Wedel, Irina; Palk, Michael; Voß, Stefan

A Bilingual Comparison of Sentiment and Topics for a Product Event on Twitter

Information Systems Frontiers

Provided in Cooperation with:
Springer Nature

Suggested Citation: Wedel, Irina; Palk, Michael; Voß, Stefan (2021) : A Bilingual Comparison of Sentiment and Topics for a Product Event on Twitter, Information Systems Frontiers, ISSN 1572-9419, Springer US, New York, NY, Vol. 24, Iss. 5, pp. 1635-1646, https://doi.org/10.1007/s10796-021-10169-x

Information Systems Frontiers (2022) 24:1635–1646
https://doi.org/10.1007/s10796-021-10169-x

A Bilingual Comparison of Sentiment and Topics for a Product Event on Twitter

Irina Wedel[1] · **Michael Palk**[2] · **Stefan Voß**[2]

Accepted: 25 June 2021 / Published online: 31 July 2021

Abstract

Social media enable companies to assess consumers' opinions, complaints and needs. The systematic and data-driven analysis of social media to generate business value is summarized under the term Social Media Analytics which includes statistical, network-based and language-based approaches. We focus on textual data and investigate which conversation topics arise during the time of a new product introduction on Twitter and how the overall sentiment is during and after the event. The analysis via Natural Language Processing tools is conducted in two languages and four different countries, such that cultural differences in the tonality and customer needs can be identified for the product. Different methods of sentiment analysis and topic modeling are compared to identify the usability in social media and in the respective languages English and German. Furthermore, we illustrate the importance of preprocessing steps when applying these methods and identify relevant product insights.

Keywords Social media analytics · Natural language processing · Topic modeling · Sentiment analysis

1 Introduction

Around half of the world population uses social media nowadays, where the about 3.8 billion users satisfy their needs for communication, information and entertainment (Kemp, 2020). It can be seen as a source of information about opinions, interests, activities and interactions among the users. Especially companies recognized the potential of social media as a tool for market research, competitive analysis, customer relationship management or product research and development. The field Social Media Analytics (SMA) is an interdisciplinary research field where computational methods are important tools to handle the massive amount of data and transform user posts into helpful insights, which is used across various areas like public administration, politics, the industrial and financial sector, retail, education, healthcare (Rathore et al., 2017), supply chain intelligence (Swain & Cao, 2019), tourism (Chang & Chen, 2019) or road traffic (Vallejos et al., 2021). Previous research focused mainly on user behavior (Awal & Bharadwaj, 2019), the value and risk associated with social media, comparison with traditional media, social media as marketing tool or as communication platform during catastrophes (Kapoor et al., 2018). Stieglitz et al. (2018) contribute to this field with a SMA framework which includes the whole process from data collection, preparation and up to analysis. They also identify common approaches and methods in SMA like statistical, social network, sentiment, content and trend analysis. Especially the methods sentiment, content and trend analysis to capture opinions, complaints or needs of consumers require textual data as input, where Twitter can be seen as a leading social media platform for microblogging and thus suitable as a data source. An advantage compared to classical market research methods like interviews is that social media can be described as a laboratory for observations where behaviors or opinions can be assessed in an unbiased way (Stieglitz et al., 2014). A use case for companies is to monitor user reactions during an event, where a new product is introduced to the public, called product

✉ Michael Palk
michael.palk@uni-hamburg.de

Irina Wedel
irina.wedel@icloud.com

Stefan Voß
stefan.voss@uni-hamburg.de

[1] University of Hamburg, Hamburg, Germany

[2] Institute of Information Systems, University of Hamburg, Hamburg, Germany

event. Opinions about technical specification or design of the product can be assessed in real time via the Twitter API during the event, allowing to capture direct reactions of the consumers. Approaches from Natural Language Processing (NLP) can be applied for this task like sentiment analysis or topic modeling, which have to be adapted in the context of social media, where abbreviations, emoticons and various slangs are common.

In this paper, we analyze reactions and sentiment on Twitter for a product event from Apple in October 2020, where new models of the iPhone 12 were presented. Various approaches and methods of NLP are tested and compared. In addition, we analyze how the sentiment and conversation topics differ across two languages and four countries. The remainder of this paper is structured as follows.

In Section 2, approaches and previous works of sentiment analysis and topic modeling in the context of social media are reviewed from the literature. Afterwards, we present the design of the computational study, where various models are compared on data retrieved during a product event on Twitter in Section 3. The results across different countries and languages and derived insights for the products are presented in Section 4. This paper ends with a conclusion and future outlook in Section 5.

2 Natural Language Processing in Social Media Analytics

In recent years, the area of NLP improved through better data storage and process capabilities, increased computing speed and advanced algorithms. With millions of posts every day, social media delivers a large amount of data which can be analyzed with these tools. Especially Twitter, where the focus of this microblogging service is on short text messages, called Tweets, seems to be a suitable platform for applying various approaches from NLP. We refer to Giachanou & Crestani (2016) for further terminology of Twitter which is used throughout this paper. Furthermore, Twitter offers an API for scraping Tweets which include a hashtag, i.e. a kind of identifier for classifying a Tweet to a specific topic. Another possibility is to search for Tweets which include mentions or specific key words. In contrast to Facebook or Instagram,[1] the API of Twitter does not suffer from several restrictions due to various data scandals, such that practitioners and researchers still have potential access to uncensored data. This could be the reason, that Twitter is the most used platform in SMA research (Rathore et al., 2017). Two important topics in the field of NLP in SMA are sentiment analysis and topic modeling, which are introduced in the next subsections.

2.1 Sentiment Analysis

Sentiment analysis refers to extracting opinions and/or emotions from documents or just simple pieces of text. While sentiment analysis can be used to investigate diverse media such as audio, video or images, we are mainly focused on extracting sentiments from a collection of texts, or more specific Tweets. This could be important for assessing the general sentiment or polarity towards a certain product or service to estimate product preferences of consumers. Examples for general surveys about sentiment analysis and opinion mining are Pang & Lee (2008) and Liu (2012), a more recent one especially for Twitter is Giachanou and Crestani (2016).

Four approaches in this area are commonly applied, a lexicon-based and a learning-based approach, a hybrid of the two previous ones and a graph-based approach (see Giachanou & Crestani (2016) for an overview). The idea of the lexicon-based approach is to compare the words of a text with a semantic lexicon, which is a collection of unique words with their corresponding polarities. The challenge with this approach is that the polarity of words can differ in dependence of the context, for example, 'long' can be associated positive in context of battery life but negative in context of waiting time in a queue. Early work to avoid such problems can be found in Hatzivassiloglou & McKeown (1997), where it is shown that conjunctions like 'but' can help to determine if an adjective should be regarded as positive or negative. A similar approach from (Popescu & Etzioni, 2005) is to include the corresponding noun of an adjective to determine the polarity, other examples of such rules can be found in Ding et al. (2008), Fahrni & Klenner (2008), Qiu et al. (2009) or Cruz et al. (2013). Usually, when applying sentiment analysis or other NLP algorithms, researchers use existing lexicons (for example, Esuli & Sebastiani (2006) or Tausczik & Pennebaker (2010) to avoid building a lexicon from scratch.

However, previous lexicons are not really suitable in the context of social media since users tend to include abbreviations, acronyms, emoticons or slangs in their posts. This is especially the case on Twitter since a posting can have a maximum of 280 characters which motivates even more to use a more informal speech; other challenges of Twitter sentiment analysis are stated in Giachanou & Crestani (2016). Therefore, Hutto & Gilbert (2014) developed a lexicon specific for social media usage for the English language, a German equivalent was adopted by (Tymann et al., 2019).

Another approach is the learning-based approach where an algorithm learns the polarity in a supervised manner, i.e.

[1] See, for example, the changelog in the case of Instagram's API: https://www.instagram.com/developer/changelog/

on a training corpus with labeled sentiment for every text in this corpus. An advantage compared to the lexicon-based approach is that suitable machine learning algorithms can learn the sentiment independently from the domain.

Previous work includes, for example, Cheong & Lee (2011) who track sentiment from civilians towards terrorism events and more general (Zhang et al., 2011), where a lexicon- and learning-based method is combined to determine sentiment for various themes on Twitter. Agarwal et al. (2011) and Kouloumpis et al. (2011) focus on feature representation of Tweets to detect the sentiment, while Chamlertwat et al. (2012) include lexicon- and learning-based approaches in their solution for detecting sentiment related to products. Saif et al. (2012) include semantics as additional feature for the prediction, while Severyn and Moschitti (2015) utilize deep convolutional neural networks.

2.2 Topic Modeling

In contrast to sentiment analysis, topic modeling is an unsupervised technique to identify topics and does not require a lexicon as input. It is assumed that every document contains several topics which are defined as a distribution of corresponding words. Furthermore, every document can be seen as a probability distribution of several topics. One approach for this task is the Latent Dirichlet Allocation (LDA) by Blei et al. (2003), where the words in the documents are imagined as observable objects while the probability distribution of topics in a document and the relation of words to these topics are hidden, i.e. latent structures. The Dirichlet distribution is a family of continuous multivariate distributions where in the case of topic modeling the topic distribution θ_i for a document $i \in \{1, \ldots, M\}$ and the word distribution φ_k for a topic $k \in \{1, \ldots, K\}$ are drawn from a Dirichlet distribution $Dir(\alpha)$ and $Dir(\beta)$, respectively. Thus, each of the M documents has its own multinomial distribution of the K topics and each topic k its own multinomial distribution of the words from a corpus. It is assumed that documents are created by a generative process where for each word position i, j in a document i and $j \in \{1, \ldots, N_i\}$, where N_i is the number of words in document i, first a topic $z_{i,j}$ is drawn from $Multinomial(\theta_i)$ and from this topic a word $w_{i,j}$ is chosen from $Multinomial(z_{i,j})$. Under such assumption, the aim is to identify topics which have most likely generated the documents under such generative process. This problem of statistical inference can be solved, for example, with Monto Carlo Simulation or with Likelihood Maximization.

Another approach is the Latent Semantic Analysis (LSA) by Deerwester et al. (1990), where a Singular Value Decomposition (SVD) is applied to a document-term matrix X, where an entry x_{ij} indicates that term or word j occurs in document i. If SVD is applied on such a matrix, two orthogonal matrices T, D and a diagonal matrix S can be retrieved. The matrix T contains information about the relation of words to documents, D about the distribution of topics across documents and S about the relevance of topics. Via dimensionality reduction techniques, it is possible to identify latent topics and allow to display the similarity of words and documents in a semantic space. A similar approach is the Non-negative Matrix Factorization (NMF) which also uses a document-term matrix as input but in contrast to the SVD, two non-negative matrices W and H will be produced as output of the factorization. The values of these matrices can also be displayed in a semantic space to identify word-document similarity.

Previous work where these approaches were applied to Twitter data is, for example, Zhao et al. (2011), where LDA is used to compare dominant topics in Twitter and traditional media outlets like the New York Times. Also Lau et al. (2012) utilize LDA to detect trending topics in Twitter, while Ostrowski (2015) identify product-related topics for an automotive and Kostygina et al. (2016) make use of LDA to identify content about cigars on Twitter. Xie et al. (2016) identify the need for a real-time detection for topics which are prevalent for just a short time or for news which is spread on Twitter faster than it is reported about in the mainstream media. Their solution approach is a modified version of LSA which also includes SVD and dimensionality reduction techniques. Other examples are Park et al. (2016) for an analysis of tourism marketing related content on Twitter, Haghighi et al. (2018) for an application in public transport, Curiskis et al. (2020) for an evaluation of different NLP techniques for the platforms Twitter and Reddit and Ustek-Spilda et al. (2021), who investigated topics from Twitter in the context of Internet of Things,

Most of the previous literature focus, for sentiment analysis and topic modeling, just on one language and on one specific approach. Our contribution is to compare several methodologies for two languages across four countries, while also highlighting the relevance of preprocessing and feature extraction for the efficiency of the methods.

3 Topic Modeling and Sentiment Analysis of a Product Event on Twitter

In this section, we present the data collection, preprocessing, model building and considered evaluation measures for the computational study to compare the introduced methods on a data set of Tweets about a product event of Apple. This event took place on the 13th of October in 2020, where new models of the iPhone 12 were introduced. Since the CEO of Apple, Tim Cook, was presenting these new smartphones and due to the general popularity of Apple's products, there is a typical high resonance for such events, probably

resulting in a high engagement of users on Twitter. The goal is to assess the overall sentiment towards the new product across countries and also identify relevant product attributes which are perceived usable or bothering.

3.1 Data Collection

We utilize the Search- and Streaming-API of Twitter to collect Tweets about the product event. The Search-API gives a representative sample of Tweets in the last seven days, while the Streaming-API gives all Tweets in real-time with pre-defined filtering criteria. To capture cultural differences, not only German Tweets from Germany and English Tweets are collected, but we also differentiate English Tweets by country, namely the United States (US), the United Kingdom (UK) and Australia. Since adaption of the iPhone in these countries is quite high, we expect still variations in the sentiment and topics due to cultural differences. Another aspect is that we restrict the data collection to a representative sample for the Englisch speaking countries due to data retrieving and storage issues, while due to the the smaller popularity of Twitter in Germany it is possible to access the data via the Streaming-API. Hence, we collect real-time data for German Tweets and augment this data set with a sample from the Search-API. The data collection is executed with the Python package Tweepy which is able to handle all necessary steps like authentication, connection to the API, starting or ending a session. For a detailed overview of the technical specifications, we refer to the official documentation of Twitter's API.[2]

The real-time data from the Streaming-API with the key word 'iPhone' was collected around the time window of the iPhone presentation during the Apple event, where the streaming session started at 17:13 (according to the Coordinated Universal Time) until 19:00, such that an additional buffer time was included after the official ending of the iPhone presentation at 18:10. A total of 2422 German Tweets could be retrieved, which was augmented by Tweets from the Search-API. Here, a sample of Tweets from the last seven days can be collected which is useful to include reactions from users which did not follow the event live but at a later point in time. The search parameters were specified in a way that we include a sample with the search term 'iPhone OR iphone OR #iphone' from 13 to 15 October since it is plausible to assume that the interest about the event declines rapidly after a longer time frame. In total, 7735 German Tweets, 7515 from the UK, 55040 from the US and 1771 from Australia were collected.

[2] https://developer.twitter.com/en/docs

3.2 Data Preprocessing

After the data collection, the Tweets are preprocessed within a Python environment. First, elements like mentions of other users, URLs, punctuation marks, empty paragraphs, hashtags and emoticons are removed, as such symbols can not be processed by existing algorithms for topic modeling and sentiment analysis. Additionally, all characters are set to lowercase and the resulting strings of words are tokenized. These steps can be done by basic Python operations, while we use the library SpaCy for stop word elimination which includes a list of 543 German stop words and a list of 326 English stop words. Again with the use of the SpaCy library and also the NLTK (Natural Language Toolkit) package, we apply stemming and lemmatization to the Tweets which results in an additional reduction of the text, since variants of words are reduced to a common word stem. These preprocessing steps are necessary to create a basic data set on which all considered methods can be applied, as for example hashtags often contain abbreviations where domain knowledge is needed for deciphering and emoticons can not be interpreted by existing lexicons. Another advantage of such preprocessing is that the computational effort through reduced data sets can be lowered. The next step is to extract numerical features from the texts. Therefore, the Bag-of-Words (BOW) and Term Frequency - Inverse Document Frequency (TF-IDF) approach are applied. In the case of the BOW approach, a document-term matrix D is created where the rows contain the Tweets and the columns all words from the vocabulary which contains every word of every collected Tweet. Thus, an entry d_{ij} displays the number of times a word j from the vocabulary exists in Tweet i. A problem with this approach is that words which are common in many documents, do not add much explanatory value. The ideas based on Spärck Jones (1972) resulted in the commonly used TF-IDF metric. Here, terms which are used often but just in few documents get a higher weight. First, the inverse document frequency $idf(t, d)$ of term t in the documents can be calculated with

$$idf(t, d) = log \frac{N}{df(d, t)}$$

where N is the number of documents and $df(d, t)$ the frequency of documents d including term t. Then the TF-IDF metric is expressed as follows:

$$tfidf(t, d) = tf(t, d) * idf(t, d)$$

where $tf(t, d)$ is the term frequency of term t in the documents.

For the learning-based sentiment analysis approaches, it is necessary to manually label the data. Therefore, all 7735

German Tweets are labeled according to their sentiment in 'positive', 'neutral' or 'negative', as well as all 7515 Tweets from the UK to have a representative training and test set for the English language. Due to a prohibitively large number of Tweets from the US and Australia, we decided to choose Germany and the UK for labeling as nearly the similar amount of Tweets are coming from these countries. The trained machine learning models for English Tweets are used to detect sentiment for the US and Australia. 730 German Tweets are positive, 525 negative and 4552 neutral, whereby 1929 from these Tweets are deleted since they can be identified as non product-related spam. To avoid problems due to imbalanced classes, we reduce the number of neutral German Tweets to 700 by deleting randomly chosen ones. A similar approach is done for the English Tweets from the UK, where 717 Tweets are positive, 632 negative and 5748 neutral, thus we reduce the number of neutral Tweets again to 700.

3.3 Model Implementation and Evaluation

For the topic modeling, we utilize the Python library Gensim which includes algorithms for LDA, LSA and NMF. To evaluate such unsupervised tasks, the coherence value is a metric which measures the relative distance of words inside a topic. The idea is that words with a similar meaning tend to appear in a similar context, such that a topic gets a high coherence value if the most common words are strongly related to each other in a semantic sense. Another measure for topic modeling is the optimal number of topics to maximize the coherence value (called K-value). While there are several coherence metrics, we use the CV-coherence from Röder et al. (2015), which is declared as a reliable measure in their comparisons. To compare various methods from the lexicon- and learning-based sentiment analysis approaches, we consider for the lexicon-based approach four open accessible lexicons, namely SentiWordNet and VADER for the English language and SentiWS and GerVADER for German. For the learning-based approach, the methods Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Nearest-Neighbors (NN), Decision Tree (DT) and Random Forest (RF) are included into the computational study. For training and evaluating these models, we use 80% of the data as training set and 20% for testing, which are divided randomly while ensuring that the relation of classes inside the training and test set is balanced. As metric for the classification task to determine sentiment, we use the typical accuracy metric based on the True Positive and Negative, False Positive and Negative values for predicted class affiliation.

4 Results

In this section, we show the results of the applied topic modeling and sentiment analysis. Besides a quantitative comparison and discussion of the various approaches and algorithms, we illustrate the importance of preproessing and feature extraction and suggest how those insights are useful for practitioners.

4.1 Results of Topic Modeling

The output of the top-N-words with $N = 10$ topics, for example, can give an initial overview which topics could be potentially retrieved by topic modeling. While few topics seem to include incoherent words, there are some product-related topics about the camera quality, design, display, issues about the missing charger and headphones, or more generally, country-specific topics like the upcoming 5G network.

While a manual choice of the number of topics can lead to misleading results, a way to find the most suitable model and also the number of topics which allows to represent the Tweets in a meaningful way, is to utilize the CV-coherence value as a metric. We iterate the number of topics from 5 to 20 as less topics will probably have no specific meanings and more topics too specific meanings. For every number of topics, the coherence value for the LDA, LSA and NMF is computed while differentiating for every model whether the BOW or TF-IDF approach is applied for feature extraction. The results are displayed in Fig. 1 for English on the top and for German Tweets at the bottom. In both cases, the TF-IDF feature extraction approach leads to higher coherence values, while the algorithms score different values for each language. For the English Tweets, the LDA gives high coherence values for few topics but lower values as the number of topics rise. In contrast, the NMF performs better with a higher number of topics while LSA does not give high results at all. For German Tweets, NMF also performs well for a lower number of topics and outperforms LSA and LDA, while LSA is slightly better than LDA. This result suggests that there seems to be no algorithm for topic modeling which should be preferred in SMA, since LDA, LSA and NMF achieve various performances dependent on the number of topics and language. Another aspect we investigate, is the impact of different preprocessing step combinations on the results. Therefore, the averages of the coherence values are taken for a different number of topics (5 to 20), different algorithms (LDA, LSA, NMF) and different feature extraction methods (BOW, TF-IDF), resulting in an average of 90 values for each combination of preprocessing steps. We consider the

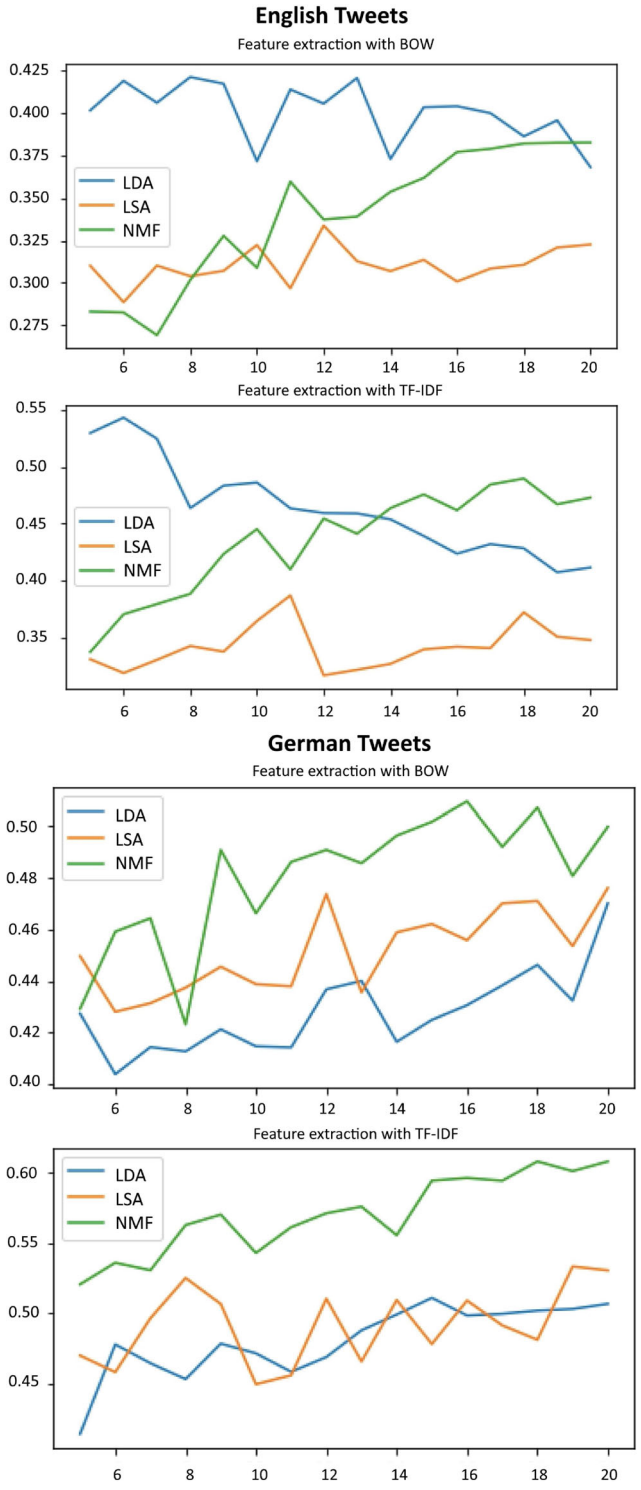

Fig. 1 Comparison of coherence values for different algorithms, number of topics and feature extraction methods for English and German Tweets

following preprocessing steps: Tokenization (1), lower case writing (2) and the removal of URLs and mentions (3), hashtags (4), emoticons (5), punctuation marks (6), numbers (7) and stop words (8), which we combine arbitrarily. Additionally, we add to each combination exactly one stemming or lemmatization approach, choosing from the Snowball (9) and Cistem (10) stemming and SpaCy (11) lemmatization for German and from the Porter (12) and Snowball (13) stemming and SpaCy (14) and NLTK (15) lemmatization for the English language.

In Table 1, we present the results of the average coherence values (ACV) for chosen combinations of different preproceesing steps for English Tweets. The aggregation of different numbers in brackets indicates which from the above steps are considered together. In every combination, we include tokenization as a basic requirement for the other steps. As lower case writing is typical in English, the preprocessing step (2) does not contribute to a better coherence such that we excluded it from the next combinations. While the adding of steps (3) to (7) just improve the result slightly, a bigger positive impact on the coherence value can be reached by including the removal of stop words, step (8), which leads to the best combination in this scenario of (1) + (3) + (4) + (7) + (8) with a coherence value of 0.39817 (indicated as bold value in Table 1). In contrast, the addition of different stemming and lemmatization steps on this combination have a negative impact on the coherence value, but among these different approaches, the Porter stemming method (Porter, 1980) with an ACV of 0.37274 and the NLTK lemmatization method with an ACV of 0.38534 give just slightly worse values. This result could indicate that it is not always advisable to reduce words to their word stem in the context of social media. This is in accordance with previous findings from Bao et al. (2014). Similarly, the removal of stop words also has a large impact for German Tweets, illustrated in Table 2, as such words probably do not add any explanatory value. Here, the inclusion of more preprocessing steps gives in general better results such

Table 1 Impact of different combinations of preprocessing steps on the average coherence value (ACV) for English Tweets

Combination	ACV
(1)	0.32151
(1) + (2)	0.31008
(1) + (3)	0.32310
(1) + (3) + (4)	0.32881
(1) + (3) + (4) + (5)	0.32797
(1) + (3) + (4) + (6)	0.32482
(1) + (3) + (4) + (7)	0.33307
(1) + (3) + (4) + (7) + (8)	**0.39817**
(1) + (3) + (4) + (7) + (8) + (12)	0.37274
(1) + (3) + (4) + (7) + (8) + (13)	0.35651
(1) + (3) + (4) + (7) + (8) + (14)	0.33933
(1) + (3) + (4) + (7) + (8) + (15)	0.38534

Table 2 Impact of different combinations of preprocessing steps on the average coherence value (ACV) for German Tweets

Combination	ACV
(1)	0.28906
(1) + (2)	0.30577
(1) + (2) + (3)	0.31077
(1) + (2) + (3) + (4)	0.31850
(1) + (2) + (3) + (4) + (5)	0.31785
(1) + (2) + (3) + (4) + (6)	0.32018
(1) + (2) + (3) + (4) + (6) + (7)	0.33501
(1) + (2) + (3) + (4) + (6) + (7) + (8)	**0.47827**
(1) + (2) + (3) + (4) + (6) + (7) + (8) + (9)	0.45638
(1) + (2) + (3) + (4) + (6) + (7) + (8) + (10)	0.44331
(1) + (2) + (3) + (4) + (6) + (7) + (8) + (11)	0.45425

that the best result is achieved by including the steps (1) to (8) with a coherence value of about 0.47827 (indicated as bold value in Table 2). As capitalization is common in the German language, the lower case preprocessing has a positive impact on the result as this step probably reduces the differentiation between lower case and capitalized words. Just as for English Tweets, the inclusion of stemming and lemmatization does not improve the results and all three considered approaches of the Snowball, Cistem and SpaCy Python packages deliver nearly the same ACV.

Next, we start again an iteration with the different algorithms LDA, LSA and NMF, different feature extraction methods and different number of topics, ranging from 5 to 20, with the best determined combination of preprocessing steps for English and German Tweets. In unison with previous results, we find that the TF-IDF feature extraction approach gives in general better results. For the English Tweets, LDA still performs best and gives with 7 topics and TD-IDF feature extraction the best overall result among all iterations with a coherence value of 0.57510. For the German Tweets, NMF with TD-IDF and 19 topics gives the best result among the iterations with a coherence value of 0.60704.

4.2 Results of Sentiment Analysis

First, we present the results of the lexicon-based approach. We compare the accuracy of the lexicons VADER with an accuracy value of 0.56908 and SentiWordNet with 0.35843 for the English language and GERVADER with 0.60593 and SentiWS with an accuracy of 0.49539 for German Tweets. We can see that the two lexicons VADER and GerVADER, which are specialized for social media texts, give better results than the traditional ones. But in comparison to learning-based, lexicon-based approaches give in general worse accuracy results. In Fig. 2, the accuracy of NB, LR, SVM, NN, DT and RF is displayed for English Tweets, differentiating BOW and TF-IDF feature extraction. While RF with BOW gives the best result with an accuracy of approximately 0.77, NB and NN perform worst while LR and SVM lead to a high accuracy. With TD-IDF feature extraction, the results are nearly identically. In contrast, RF could not give a same high accuracy for German Tweets, displayed in Fig. 3, but better results with TD-IDF. LR and SVM, for both BOW and TD-IDF, give the best results with an accuracy around 0.9 for German Tweets, indicating that these methods seem to be robust across these two languages. Another interesting aspect is that the accuracy of NN increases by around 15 % when TD-IDF instead of BOW is applied, demonstrating the importance of the feature extraction method for some models.

In a similar way as we did for topic modeling, we want to investigate the impact of preprocessing steps on the results, displayed in Table 3 for English and Table 4 for German Tweets, and using the same notation for the single steps as in Section 4.1. The average accuracy on the right side of the tables is calculated based on the six accuracy values of the previously tested methods NB, LR, SVM, NN, DT and RF with BOW and TF-IDF feature extraction, resulting in an

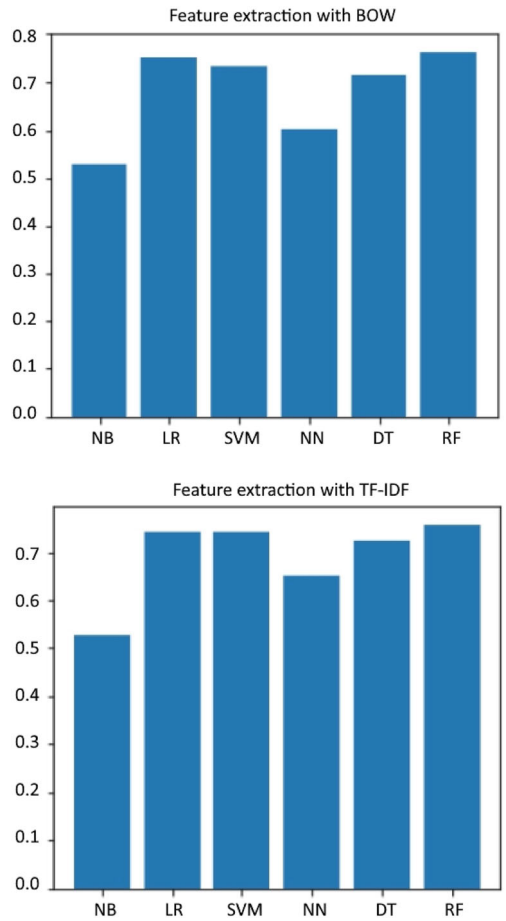

Fig. 2 Comparison of accuracy of different classification models for English Tweets

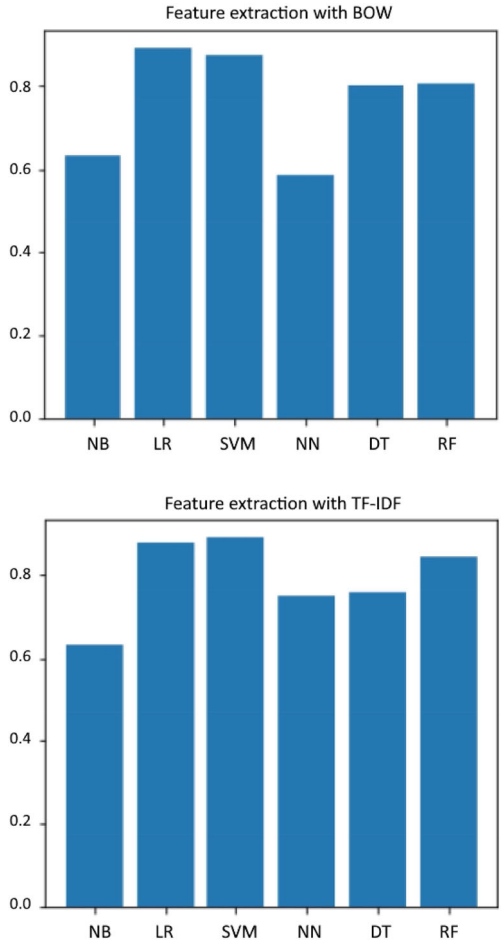

Fig. 3 Comparison of accuracy of different classification models for German Tweets

average of 12 values for every displayed preprocessing step combination. From Table 3 we can see that the inclusion of more preprocessing steps gives in general better results for English Tweets, except for steps (3), (5) and (6). In contrast

Table 3 Impact of different combinations of preprocessing steps on the accuracy for English Tweets

Combination	Average accuracy
(1)	0.66930
(1) + (2)	0.66930
(1) + (2) + (3)	0.66646
(1) + (2) + (4)	0.67193
(1) + (2) + (4) + (5)	0.67144
(1) + (2) + (4) + (6)	0.67011
(1) + (2) + (4) + (7)	0.68750
(1) + (2) + (4) + (7) + (8)	0.69397
(1) + (2) + (4) + (7) + (8) + (12)	0.69640
(1) + (2) + (4) + (7) + (8) + (13)	0.69417
(1) + (2) + (4) + (7) + (8) + (14)	**0.70044**
(1) + (2) + (4) + (7) + (8) + (15)	0.69579

Table 4 Impact of different combinations of preprocessing steps on the accuracy for German Tweets

Combination	Average accuracy
(1)	0.76768
(1) + (2)	0.77024
(1) + (2) + (3)	0.78324
(1) + (2) + (3) + (4)	0.77834
(1) + (2) + (3) + (5)	0.78260
(1) + (2) + (3) + (6)	0.78324
(1) + (2) + (3) + (7)	0.77621
(1) + (2) + (3) + (8)	0.79305
(1) + (2) + (3) + (8) + (9)	0.79092
(1) + (2) + (3) + (8) + (10)	0.78388
(1) + (2) + (3) + (8) + (11)	**0.81160**

to topic modeling, the usage of a stemming or lemmatization method results in a higher accuracy, where the combination with the SpaCy lemmatization leads to the best result of an accuracy of 0.70044 (indicated as bold value in Table 3). Nonetheless, the differences in the accuracy are very small, such that the combination (1) + (2) + (4) + (7) + (8) with an accuracy of 0.69397 is nearly identical to the best result. Thus, it could be argued if stemming or lemmatization is even necessary and, like in the case of topic modeling, the rest of the preprocessing steps are already sufficient. A similar pattern can be identified for German Tweets in Table 4, since the addition of lemmatization gives the best result of 0.81160 (indicated as bold value in Table 4), but just slighly better than the combination (1) + (2) + (3) + (8) of 0.79305 without it. Interestingly, for both English and German, the removal of stop words does not have such a big impact for sentiment analysis than for topic modeling. This could be the case since stop words could represent noisy data which interfere a clear distinction across topics, but do not have a critical impact on the sentiment.

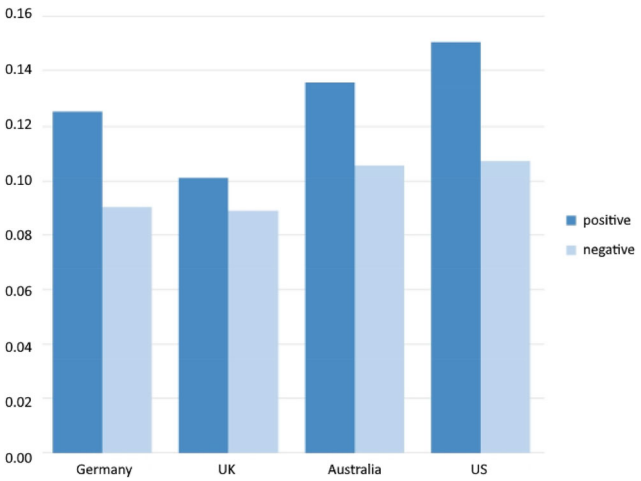

Fig. 4 Comparison of sentiment between different countries

4.3 Comparison between Languages and Countries for Product Insights

In Fig. 4, the sentiment between the four countries Germany, the UK, Australia and the US is compared. We can see that in every country, the number of positive tweets exceeds the negative ones, while the consumers interested in Apple products on Twitter in the US seem to have the most positive attitude towards the new models of the iPhone 12, while the share of positive opinions from the UK is the lowest among the considered countries. For gaining deeper insights about the causes of positive and negative sentiment towards the products, we generate word clouds which display the most common words among a selected corpus. Thus, we can get valuable hints about the relevance of certain product attributes which are perceived good or bad. In Fig. 5, the word clouds for positive (left side) and negative (right side) sentiment are displayed for Germany in the first row, US in the second, UK in the third and Australia in the fourth row. Some exemplary insights about certain product attributes can be derived from Fig. 5. The iPhone 12 Pro model is favored the most in all four countries according to these word clouds, the Mini model is perceived as quite positive in Germany, the UK and Australia, while the Max and Mini Model seem to be equally likable in the US. This hint can give Apple executives important feedback which models are especially popular. The design in general is a positive aspect across all countries, especially in Germany and Australia, also blue seems to be the favorite color for the product among all considered countries. In the English speaking countries, the camera of the iPhone is getting approval on Twitter. Bad aspects are the missing charger and headphones across all countries, also a non-integrated USB-C-port is a main reason for negative sentiment. Aspects, which were perceived as positive like certain models are also criticized to a small extent as well as the display, while the upcoming,

Fig. 5 Wordclouds in four rows for Germany, US, UK and Australia, respectively, where the left side represents positive and the right side negative sentiment

not product-related 5G standard seems to affect only in Germany in a negative way.

4.4 Managerial Implications

The results suggest that a sentiment analysis is possible with a high accuracy and together with topic modeling, relevant aspects which are considered good or bad by consumers could be identified. Thus, such approach can be applied to any other product or service and in a frequent manner, say every week, to assess the sentiment and upcoming complaints on a regular basis. Not only for own products and services such kind of market research could be possible, also opposing businesses could be monitored to compare competing assortment and draw conclusions for future strategy and product development. An advantage of such approach in contrast to consumer interviews is that opinions can be observed without interference or interviewer-bias. This could be extended for multiple languages, countries and domains to capture the overall picture and brand image of an international company which could be helpful to adjust the global strategy. Furthermore, such analysis could be part of a decision support system, where consumers' perceptions of products or services can be assessed in real time. In the case of the insights collected on Twitter for the iPhone 12 models of Apple, there is evidence that there seems to be potential for improvement for the display and USB-port while consumers were dissatisfied with the missing charger and headphones.

5 Conclusion

In this paper, we investigated how data from Twitter can be utilized to assess opinions and sentiment of consumers towards a new product (in this case the new iPhone 12 models from Apple) via topic modeling and sentiment analysis. The single steps data collection, preprocessing, model building and evaluation were described and illustrated with Tweets from four countries and two languages. A lexicon-based and learning-based approach were applied and compared for the sentiment analysis and the algorithms LDA, LSA and NMF for topic modeling. The importance of preprocessing steps and feature extraction were highlighted to improve the accuracy of the models, in the case of sentiment analysis, the learning-based approach seems to be superior to the lexicon-based, probably also due to the specialties in such social media context. Also, the TD-IDF feature extraction approach seems to be superior to the BOW method. Through the comparison between English and German Tweets, we could find that neither one specific algorithm for topic modeling, nor a certain combination of preprocessing steps can be identified as best fitting for both languages, the same applies for the choice of a machine learning algorithm for sentiment analysis. Although there exist substantially steps towards adjusted NLP approaches for handling the specialties of social media, several improvements are desirable. One of them would be to make sense of the often used abbreviations which are partly necessary in the case of Twitter where a restricted number of symbols are allowed per Tweet. Also the interpretation of emoticons could help to identify the right sentiment, especially in cases where sarcasm or irony is involved. The processing of hashtags and correction of spelling checks to identify existing words would be another aspect which could further improve the previous methods. Thus, for future research, instead of just removing URLs and mentions, hashtags and emoticons, new methods could be developed to include these aspects into the preprocessing to gain deeper insights into texts from social media. In accordance to previous research, we could confirm that in our case, the inclusion of stemming and lemmatization gives just slightly better results for sentiment analysis or even worse results for topic modeling, thus it could be argued if these steps are necessary, since valuable information could get lost by reducing a word to its word stem. In case of the German adaption of the social media NLP tool VADER, several improvements are necessary to remove issues like the non-detection of negations. A further challenge is the data itself, since it is not always clear if a Tweet is coming from a real person or a bot. When even political manipulation could be achieved through the usage of bots (Badawy et al., 2018) or general spam (Aswani et al., 2018), opposing businesses could also utilize such strategies to damage the reputation of competing products. In such cases, Twitter is required to prevent further emerge of such bots since the whole sentiment could be distorted. Especially for products, where many enterprises compete for a limited market share, malicious bots could be programmed to lower the sentiment of a competitor. Nonetheless, the general quality in terms of readability, completeness, usefulness and trustworthiness of Twitter data seems to be high (Arolfo et al., 2020). Despite the deprecation of several social media APIs like in the case of Facebook or Instagram due to data scandals, Twitter even extended the functionality of their API in the last years and plans to improve it further. Thus, one can expect that SMA especially on Twitter will stay not only for practitioners a relevant tool, but also allow researchers to conduct further experiments on real data.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. In *LSM '11: proceedings of the workshop on languages in social media, (pp. 30–38)*.

Arolfo, F., Rodriguez, K. C., & Vaisman, A. (2020). Analyzing the quality of twitter data streams information systems frontiers. https://doi.org/10.1007/s10796-020-10072-x.

Aswani, R., Kar, A. K., & Ilavarasan, P.V. (2018). Detection of spammers in twitter marketing a hybrid approach using social media analytics and bio inspired computing. *Information System Frontiers*, *20*, 515–530. https://doi.org/10.1007/s10796-017-9805-8.

Awal, G. K., & Bharadwaj, K. K. (2019). Leveraging collective intelligence for behavorial prediction in signed social networks through evolutionary approach. *Information Systems Frontiers*, *21*, 417–439. https://doi.org/10.1007/s10796-017-9760-4.

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation the 2016 russian interference twitter campaign. In *2018 IEEE/ACM International conference on advances in social networks analysis and mining (ASONAM), (pp. 258–265). https://doi.org/10.1109/ASONAM.2018.8508646*.

Bao, Y., Quan, C., Wang, L., & Ren, F. (2014). The role of pre-processing in twitter sentiment analysis. In D. S. Haung, K. H. Jo, & L. Wang (Eds.) *Intelligent computing methodologies, volume 8589 of lecture notes in computer science* (pp. 615–624). Cham: Springer. https://doi.org/10.1007/978-3-319-09339-0_62.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 933–1022.

Chamlertwat, W., Bhattarkosol, P., & Rungkasiri, T. (2012). Discovering consumer insight from twitter via sentiment analysis. *Journal of Universal Computer Science*, *18*(8), 973–992.

Chang, W.-L., & Chen, Y.-P. (2019). Way too sentimental? a credible model for online reviews. *Information Systems Frontiers*, *21*, 453–468. https://doi.org/10.1007/s10796-017-9757-z.

Cheong, M., & Lee, V. C. S. (2011). A microblogging-based approach to terrorism informatic: exploration and chronicling civilian sentiment and response to terrorism events via twitter. *Information Systems Frontiers*, *13*, 45–59. https://doi.org/10.1007/s10796-010-9273-x.

Cruz, F. L., Troyano, J., Enríquze, F., Ortega, J., & Vallejo, C.G. (2013). 'Long autonomy or long delay?' the importance of domain in opinion mining. *Expert Systems with Applications*, *40*(8), 3174–3184. https://doi.org/10.1016/j.eswa.2012.12.031.

Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P.J. (2020). An evaluation of document clustering and topic modelling in two online social networks: twitter and reddit. *Information Processing & Management*, *57*(2), 102034. https://doi.org/10.1016/j.ipm.2019.04.002.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Ding, X., Liu, B., & Yu, P.S. (2008). A holistic lexicon-based approach to opinion mining. In *WSDM '08: Proceedings of the 2008 international conference on web search and data mining, (pp. 231–240). https://doi.org/10.1145/1341531.1341561*.

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the fifth international conference on language resources and evaluation (LREC'06), (pp. 417–422)*.

Fahrni, A., & Klenner, M. (2008). Old wine or warm beer: target-specific sentiment analysis of adjectives. In *Symposium on affective language in human and machine, (pp. 60–63)*.

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys*, *49*(2), 28. https://doi.org/10.1145/2938640.

Haghighi, N. N., Liu, X. C., Wei, R., Li, W., & Shao, H. (2018). Using twitter data for transit performance assessment: a framework for evaluating transit riders' opinions about quality of service. *Public Transport*, *10*, 363–377. https://doi.org/10.1007/s12469-018-0184-4.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *35Th annual meeting of the association for computational linguistics and 8th conference of the european chapter of the association for computational linguistics, pp. (174–181). https://doi.org/10.3115/976909.979640*.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the eighth international AAAI conference on weblogs and social media, (pp. 216–225)*.

Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., & Nerur, S. (2018). Advances in social media research past, present and future. *Information Systems Frontiers*, *20*, 531–558. https://doi.org/10.1007/s10796-017-9810-y.

Kemp, S. (2020). Digital 2020: global digital overview. https://datareportal.com/reports/digital-2020-global-digital-overview.

Kostygina, G., Tran, H., Shi, Y., Kim, Y., & Emery, S. (2016). 'Sweeter Than a Swisher': Amount and themes of little cigar and cigarillo content on twitter. *Tobacco Control*, *25*, 75–82.

Kouloumpis, E., Wilson, T., & Moore, J.D. (2011). Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of the fifth international conference on weblogs and social media, (pp. 538–541), AAAI Press*.

Lau, J. H., Collier, N., & Baldwin, T. (2012). On-line trend analysis with topic models: #twitter trends detection. In *Proceedings of COLING 2012, (pp. 1519–1534)*.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016.

Ostrowski, D. A. (2015). Using latent dirichlet allocation for topic modelling in twitter. In *Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015), (pp. 493–497). https://doi.org/10.1109/ICOSC.2015.7050858*.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1–135. https://doi.org/10.1561/15000000011.

Park, S. B., Ok, C. M., & Chae, B.K. (2016). Using twitter data for cruise tourism marketing and research. *Journal of Travel & Tourism Marketing*, *33*(6), 885–898. https://doi.org/10.1080/10548408.2015.1071688.

Popescu, A.-M., & Etzioni, O. (2005). Extracting product features and opinions from reviews. In *HLT '05: Proceedings of the conference*

on human language technology and empirical methods in natural language processing, (pp. 339–346). https://doi.org/10.3115/1220575.1220618.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems*, *14*(3), 130–137. https://doi.org/10.1108/eb046814.

Qiu, G., Liu, B., Bu, J., & Chen, C.L.P. (2009). Expanding domain sentiment lexicon through double propagation. In *Proceedings of the 21st international joint conference on artificial intelligence (IJCAI 2009), (pp. 1199–1204)*.

Rathore, A. K., Kar, A. K., & Ilavarasan, P.V. (2017). Social media analytics: literature review and directions for future research. *Decision Analysis*, *14*(4), 229–249. https://doi.org/10.1287/deca.2017.0355.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *WSDM '15: Proceedings of the eighth ACM international conference on web search and data mining, (pp. 399–408). https://doi.org/10.1145/2684822.2685324*.

Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. In P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. X. Parreira, J. Hendler, G. Schreiber, A. Bernstein, & E. Blomqvist (Eds.) *The Semantic Web - ISCW 2012, volume 7649 of lecture notes in computer science* (p. Berlin). Springer. https://doi.org/10.1007/978-3-642-35176-1_32.

Severyn, A., & Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. In *SIGIR '15: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, (pp. 959–962). https://doi.org/10.1145/2766462.2767830*.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526.

Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social media analytics - An interdisciplinary approach and its implications for information systems. *Business & Information Systems Engineering*, *6*(2), 89–96. https://doi.org/10.1007/s12599-014-0315-7.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics - challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002.

Swain, A. K., & Cao, R. Q. (2019). Using sentiment analysis to improve supply chain intelligence. *Information Systems Frontiers*, *21*, 469–484.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676.

Tymann, K. M., Lutz, M., Palsbröker, P., & Gips, C. (2019). GerVADER - a german adaption of the VADER sentiment analysis tool for social media texts. In *Proceedings of the conference "Lernen, Wissen, Daten, Analysen" (LWDA 2019), (pp. 178–189)*.

Ustek-Spilda, F., Vega, D., Magnani, M., Rossi, L., Shklovski, I., Lehuede, S., & Powell, A. (2021). A twitter-based study of the european internet of things. *Information Systems Frontiers*, *23*, 135–149. https://doi.org/10.1007/s10796-020-10008-5.

Vallejos, S., Alonso, D. G., Caimmi, B., Berdun, L., Armentano, M. G., & Soria, A. (2021). Mining social networks to detect traffic incidents. *Information Systems Frontiers*, *23*, 115–134. https://doi.org/10.1007/s10796-020-09994-3.

Xie, W., Zhu, F., Jiang, J., Lim, E., & Wang, K. (2016). TopicSketch, real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering*, *28*(8), 2216–2229. https://doi.org/10.1109/TKDE.2016.2556661.

Zhang, L., Ghosh, R., Dekhil, R., Hsu, M., & Liu, B. (2011). Combining lexicon-based and learning-based methods for twitter sentiment analysis. Technical Report 89 HP Laboratories.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, & V. Mudoch (Eds.) *Advances in information retrieval, volume 6611 of lecture notes in computer science* (pp. 338–349). Berlin: Springer. https://doi.org/10.1007/978-3-642-20161-5_34.

Irina Wedel received her M.Sc. in Information Systems at the University of Hamburg. Her current research interests are mainly in the areas of data mining, machine learning and social media analytics. She is particularly interested in the field of natural language processing with a focus on subtopics such as sentiment analysis and topic modeling.

Michael Palk is a PhD candidate in the Institute of Information Systems at the University of Hamburg. He holds a M.Sc. in Business Mathematics (passed with distinction) from the University of Hamburg. His current research interests are centered around social media, web and distributed ledger analytics. He serves as a reviewer for various journals and conferences. His industry experiences include positions at Kühne+Nagel, Ginkgo Analytics and Nordmetall.

Stefan Voß is professor and director of the Institute of Information Systems at the University of Hamburg. Moreover, he serves as dean of the Hamburg Business School (School of Business Administration). Previous positions include full professor and head of the department of Business Administration, Information Systems and Information Management at the University of Technology Braunschweig (Germany) from 1995 up to 2002. He holds degrees in Mathematics (diploma) and Economics from the University of Hamburg and a Ph.D. and the habilitation from the University of Technology Darmstadt. His current research interests are in quantitative / information systems approaches to supply chain management and logistics including public mass transit and telecommunications. He is author and co-author of several books and several hundred papers in various journals. In the German Handelsblatt ranking he is continuously within the top 20 professors in business administration within the German speaking countries. The most recent Wirtschaftswoche ranking lists him among the top 10. Stefan Voß serves on the editorial board of some journals including being Editor of Public Transport. He is frequently organizing workshops and conferences. Furthermore, he is consulting with several companies.