**Web Appendix**

1. Marketing research applying automated text classification 2
2. Overview of comparative studies in computer science 7
3. Transfer learning example 8
4. Exemplary text documents for each social media type 10
5. Preprocessing and method specifications 12
6. R script for text classification 15
7. Correlations based on classification accuracy 16
8. Details on cost computation for application scenarios 16

**A. Marketing research applying automated text classification**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Citation** | **Task (Classes)** | **DV** | **Dataset types**  **(Source)** | **Text length** | **Method(s)** | **Findings related to classification task** |
| Barasch & Berger (2014) | Sentiment  (pos, neg) | Likelihood of sharing | Free talks about restaurants and self-presentation | Medium | DICT (LIWC) | Relative to communication to few recipients, broadcasting to many makes individuals use less negative speech |
| Berger & Milkman (2012) | Sentiment  (pos) | Virality | News articles (New York Times) | Long | DICT (LIWC) | Arousing content has a positive impact on virality |
| Borah & Tellis (2016) | Sentiment  (pos, neg) | Negative chatter, financial performance | Consumer posts (forums) | Medium | Undisclosed ML method | Negative chatter about a brand increases negative chatter about competing brands |
| Chen (2017) | Sentiment  (pos, neg) | Sentiment (net positivity index) | Reviews (Yelp) | Medium | DICT (LIWC) | Reviews on Yelp become more negative as the number of friends increases |
| Colicev et al. (2018) | Sentiment  (pos, neg) | Brand awareness,  satisfaction | Posts from social network (Facebook) | Short | NB | Valence of social media posts is related to both brand awareness and customer satisfaction |
| Das & Chen (2007) | Sentiment  (pos, neg, neut) | Stock prices | Comments (Yahoo) | Medium | NB, different DICT | Social media sentiment is positively related to stock returns |
| Heimbach & Hinz (2016) | Sentiment  (pos) | Virality | News articles (Spiegel Online) | Long | DICT (LIWC) | There is an inverse U-shaped relationship between positive content and virality |
| Hennig-Thurau et al. (2015) | Sentiment  (pos, neg) | Movie adoption | Movie-related tweets (Twitter) | Short | SVM | Negative WOM has a negative impact on movie adoption |
| Hewett et al. (2016) | Sentiment  (pos, neg) | Customer deposits | Echoverse of corporate communication, news stories, UGC | Long | DICT (LIWC) | Consumer and press release sentiment drive financial product demand |
| **Citation** | **Task (Classes)** | **DV** | **Dataset types**  **(Source)** | **Text length** | **Method(s)** | **Findings related to classification task** |
| Homburg et al. (2015) | Sentiment  (pos, neg) | Sentiment | Posts from online forums (e.g., do-it-yourself, airline) | Medium | SVM | Inverse U-shaped impact of firm engagement on sentiment of consumer comments |
| Ilhan et al. (2018) | Sentiment  (pos, neg, neut) | Volume, sentiment | Posts from social network (Facebook); Reviews (Amazon, Yelp, Twitter) | Short; Medium | SVM | Negative sentiment from fans of competing brands results improves social media performance of attacked brand due to fans defending their own brand |
| Klostermann et al. (2018) | Sentiment  (pos, neg, neut) | Brand perceptions | Posts from social network (Instagram) | Short | DICT (VADER) | Customer evaluations are contingent on product type and situational context; sentiment analysis can help identify strengths and weaknesses of products |
| Kuhnen & Niessen (2012) | Sentiment  (neg) | CEO compensation | News articles (Factiva) | Long | DICT (LIWC) | Negative press coverage of executive compensation has positive impact on CEO salary and negative impact on option grants |
| Ludwig et al. (2013) | Sentiment  (pos, neg) | Conversion rate | Reviews (Amazon) | Medium | DICT (LIWC) | Positive content has a non-linear impact on conversion rates |
| Marchand et al. (2017) | Sentiment  (pos, neg, neut) | New product success (sales) | User comments (Twitter); Reviews (Amazon) | Short; Medium | SVM | Valence of consumer reviews drives purchases only after an initial period after product introduction |
| Ordenes et al. (2017) | Sentiment  (pos, neg) | Product rating | Reviews (Amazon, BN, TripAdvisor) | Medium | DICT (LIWC) | Consumer evaluations are driven by weekly sentiment changes in consumer reviews |
| Rooderkerk & Pauwels (2016) | Sentiment  (pos, neg) | Comments | Posts from online discussion forum (LinkedIn) | Medium | DICT (LIWC) | Posts without user comments contain less negative words |
| **Citation** | **Task (Classes)** | **DV** | **Dataset types**  **(Source)** | **Text length** | **Method(s)** | **Findings related to classification task** |
| Schweidel & Moe (2014) | Sentiment  (pos, neg) | Sentiment | Comments (Converseon) | Medium | DICT (LIWC) | Listening to single social media source may lead to imprecise brand sentiment monitoring |
| Tirunillai & Tellis (2012) | Sentiment  (pos, neg) | Stock returns, risk, trading volume | Reviews (Amazon, Epinions, Yahoo) | Medium | SVM, NB | Negative chatter volume has a negative impact on stock returns and a positive impact on risk and trading volume |
| Tang et al. (2014). | Sentiment  (pos, neg, neut) | Product sales | Comments (Facebook, YouTube) | Short | DICT (SentiStrength) | The impact of neutral UGC on product sales depends on both the type of UGC and the distribution of pos. and neg. UGC |
| Cavanaugh et al. (2015) | Content (love, hope, neutral, compassion) | Prosocial behavior | Emotion induction procedure, choice survey | Long | DICT (LIWC) | Specific positive emotions (e.g., love and hope) predict unique patters of prosocial behavior (e.g., prosocial consumption) |
| Chan & Mogilner (2017) | Content (% of emotion words per participant) | Emotional response (e.g., liking, relationship strength), price | Written description of gift by participants | Medium | DICT (LIWC) | Experiential gifts evoke more emotional reactions and hence strengthen the relationship to the giver more than material gifts |
| Felbermayr & Nanopoulos (2016) | Content (e.g., anger, disgust, fear, surprise) | Review quality (i.e., usefulness) | Reviews (Amazon) | Medium | DICT (NRC) | Trust, joy, and anticipation drive perceived usefulness of reviews |
| Ghose et al. (2012) | Content (subjective, objective) | Hotel ranking | Reviews (Travelocity, Tripadvisor) | Medium | Undisclosed ML method | Positive relationship between objective reviews and hotel demand |
| Hansen et al. (2018). | Content (firestorm y/n) | Brand perception | Firestorm tweets (Twitter) | Short | DICT (customized) | Firestorms linked to product or social failure induce more negative brand associations |
| **Citation** | **Task (Classes)** | **DV** | **Dataset types**  **(Source)** | **Text length** | **Method(s)** | **Findings related to classification task** |
| Huang & Luo (2016) | Content (consider y/n) | Consumer preferences | Digital cameras, tablets and synthetic data | Short | SVM | SVM can be used to elicit consumer preferences for complex products, even for high-dimensional problems |
| Kanuri et al. (2018) | Content (emotionality, cognitive processing) | Link clicks | Posts from social network (Facebook) | Short | DICT (LIWC) | Posting high-arousal content with negative emotions in the evening is less effective than in the morning |
| Kupfer et al. (2018) | Content (authenticity, exclusiveness, persuasiveness) | Movie sales | Posts from social network (Facebook) | Short | DICT (LIWC, customized) | Authenticity, exclusiveness and persuasiveness in product-related posts drives weekly movie sales |
| Lee et al. (2018) | Content (info, brand personality, mixed) | Engagement (e.g., likes) | Posts from social network (Facebook) | Short | SVM, NB, logit | Brand personality-related content (emotional, humorous) drives higher engagement (likes, comments, shares, click-throughs) |
| Netzer et al. (2016) | Content (loan repayment, default) | Default predictions | Loan requests (Prosper) | Medium | NB, logit, decision trees, DICT (LIWC) | Borrowers word usage is predictive of future repayment behavior at a similar scale as their demographic information |
| Ordenes et al. (2018) | Content (asser-tive, expressive, directive) | Consumer sharing | Posts from social network (Facebook, Twitter) | Short | SVM | Directive posts lead to lower levels of consumer sharing compared to assertive or expressive messages |
| Packard et al. (2018) | Content (pronouns) | Customer satisfaction, purchase intent | Customer-firm emails (contact-us link on website) | Long | LIWC | Service agents who refer to themselves using “I” rather than “we” increase satisfaction and purchase intent |
| Rutz et al. (2011) | Content (semantic properties of keywords) | Online shop visitors | Daily paid search traffic (Google AdWords) | Short | DICT (WordNet, customized) | Branded and broader search keywords result in more future visits of an online shop |
| **Citation** | **Task (Classes)** | **DV** | **Dataset types**  **(Source)** | **Text length** | **Method(s)** | **Findings related to classification task** |
| Timoshenko & Hauser (2018) | Content (non-/ informative) | Importance | Reviews (Amazon) | Medium | ANN, SVM | Consumer needs can be identified from UGC using machine learning methods |
| Van Laer et al. (2018) | Content (narrativity categories) | Positive feedback, attitude, purchase intent | Reviews (Tripadvisor) | Medium | DICT (LIWC) | Narrative content and discourse categories in consumer reviews are more persuasive and have a positive effect on consumer response |
| Yadav et al. (2007) | Content (attentional foci of CEOs) | Innovation outcome (e.g., detection speed) | Shareholder letters in annual reports (Compact D/SEC) | Long | DICT (N6, DICTION, customized) | CEO attention has a positive impact on innovation success |
| Yoganarasimhan (2018) | Content (search personalization) | Search quality, (e.g., click-through rate) | Searches (Yandex) | Short | (Boosted) decision trees | Quality of online search results can be improved to a larger extent by training machine learning based on search specific features (e.g., similar search requests of others) rather than individual specific features. |

Note: Text length is categorized as follows: short (1-30 words), medium (31-150 words), long (>150 words).

**B. Overview of comparative studies in computer science**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Classification problem | | Dataset characteristics | | | Text classification methods | | | | | | Performance evaluation | | | Results for marketing metrics |
| Citation | **Task** | **# classes** | **# social media types (datasets)** | **Dataset variety** | **Multi lang.** | **ANN** | **kNN** | **NB** | **RF** | **SVM** | **LIWC** | **Winning text classification methods** | **Signi-ficance tests** | **Statistical inference of text and task types** |  |
| This paper | S&C | 2&3 | 12 (41) | RE, MB, LB, SN, TM | x | x | x | x | x | x | x | RF, NB, ANN | x | x | Customer behavior |
| Annett & Kondrak (2008) | S | 2 | 2 | RE |  |  |  | x |  | x |  | NB |  |  |  |
| Bermingham & Smeaton (2010) | S | 2 | 4 | RE, MB, LB |  |  |  | x |  | x |  | NB, SVM |  |  |  |
| Boiy et al. (2007) | S | 2&3 | 2 | RE, LB |  |  |  | x |  | x |  | ME, SVM |  |  |  |
| Dave et al. (2003) | S | 2 | 2 | RE |  |  |  | x |  | x |  | NB, SVM | x |  |  |
| Dumais et al. (1998) | C | >3 | 1 | NE |  |  |  | x |  | x |  | SVM |  |  |  |
| Durant & Smith (2006) | S | 2 | 1 | LB |  |  |  | x |  | x |  | NB | x |  |  |
| Fang & Zhan (2015) | S | 2&3 | 1 | RE |  |  |  | x | x | x |  | RF, SVM |  |  |  |
| Gautam & Yadav (2014) | S | 2 | 1 | MB |  |  |  | x |  | x |  | NB |  |  |  |
| Go et al. (2009) | S | 2 | 1 | MB |  |  |  | x |  | x |  | NB, SVM |  |  |  |
| Joachims (1998) | C | >3 | 2 | NE, MD |  |  | x | x |  | x |  | SVM | x |  |  |
| Kübler et al. (2017) | S | 2&3 | 2 | SN |  |  |  |  |  | x | x | SVM | x |  | Customer mindset |
| Melville et al. (2009) | S | 2 | 3 | LB |  |  |  | x |  |  |  | NB + lexicon | x |  |  |
| Moraes et al. (2013) | S | 2 | 4 | RE |  | x |  | x |  | x |  | ANN | x |  |  |
| Neethu & Rajasree (2013) | S | 2 | 1 | MB |  |  |  | x |  | x |  | ME, SVM |  |  |  |
| Pang et al. (2002) | S | 2 | 1 | RE |  |  |  | x |  | x |  | SVM |  |  |  |
| Yang (1999) | C | >3 | 5 | NE |  | x | x | x |  |  |  | ANN, kNN, LLSF | x |  |  |
| Yang & Liu (1999) | C | >3 | 1 | NE |  | x | x | x |  | x |  | kNN, LLSF, SVM | x |  |  |
| Ye et al. (2009) | S | 2&3 | 1 | RE |  |  |  | x |  | x |  | N-gram model, SVM | x |  |  |

Note: S = sentiment, C = content, RE = reviews, MB = microblog, LB = long blogs, SN = social network, TM = text messages, NE = news articles, MD = medical documents, ME = maximum entropy, LLSF = linear least squares fit.

**C. Transfer learning example**

To evaluate the performance of the supervised machine learning methods, we use sample sizes in ranges common for comparative studies (e.g., Pang et al. 2002; Boiy et al. 2007; Melville et al. 2009). For some application scenarios (e.g., Kübler et al. 2017), it is possible that constraints such as limited training data exist, so that classifiers must be trained on similar classification tasks, using data from related domains. This approach is called transfer learning (Do & Ng 2006). To illustrate transfer learning with our data, we train all machine learning classifiers on 3,000 Amazon reviews of the “Movies and TV” dataset from McAuley et al. (2015) as our training set and predict the sentiment of 200 examples from our movie reviews dataset (IMD) from the UCI repository (see Table 1). Both datasets are publicly available, allowing researchers to reproduce our results. Although from two distinct sources, both datasets are from a similar domain, i.e., movie reviews. Before training our classifiers, we transform the star ratings to binary sentiment classes with a one- and five-star rating representing negative and positive sentiment.

Overall, the findings resemble our results from training all classifiers on 80% of the IMDb dataset (IMD) and predicting a 20% out-of-sample test set. In both cases NB produces the highest accuracy with 78.0% using IMD data compared to 71.5% using Amazon reviews (significantly higher than all four methods at *p* < .05 using paired *t*-tests). However, all methods perform worse when trained on the Amazon data, although the training set size is three times larger, implying that sentiment classification is a highly domain-specific problem (Boiy et al. 2007). Specifically, the accuracies are 67.5%, 62.0%, 67.0%, and 66.5% for ANN, kNN, RF, and SVM, respectively.

While transfer learning clearly has merits for selected use cases, manually annotating an amount of 800 observations as training data should pose no obstacle for most applications. For example, Lee et al. (2018), argue in favor of Amazon Mechanical Turk and demonstrate how to utilize those tags in combination with supervised machine learning methods. In our applications, a smaller training set from the actual text source of interest outperforms larger datasets form other but highly similar domains. Moreover, finding suitable labeled data applicable for transfer learning can be a challenge, especially for content classification where application-specific, custom classes are of interest. Transfer learning is likely to be of high value when classes share identical meanings across applications (e.g., sentiment) and text structures are highly comparable. Since social media texts are often highly heterogeneous, transfer learning is less often applied than in other machine learning applications, e.g., for object recognition in digital images where objective and immutable characteristics are associated with each object (Guillaumin & Ferrari 2012).

|  |  |  |  |
| --- | --- | --- | --- |
| **D. Exemplary text documents for each social media type** | | | |
| **ID** | **Language** | **Text (English)** | **Text (original)** |
| AMT | DE | I can't recommend it to anyone | Kann ich niemandem empfehlen |
|  | EN | Save your money and time! |  |
| AMR | DE | I'm totally disappointed in this novel. Despite the 521 pages he seems to me cobbled together quickly and mostly completely unbelievable. What happened to Lou's plans, she wanted to study fashion design? Instead, she works in a London bar with a silly costume. And falls off the roof. Of course, the paramedic is a smart single. Then someone from Will's former life appears, her mother reinvents herself, and Sister Tina's high-flyer career is also looked for in vain. This stringing together of clichés is hair-raising. At least it's well written, that's all the positive I can say. | Ich bin von diesem Roman total enttäuscht. Auf mich wirkt er trotz der 521 Seiten schnell zusammengeschustert und überwiegend völlig unglaubwürdig. Was ist aus Lou's Plänen geworden, sie wollte doch Modedesign studieren? Anstatt dessen arbeitet sie in einer Londoner Bar mit einer albernen Kostümierung. Und fällt vom Dach. Selbstverständlich ist der Rettungssanitäter ein smarter Single. Dann erscheint jemand aus Wills früherem Leben, ihre Mutter erfindet sich neu, und Schwester Tinas Überflieger-Karriere sucht man ebenfalls vergeblich. Diese Aneinanderreihung von Klischees ist haarsträubend. Wenigstens ist es gut geschrieben, das ist alles an Positivem, was ich sagen kann. |
|  | EN | I do like this product but it wasn't what I was expecting. That's not to say the description is misleading, but I think I had it in my head that this was just a blanket with a waterproof backing. It's not! It's all made of a nylon/plasticky material. The top is silky but not fabric. It wipes off easily and it's not hard to fold up like some foldable blankets I've used. The carrying pouch is handy too. We keep it in the car for times when we need a little stretch on the grass. We have the Cocoa Bubble pattern, which is quite cute. While I do like this, I wouldn't use it if we wanted a soft place to snuggle up for a long time, like watching an outdoor movie or something. For that, I would just bring a softer regular blanket to lay on top. But if you're looking for a tough, waterproof ground covering you can easily tote around, this is perfect. |  |
| IMD | EN | I could not stand to even watch it for very long for fear of losing I.Q. |  |
| YEL | EN | Today is the second time I've been to their lunch buffet and it was pretty good. |  |

|  |  |  |  |
| --- | --- | --- | --- |
| FBK | DE | With AIDA, every place on earth is something special. But with us, the ship is also the target. Nevertheless, we are pleased that so many other goals are possible with AIDA Selection. | Mit AIDA ist jeder Ort dieser Erde etwas besonderes. Aber bei uns ist ja auch das Schiff das Ziel. Dennoch freut es uns das so viele andere Ziele mit AIDA Selection möglich sind. |
| CBC | DE | "If Opel wants to stay afloat, then it should do everything in its power to appeal to a wide range of buyers. My opinion..." 100% approval. In my opinion, the "success" of some competitors lies precisely in this. The greatest possible (!) avoidance of non-combinable equipment variants would be a start. | „Wenn Opel sich über Wasser halten will, dann sollte man auch alles daran setzen, eine breite Käuferschicht anzusprechen. Meine Meinung…” 100 % Zustimmung. Der “Erfolg” mancher Konkurrenz liegt meiner Meinung nach genau darin begründet. Die größtmögliche (!) Vermeidung von nicht kombinierbaren Ausstattungsvarianten wäre ein Anfang. |
| TWS | EN | Fried pickles fried mac and cheese fried chicken.. reality shows and #beer.. yes it is that kind of… |  |
|  | ES | Evening with the people, who bring me sweet gifts from their holiday #SuperToblerone | Tarde-noche con la gente,que me traen regalos dulces de sus vacaciones #SuperToblerone |
|  | DE | Enjoying #PokemonGO with friends and a #heineken in #Meiningen enjoy | #PokemonGO mit Freunden und einen #heineken in #Meiningen genießen |
| YTU | EN | Rihanna and Eminem together are unstoppable.﻿ |  |
| SMS | EN | Hello! How's you and how did saturday go? I was just texting to see if you'd decided to do anything tomo. Not that i'm trying to invite myself or anything! |  |
| ROT | EN | when the film reaches its dramatic climax , a varied cast of characters must all figure out a way to bridge the chasm between their dreams and reality . |  |
| CBP | EN | The LG Style Icon contest tour is almost over â€“ just two more chances to make your impression! This weekend we’re heading to St David’s Centre for the ninth & penultimate of the roadshow events. LG will be hosting an open-call photo-shoot at the shopping centre from 10am on Saturday 9th and Sunday 10th August. Any budding Irish models should check your schedule for next weekend when the roadshow hits Dublin, don’t miss it! For more information about the competition, check out the official LG StyleIcon site. |  |
| TWC | EN | This place never disappoints #NoEraPenal @ Lomeli's Italian Restaurant |  |

**E. Preprocessing and method specifications**

As part of our preprocessing, we transform all words to lower case, so identical words in different capitalizations are counted only once, and remove all numbers and punctuation. Based on this, we generate n-grams, i.e., frequently occurring term combinations, with a maximum length of two words. Such bigrams are typically used because they can capture more local context than unigrams alone (Boiy et al. 2007). Also following prior research and to reduce dimensionality and consequently computational complexity, we omit all words shorter than three characters as well as all unigrams and bigrams occurring less than four and seven times, respectively (e.g., Pang et al. 2002). According to research conventions, we retain stop-words (e.g., Pang et al. 2002; Bermingham & Smeaton 2010) and do not conduct stemming (e.g., Annett & Kondrak 2008; Durant & Smith 2006). Dave et al. (2003) argue that stemming may lead to overregularization, potentially removing information relevant for the classification task. Similarly, Bermingham and Smeaton (2010) find no benefit from stop-word removal and stemming and even suggest that “noisy artefacts” of the microblog domain may convey information that helps classifiers distinguish between classes. Several researchers report decreased performance for part-of-speech (POS) tags (e.g., Go et al. 2009; Kouloumpis et al. 2011; Pang et al. 2002). Hence, we do not include them in this research.

After preprocessing, the document-term matrix (DTM) is generated, where each term and term pair above the minimum frequency thresholds are used as features of the DTM, disregarding word order. This bag-of-words (BOW) approach is widely used for text classification problems (Aggarwal & Zhai 2012). The columns of the DTM represent the terms, and the rows represent the text documents. Following research convention (e.g., Go et al. 2009; Xia et al. 2011; Durant & Smith 2006), we use binary presence weighting to represent each document as a word vector. Alternative feature weighting approaches such as term frequency-inverse document frequency (tf-idf) and frequency weighting exist. However, presence weighting has been shown to perform well for classification tasks similar to ours (e.g., Pang et al. 2002) and is also more intuitive to interpret.We implement all five machine learning methods in R, using the established machine learning package *caret* (Kuhn 2008). For preprocessing, we use the popular text mining framework of the *tm* package (Feinerer 2017).

**R package details for method implementation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Name (caret)** | **R package** | **Author** | **CRAN URL** |
| **ANN** | mlpWeight DecayML | RSNNS | Bergmeir (2017) | https://cran.r-project.org/web/packages/RSNNS/RSNNS.pdf |
| **kNN** | knn | class | Ripley (2015) | https://cran.r-project.org/web/packages/class/class.pdf |
| **NB** | nb | klaR | Ligges (2018) | https://cran.r-project.org/web/packages/klaR/klaR.pdf |
| **RF** | ranger | ranger | Wright (2018) | https://cran.r-project.org/web/packages/ranger/ranger.pdf |
| **SVM** | svmLinear | kernlab | Karatzoglou (2016) | https://cran.r-project.org/web/packages/kernlab/kernlab.pdf |

z

The large amount of potential combinations of methods, datasets, classification tasks, and parameter values makes exhaustive testing of all conceivable parameter values for each method challenging (Yang 1999). Hence, we follow method comparisons in computer science and optimize the single most important parameter per method, e.g., *k* in the range of 1 to 65 for kNN, the same parameter grid Joachims (1998) applies. As a distance measure for kNN, we apply Euclidean distance. For SVM, we use a linear kernel (e.g., Boiy et al. 2007) and tune the cost parameter *C* by factors of ten from 10-2 to 102 (see Moraes et al. 2013 for a similar parameter grid). For RF, we grow 500 uncorrelated decision trees and test two specifications for the number of randomly selected features (e.g., Caruana et al. 2008; Wright & Ziegler 2015). For NB, we test models both with and without Laplace smoothing, avoiding non-zero frequencies for previously unseen terms (Dave et al. 2003). For all parameters not otherwise specified, we adhere to their default values.

**Parameter grids**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Parameter | # specifications tested | Parameter grid | Source using similar specifications |
| ANN | # neurons in hidden layer | 6 | {20,…,25} | Caruana et al. (2008) |
| kNN | # nearest neighbors (*k*) | 5 | {1,15,30,45,65} | Joachims (1998) |
| NB | Laplace smoothing | 2 | {0,1} | Pang et al. (2002) |
| RF | # random features | 2 | {s/2, s}, where s equals the square root of the total number of features | Breiman (2001) |
| SVM | Cost parameter (*C*) | 5 | {10-2,…,102} | Moraes et al. (2013) |

**Model specifications for each dataset**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | ANN | kNN | NB | RF | SVM |
| ID | **Language** | **# sample sizes** | # neurons in hidden layer | # nearest neighbors (*k*) | Laplace smoothing | # random features | Cost parameter (*C*) |
| AMT | DE | 3 | 1/1/1 | 1/1/1 | 0/0/1 | 5/6/22 | .01/.1/1 |
|  | EN | 3 | 8/4/2 | 1/1/1 | 1/1/1 | 5/15/13 | 10/.1/1 |
| AMR | DE | 3 | 16/8/8 | 1/15/1 | 0/0/0 | 20/56/92 | .1/.1/.1 |
|  | EN | 3 | 4/1/32 | 15/15/30 | 0/0/0 | 43/58/47 | .01/.01/.01 |
| IMD | EN | 2 | 4/1 | 1/1 | 0/0 | 17/12 | 1/1 |
| YEL | EN | 2 | 8/8 | 1/1 | 1/0 | 16/11 | .1/.1 |
| FBK | DE | 3 | 32/2/16 | 1/2/1 | 0/0/0 | 17/23/39 | 1/1/1 |
| CBC | DE | 3 | 2/8/16 | 15/65/65 | 0/0/1 | 12/18/62 | .01/.1/.1 |
| TWS | EN | 3 | 2/1/4 | 1/1/1 | 0/1/1 | 13/9/17 | .01/.1/1 |
|  | ES | 2 | 1/4 | 15/65 | 1/1 | 6/9 | .01/1 |
|  | DE | 2 | 2/32 | 65/65 | 0/1 | 13/9 | .01/1 |
| YTU | EN | 2 | 4/4 | 1/1 | 0/0 | 17/25 | .01/.1 |
| SMS | EN | 2 | 1/1 | 1/1 | 0/0 | 10/15 | .1/.1 |
| ROT | EN | 3 | 1/16/8 | 65/65/1 | 1/0/0 | 10/15/26 | .01/.01/.1 |
| CBP | EN | 2 | 8/8 | 65/65 | 0/0 | 75/102 | 1/.01 |
| TWC | EN | 3 | 1/32/8 | 1/65/1 | 1/1/1 | 7/19/35 | 1/1/1 |

Note: Each cell contains parameter values for different sample sizes: N=500/1,000/3,000.

**F. R script for text classification**

To support further research interested in applying machine learning methods for text classification, we provide all technical details and our R script. As an intuitive example, researchers and practitioners can run this script on the publicly available “Sentiment Labelled Sentences Data Set” from the UCI repository (Kotziats et al. 2015), which contains three separate files.[[1]](#footnote-1) The file “yelp\_labelled.txt” corresponds to YEL, consisting of 1,000 observations (see Table 1). Thus, following the steps below, our results can be replicated.

The script we provide is fully annotated, mirroring the overall structure of the manuscript. After loading all required packages, e.g., *caret* (Kuhn 2008), and initializing parallel processing, the raw text data is loaded, preprocessed, and transformed into a document-term-matrix (DTM) with unigrams and bigrams as features using the function “pre\_process\_data” (see Web Appendix B). The labeled DTM is partitioned into a training and hold-out test set, using an 80/20 split, allowing an unbiased parameter tuning and performance assessment. We set a seed before splitting to make sure replications arrive at identical results. Following this step, we define the “trainControl” functions for cross-validation and define the parameter grids for each method (see chapter 4.3 and Web Appendix E).

After concluding the preprocessing and method specifications, training is initialized. Again, we set a seed to ensure replicability. The training function is implemented as a loop over all five methods, which are specified in the vector “set\_of\_models”. The tuned parameters for each method are consolidated in the dataframe “parameters”. Using another loop over all methods with tuned parameters, we let them predict on the hold-out test set, storing predictions in the dataframe “predictions”. Lastly, we consolidate all results in the dataframe “results” and plot the accuracies per method.

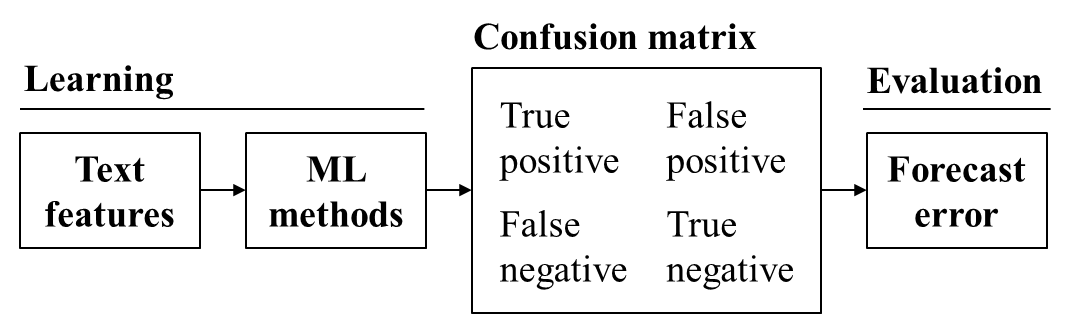
**G. Correlations based on classification accuracy**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | kNN | NB | RF | SVM |
| ANN | .35 | .56 | .64 | .61 |
| kNN |  | .40 | .47 | .40 |
| NB |  |  | .61 | .51 |
| RF |  |  |  | .55 |

**H. Details on cost computation of application scenarios**

For application scenario 1, the misclassification costs equal the accumulated efficiency gains forgone through suboptimal method choice. In contrast, application scenarios 2 and 3 are slightly more complex. They quantify the misclassification costs as the difference in forecast errors between different machine learning methods. The objective of the latter two scenarios is to predict demand based on textual online communication. As illustrated below, we apply the same two initial steps for both datasets, i.e., train each method to assign a set of text features (document-term matrix) to demand and website visit classes. Since the primary objective of both applications is prediction rather than causal inference, endogeneity and omitted variable bias are of less concern in particular when reasonable levels of accuracy can be achieved.

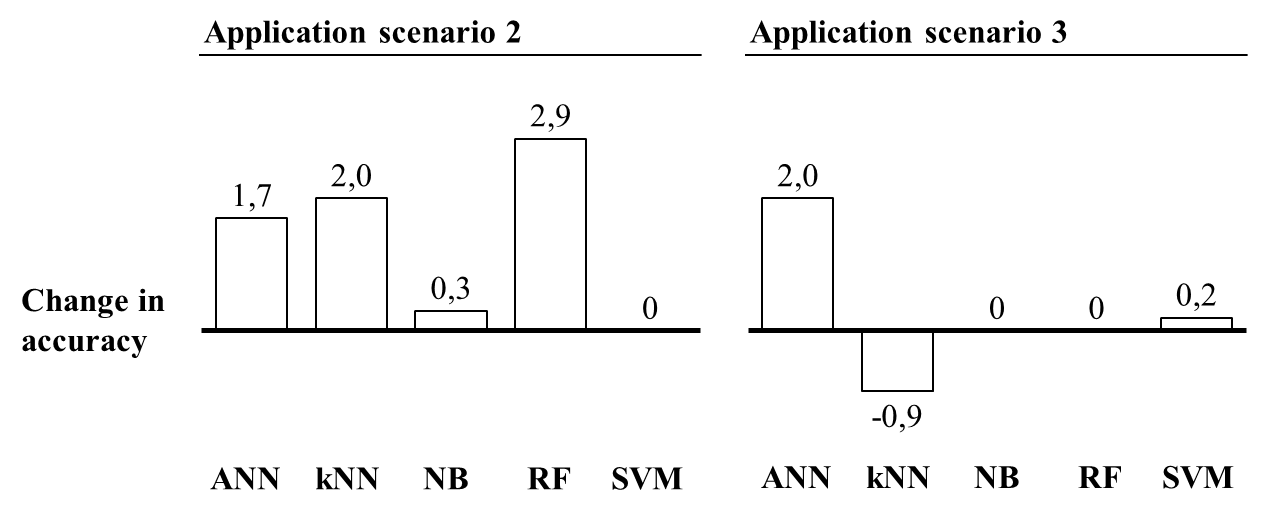
Comparing the predictions of all methods against the true outcomes produces a 2x2 confusion matrix. The sum of false positives and false negatives represent the cost of wrong predictions. We assume the forecast error to be symmetric along the diagonal of the confusion matrix, i.e., the costs of over- and underestimating demand to be identical.

**Process from text features to forecast error**

Note the values in Fig. 3 are based on text-only data. Forecasting accuracy can be further improved adding further information. Since this further increases dimensionality while keeping number of observations constant, methods like ANN or RF are likely to benefit more than methods like kNN. This would further increase the difference in performance and associated economic consequences. The estimated differences reported in the paper are therefore likely to be conservative. To illustrate this, we run additional models for both scenarios adding relevant factors available to us. Specifically, for scenario 2, we add the number of song uploads, network centrality, and previous song downloads in the models. For scenario 3, we control for the time since blog post as well as age, experience, and gender of the corporate blog post author.

The bar chart below summarizes the percentage changes of forecasting accuracies, including the additional variables compared to the baseline models limited to text features. As can be seen, adding the controls improves the accuracies only marginally with absolute percentage changes between 0 and 2.9 percentage points, presumably because the text-only analysis already results in good levels of forecasting accuracy. In both cases the difference between the RF performance (best performing method) and kNN (worst performing method) further increases, which would result in larger economic effects of inferior method choice. However, as the effect is quite small and the implications remain the same, we report the results of the base model in the manuscript.

**Changes in accuracies (in % points) when controlling for additional variables**



**Additional References**

Borah, A., & Tellis, G. J. (2016). Halo (spillover) effects in social media: Do product recalls of one brand hurt or help rival brands?. *Journal of Marketing Research*, *53*(2), 143-160.

Bergmeier, C. (2017). *Package ‘RSNNS’*. 1-74.

Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. *Proceedings of the 25th International Conference on Machine learning*, 99-103.

Chan, C., & Mogilner, C. (2017). Experiential gifts foster stronger social relationships than material gifts. *Journal of Consumer Research*, *43*(6), 913-931.

Chen, Z. (2017). Social acceptance and word of mouth: How the motive to belong leads to divergent WOM with strangers and friends. *Journal of Consumer Research*, *44*(3), 613-632.

Colicev, A., Malshe, A., Pauwels, K., & O’Connor, P. (2018). Improving consumer mindset metrics and shareholder value through social media: The different roles of owned and earned media. *Journal of Marketing*, *82*(1), 37-56.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *In Proceedings of the 12th International Conference on World Wide Web*, 519-528.

Do, C. B., & Ng, A. Y. (2006). Transfer learning for text classification. *In Advances in Neural Information Processing Systems,* 299-306.

Durant, K. T., & Smith, M. D. (2006). Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. *International Workshop on Knowledge Discovery on the Web*, Springer, Berlin, Heidelberg, 187-206.

Feinerer, I., Hornik, K., & Feinerer, M. I. (2017). *Package ‘tm’*. 1-64.

Gautam, G., & Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. *7th Int. Conf. on* *Contemporary Computing,* 437-442.

Guillaumin, M., & Ferrari, V. (2012). Large-scale knowledge transfer for object localization in ImageNet. *In* *Computer Vision and Pattern Recognition* *(CVPR),* 3202-3209.

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* *forthcoming*, 1-51.

Heimbach, I., & Hinz, O. (2016). The impact of content sentiment and emotionality on content virality. *International Journal of Research in Marketing*, *33*(3), 695-701.

Ilhan, B. E., Kübler, R. V., & Pauwels, K. H. (2018). Battle of the brand fans: Impact of brand attack and defense on social media. *Journal of Interactive Marketing*, *43*, 33-51.

Kanuri, V. K., Chen, Y., & Sridhar, S. (2018). Scheduling content on social media: Theory, evidence and application. *Journal of Marketing forthcoming,* 1-64.

Karatzoglou, A. (2016). *Package* *‘kernlab’.* 1-108.

Klostermann, J., Plumeyer, A., Böger, D., & Decker, R. (2018). Extracting Brand Information from Social Networks. Integrating Image, Text, and Social Tagging Data. *International Journal of Research in Marketing*, 35(4).

Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg!. *In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 11(164), 538-541.

Kuhn, M. (2008). Caret package. *Journal of Statistical Software*, 28(5), 1-26.

Kupfer, A. K., Pähler vor der Holte, N., Kübler, R. V., & Hennig-Thurau, T. (2018). The role of the partner brand’s social media power in brand alliances. *Journal of Marketing*, *82*(3), 25-44.

Ligges, U. (2018). *Package* *‘klaR’.* 1-81.

Ludwig, S., de Ruyter, K., Friedman M., Brüggen, E.C., Wetzels, M., Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing,* 77(1), 87-103.

Marchand, A., Hennig-Thurau, T., & Wiertz, C. (2017). Not all digital word of mouth is created equal: Understanding the respective impact of consumer reviews and microblogs on new product success. *International Journal of Research in Marketing*, *34*(2), 336-354.

Packard, G., Moore, S. G., & McFerran, B. (2018). (I'm) happy to help (you): The impact of personal pronoun use in customer-firm interactions. *Journal of Marketing Research forthcoming,* 1-78.

Ripley, B. (2015). *Package ‘class‘.* 1-19.

Rooderkerk, R. P., & Pauwels, K. H. (2016). No comment?! The drivers of reactions to online posts in professional groups. *Journal of Interactive Marketing*, 35, 1-15.

Rutz, O. J., Trusov, M., & Bucklin, R. E. (2011). Modeling indirect effects of paid search advertising: Which keywords lead to more future visits?. *Marketing Science*, *30*(4), 646-665.

Tang, T., Fang, E., & Wang, F. (2014). Is neutral really neutral? The effects of neutral user-generated content on product sales. *Journal of Marketing*, *78*(4), 41-58.

Timoshenko, A., & Hauser, J. R. (2018). Identifying customer needs from user-generated content. (*Working Paper).*

van Laer, T., Escalas, J. E., Ludwig, S., van den Hende, E. A., & Herr, P. M. (2018). What happens in Vegas stays on TripAdvisor? A theory and technique to understand narrativity in consumer reviews. *Journal of Consumer Research (forthcoming),* 1-64.

Wright, M. N., & Ziegler, A. (2015). *ranger: A fast implementation of random forests for high dimensional data in C++ and R*. 1-17.

Wright, M.N. (2018). *Package ‘ranger’*. 1-23.

Yadav, M. S., Prabhu, J. C., & Chandy, R. K. (2007). Managing the future: CEO attention and innovation outcomes. *Journal of Marketing*, *71*(4), 84-101.

1. https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences [↑](#footnote-ref-1)