ofconsumers. *International Journal of Research in Marketing, 34*(2), 516–535.

Chakraborty, Ishita, Kim, Minkyung, & Sudhir, K. (2022). Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes. *Journal of Marketing Research, 59*(3), 600–622.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, https://arxiv.org/abs/1810.04805.

Ghosal, Tirthankar, Das, Sajal K., Bhattacharjee, Saprativa (2015). "Sentiment analysis on (Bengali horoscope) corpus." In 12th IEEE International Conference Electronics, Energy, Environment, Communication, Computer, Control New Delhi, India IEEE, 1–6.

Hansen, Nele, Kupfer, Ann Kristin, & Hennig-Thurau, Thorsten (2018). Brand crises in the digital age: The short- and long-term effects of social media firestorms on consumers and brands. *International Journal of Research in Marketing, 35*(4), 557–574.

Hartmann, Jochen, Huppertz, Juliana, Schamp, Christina, & Heitmann, Mark (2019). Comparing automated text classification methods. *International Journal of Research in Marketing, 36*(1), 20–38.

Hartmann, Jochen, Heitmann, Mark, Schamp, Christina, & Netzer, Oded (2021). The power of brand selfies. *Journal of Marketing Research, 58*(6), 1159–1177.

Herhausen, Dennis, Ludwig, S., Grewal, D., Wulf, J., & Schoegel, M. (2019). Detecting, preventing, and mitigating online firestorms in brand communities. *Journal of Marketing, 83*(3), 1–21.

Hewett, Kelly, Rand, William, Rust, Roland T., & van Heerde, Harald J. (2016). Brand buzz in the echoverse. *Journal of Marketing, 80*(3), 1–24.

Hirschberg, Julia, & Manning, Christopher D. (2015). Advances in natural language processing. *Science, 349*(6245), 261–266.

Homburg, Christian, Ehm, Laura, & Artz, Martin (2015). Measuring and managing consumer sentiment in an online community environment. *Journal of Marketing Research, 52*(5), 629–641.
Howard, Jeremy, Ruder, Sebastian (2018), "Universal language model fine-tuning for text classification." In Proceedings of the 56th Annual Meeting of the ACL Melbourne, Australia (pp. 328–339).
Humphreys, Ashlee (2019). "Automated text analysis," in Handbook of Market Research Springer, 1–32.
Hutson, Matthew (2020). "Eye-catching advances in some AI fields are not real," https://www.sciencemag.org/news/2020/05/eye-catching-advances-some-ai-fields-are-not-real.
Hutto, C. J. J. & Gilbert, Eric (2014), "VADER: A parsimonious rule-based model for sentiment analysis of social media text." In Eighth International AAAI Conference on Weblogs and Social Media Ann Arbor, USA Association for the Advancement of Artificial Intelligence, (pp. 216–225).
Jaidka, Kokil, Salvatore Giorgi, H., Schwartz, Andrew, Kern, Margaret L., Ungar, Lyle H., & Eichstaedt, Johannes C. (2020). Estimating geographic subjective well-being from Twitter. *Proceedings of the National Academy of Sciences, 117*(19), 10165–10171.
Kübler, Raoul V., Colicev, Anatoli, & Pauwels, Koen H. (2020). Social media's impact on the consumer mindset. *Journal of Interactive Marketing, 50*, 136–155.
Lafreniere, Katherine C., Moore, Sarah G., Fisher, Robert J. (2022), "The power of profanity: The meaning and impact of swear words in word of mouth," Journal of Marketing Research. in press. https://journals.sagepub.com/doi/full/10.1177/00222437221078606.
Liu, Yang, Bi, Jian Wu, & Fan, Zhi Ping (2017). Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms. *Expert Systems with Applications, 80*, 323–339.
Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, Stoyanov, Veselin (2019). "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, https://arxiv.org/abs/1907.11692.
Manning, Christopher D., Clark, Kevin, Hewitt, John, Khandelwal, Urvashi, & Levy, Omer (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences, 117*(48), 30046–30054.
Matalon, Yogev, Magdaci, Ofir, Almozlino, Adam, & Yamin, Dan (2021). Using sentiment analysis to predict opinion inversion in Tweets of political communication. *Scientific Reports, 11*(1), 7250.
Meire, Matthijs, Hewett, Kelly, Ballings, Michel, Kumar, V., & Van den Poel, Dirk (2019). The role of marketer-generated content in customer engagement marketing. *Journal of Marketing, 83*(6), 21–42.
Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg, Dean, Jeffrey (2013). "Distributed representations of words and phrases and their compositionality." In Advances in Neural Information Processing Systems 26 (NIPS 2013).
Netzer, Oded, Lemaire, Alain, & Herzenstein, Michal (2019). When words sweat: Identifying signals for loan default in the text of loan applications. *Journal of Marketing Research, 56*(6), 960–980.
Ordabayeva, Nailya, Cavanaugh, Lisa A., Dahl, Darren W. (2022), "The upside of negative: Social distance in online reviews of identity-relevant brands," Journal of Marketing, In Press.
Ordenes, Francisco Villarroel, Grewal, Dhruv, Ludwig, Stephan, De Ruyter, Ko, Mahr, Dominik, & Wetzels, Martin (2019). Cutting through content clutter: How speech and image acts drive consumer sharing of social media brand messages. *Journal of Consumer Research, 45*(5), 988–1012.
Ordenes, Francisco Villarroel, Ludwig, Stephan, De Ruyter, Ko, Grewal, Dhruv, & Wetzels, Martin (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.
Pennebaker, James W., Boyd, Ryan L., Jordan, Kayla, Blackburn, Kate (2015), "The development and psychometric properties of LIWC2015," Technical report University of Texas at Austin Austin, USA.
Pennington, Jeffrey, Socher, Richard, Manning, Christopher D (2014). "GloVe: Global vectors for word representation." In Proceedings of the 2014 Conference on EMNLP Doha, Qatar (pp. 1532–1543).
Puranam, Dinesh, Kadiyali, Vrinda, & Narayan, Vishal (2021). The impact of increase in minimum wages on consumer perceptions of service: A transformer model of online restaurant reviews. *Marketing Science, 40*(5), 813–1007.
Ravi, Kumar, Ravi, Vadlamani, Gautam, Chandan (2015). "Online and semi-online sentiment classification." In International Conference on Computing, Communication and Automation, ICCCA 2015 number August Greater Noida, India IEEE (pp. 938–943).
Reinartz, Werner, Haenlein, Michael, & Henseler, Jörg (2009). An empirical comparison of the efficacy of covariance-based and variance-based SEM. *International Journal of Research in Marketing, 26*(4), 332–344.
Ruder, Sebastian, Matthew Peters, Swabha Swayamdipta and Thomas Wolf (2019), "Transfer learning in natural language processing tutorial," in 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Tutorial Abstracts Stroudsburg, PA, USA Association for Computational Linguistics, 15–18.
Socher, Richard, A Perelygin, J Wu, J Chuang, C Manning, A Ng and C Potts (2013), "Recursive deep models for semantic compositionality over a sentiment treebank," in Proceedings of the 2013 Conference on EMNLP Seattle, USA, 1631–1642.
Sukhwal, Prakash Chandra and Atreyi Kankanhalli (2022), "Determining containment policy impacts on public sentiment during the pandemic using social media data," Proceedings of the National Academy of Sciences, 119 (19).
Tang, Tanya, Fang, Eric, & Wang, Feng (2014). Is neutral really neutral? The effects of neutral user-generated content on product sales. *Journal of Marketing, 78*(4), 41–58.
Timoshenko, Artem, & Hauser, John R. (2019). Identifying customer needs from user-generated content. *Marketing Science, 38*(1), 1–192.
Tirunillai, Seshadri, & Tellis, Gerard J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science, 31*(2), 195–368.
Tunstall, Lewis, von Werra, Leandro, Wolf, Thomas (2022). Natural Language Processing with Transformers O'Reilly Media Inc.
Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, Polosukhin, Illia (2017). "Attention is all you need." In Proceedings of the 31st Conference on Neural Information Processing Systems Long Beach, USA.
Wang, Jianghao, Fan, Yichun, Palacios, Juan, Chai, Yuchen, Guetta-Jeanrenaud, Nicolas, Obradovich, Nick, Zhou, Chenghu, & Zheng, Siqi (2022). Global evidence of expressed sentiment alterations during the COVID-19 pandemic. *Nature Human Behaviour, 6*(3), 349–358.
Winkler, Stephan, Schaller, Susanne, Dorfer, Viktoria, Affenzeller, Michael, Petz, Gerald, & Karpowicz, Michal (2015). Data-based prediction of sentiments using heterogeneous model ensembles. *Soft Computing, 19*(12), 3401–3412.
Wooldridge, Jeffrey M. (2019). *Introductory econometrics: A modern approach* (7th edn). South Western Educ Pub.
Yang, Zhilin, Dai, Zihang, Yang, Yiming, Carbonell, Jaime, Salakhutdinov, Ruslan, Le, Quoc V. (2019), "XLNet: Generalized autoregressive pretraining for language understanding," arXiv preprint arXiv:1906.08237, https://arxiv.org/abs/1906.08237.