SN Computer Science (2023) 4:477
https://doi.org/10.1007/s42979-023-01886-y

ORIGINAL RESEARCH

A Comparison of Commercial Sentiment Analysis Services

Tatiana Ermakova[1,2,3,4] · Benjamin Fabian[2,5,6,7] · Elena Golimblevskaia[2,3] · Max Henke[7]

Received: 16 July 2022 / Accepted: 4 May 2023

Abstract
Empirical insights into promising commercial sentiment analysis solutions that go beyond the claims of their vendors are rare. Moreover, due to the constant evolution in the field, previous studies are far from reflecting the current situation. The goal of this article is to evaluate and compare current solutions using two experimental studies. In the first part of the study, based on tweets about airline service quality, we test the solutions of six vendors with different market power, such as Amazon, Google, IBM, Microsoft, Lexalytics, and MeaningCloud, and report their measures of accuracy, precision, recall, (macro)F1, time performance, and service level agreements (SLA). Furthermore, we compare two of the services in depth with multiple data sets and over time. The services tested here are Google Cloud Natural Language API and MeaningCloud Sentiment Analysis API. For evaluating the results over time, we use the same data set as in November 2020. In addition, further topic-specific and general Twitter data sets are used. The experiments show that the IBM Watson NLU and Google Cloud Natural Language API solutions may be preferred when negative text detection is the primary concern. When tested in July 2022, the Google Cloud Natural Language API was still the clear winner compared to the MeaningCloud Sentiment Analysis API, but only on the airline service quality data set; on the other data sets, both services provided specific benefits and drawbacks. Furthermore, we detected changes in the sentiment classification over time with both services. Our results motivate that an independent, critical, and longitudinal experimental analysis of sentiment analysis services can provide interesting insights into their overall reliability and particular classification accuracy beyond marketing claims to critically compare solutions based on real data and analyze potential weaknesses and margins of error before making an investment.

Keywords Sentiment analysis · Machine learning · Text classification · Commercial service · SaaS · Cloud computing

Introduction

An explosive growth of Web 2.0 applications (e.g., social media platforms) has resulted in an almost continuous stream of publicly available digital opinions [1]. Sentiment analysis enables automated opinion recognition and polarity

This article is part of the topical collection "Web Information Systems and Technologies 2021" guest edited by Joaquim Filipe, Francisco Domínguez Mayo and Massimo Marchiori.

✉ Benjamin Fabian
benjamin.fabian@th-wildau.de

Tatiana Ermakova
tatiana.ermakova@htw-berlin.de

Elena Golimblevskaia
elena.golimblevskaia@fokus.fraunhofer.de

Max Henke
max.henke@hhl.de

[1] Hochschule für Technik und Wirtschaft (HTW) Berlin, Wilhelminenhofstraße 75A, 12459 Berlin, Germany

[2] Weizenbaum Institute for the Networked Society, Hardenbergstraße 32, 10623 Berlin, Germany

[3] Fraunhofer Institute for Open Communication Systems (FOKUS), Kaiserin-Augusta-Allee 31, 10589 Berlin, Germany

[4] Technical University of Berlin, Einsteinufer 25, 10587 Berlin, Germany

[5] Technical University of Applied Sciences Wildau (TH Wildau), Hochschulring 1, 15745 Wildau, Germany

[6] Humboldt University of Berlin, Spandauer Str. 1, 10178 Berlin, Germany

[7] Hochschule für Telekommunikation Leipzig (HfTL), Gustav-Freytag-Straße 43-45, 04277 Leipzig, Germany

Published online: 24 June 2023

classification [2]. Taken together, this offers organizations unprecedented opportunities to support and improve decision-making processes [3]. Recent research shows that firms can leverage user-generated content in the form of sentiments to predict and/or explain various aspects of their performance, such as sales [4–6], profits [7], brand perception [8], customer satisfaction and market performance [9], and stock trade performance [10].

Developing proprietary sentiment analysis technologies require years of experience in data science and coding [11], as well as related sufficient resources, such as human resources, large amounts of rare data, GPU support, large storage for the data sets, etc. [12]. In contrast, suitable commercial "software as a service" (SaaS) tools provide a convenient quickly accessible, easily configurable, and cost- and time-efficient on-demand solution [13]. Indeed, no special prior knowledge is required, considering that in 2020 alone, a total of 112 papers were published on sentiment analysis, based only on the Deep Learning approach, one of the possible approaches [14]. Furthermore, the programming effort as well as the implementation and integration of the solutions into the internal processes either remain manageable or are even reduced to almost zero. Billing is based on the service provided [13].

Choosing an appropriate solution can be a challenge. Empirical findings on the sentiment services established in industry that go beyond the claims of their providers are rather limited and, due to the constant evolution of the field, are far from being able to reflect the current situation after a few years [15–19], with the notable exceptions of [20] and an investigation of ensemble approaches based on such services [21]. With this in mind, the goal of this study is to evaluate and compare current commercial SaaS solutions for sentiment analysis offered by cloud providers with varying degrees of market power, with respect to a wide range of established classification performance measures, such as accuracy, precision, recall, and (macro) F1 [22, 23], as well as usage characteristics, such as time performance and service level agreements (SLA). The well-established evaluation framework applied to the solutions in this study enables an independent comparison of the solutions in terms of their functional requirements. This study can, therefore, provide a basis or guidance for selecting a solution. In addition, this study can potentially provide motivation and ideas for further development of the solutions.

In particular, in November 2020, we test services from four major cloud platforms—IBM, Amazon, Microsoft, and Google—that have been investigated in recent studies in this area [20, 21], as well as solutions such as Lexalytics Semantria API [16, 18], and MeaningCloud Sentiment Analysis API (as of November 2020), which, to our knowledge, are still subject to recent and rigorous evaluation. We rely on a real-world Twitter data set of 14,640 airline service quality entries, which was also used in a comparative study of deep learning models in sentiment analysis [24] and is comparable to the data sets used in other related studies [20, 21].

In July 2022, we compare two of the services in depth on multiple data sets and after a longer time period. In this part, we test Google Cloud Natural Language API and MeaningCloud Sentiment Analysis API (as of July 2022) on the same data set as in November 2020 to evaluate differences in results over time. In addition, we test these services on two further real-world Twitter data sets: 7,064 service quality entries related to Southwest Airlines and 162,980 general tweets.

The paper is organized as follows: in "Background and Foundations", we provide an introduction to the applications and fundamentals of sentiment analysis to prepare the motivation and background of our experimental approach. Then, in "Related Work on Sentiment Services", we discuss previous research on industrial cloud services for sentiment analysis. We then present the experimental setup in "Experimental Design", explicitly discussing the data sets used, the sentiment analysis solutions studied, and the implementations. In "Results", we present the results of two studies. Finally, we summarize and discuss our findings, point out limitations, and make recommendations for further research.

Background and Foundations

Application Areas of Sentiment Analysis

Opinions and sentiments are of value to a wide range of stakeholder groups in politics, business, and society. For example, opinions and sentiments of citizens are of particular interest in the political environment [25, 26]. Through social media, citizens' public expressions are widely accessible. These can be used by governments and political organizations to gain insights into the needs and moods of voters. In the past, this required traditional methods such as polls conducted by opinion research institutes, which involved a great deal of effort and a certain time delay.

In the business context [27–29], there are several decisions and activities that are based on the interests of customers. Due to the wide availability of public opinions on the Internet, sentiment analysis can provide valuable insights. One of the most widespread application areas in research is marketing, as this area can benefit most from a comprehensive understanding of customer needs. For example, the long-term viability of a company depends to a large extent on its ability to satisfy customer needs with suitable products and thus create sustainable brand value. To do this, companies need information about consumer preferences and demand to understand perceptions of the products they buy. This information can be used to develop suitable strategies for branding and positioning their own products in the market. Accordingly, marketing decision

makers need to know how their own brand is perceived by the target group. Opinion mining can be used here to analyze customer perceptions in comparison with other brands in the industry and to identify the aspects most relevant to the brand image. Another application for sentiment analysis is sales forecasting, especially for product launches, by collecting sentiment data on public perception.

Technical Foundations

Sentiment analysis as one of the areas of affective computing is about detecting, analyzing, and evaluating people's state of mind towards various events, products, services, etc. [30] More precisely, this area aims at detecting opinions, moods and emotions based on human actions by means of writing, facial expressions, speech, movements, etc., without going into the analysis of these feelings. Here, our focus is exclusively on the analysis of textual feelings.

A sentiment can be defined as a triplet, (y, o, i), where y describes the target of the sentiment, o the orientation of the sentiment, and i the intensity of the sentiment [1]. In its orientation (which is also often called polarity, tonality, or semantic orientation), a sentiment can be positive, negative, or neutral. Neutrality usually means the absence of any sentiment. Furthermore, a sentiment can also differ in intensity within the same sentiment polarity (e.g., the use of *perfect* vs. *good*).

Sentiment polarity classification can be accomplished at three levels in terms of granularity: the document level, the sentence level, and the aspect level [30]. At the document level of sentiment analysis, the whole document, regardless of its length, is considered as the atomic unit, and the polarity of the whole document is studied [30]. The analysis at the document level implicitly assumes that a document expresses only one opinion about a single entity [1] and, hence, can be too coarse for practical use [5].

At the sentence level, it is first checked whether a sentence expresses opinion or only states facts without implication. Aspect-level analysis focuses directly on opinions and their target [1]. For instance, the frequency-based analysis method searches for frequent nouns or compound nouns (POS tags). An often-used rule of thumb says that when a (compound) noun occurs in 1% or more sentences, it can be considered as an aspect [14]. This level of sentiment analysis is very valuable for entrepreneurs and policy makers interested in summarizing the opinions of individuals on certain features of their products or/and services, where applying sentiment analysis at the document or sentence level is not enough [30]. A recent study of dimension-specific mood effects on product sales showed that for low-budget movies, a positive relationship with movie sales was stronger for plot sentiment than for lead sentiment, while for high-budget movies, a positive relationship with movie sales was stronger for lead sentiment than for plot or genre sentiment [5].

The approaches used in sentiment analysis can be grouped into three categories: (1) lexicon-based approaches; (2) machine learning approaches [31, 32]; (3) hybrid approaches that couple the previous ones [33]; and (4) graph-based approaches that are based on the assumption that Twitter users influence one another [22, 34]. Lexicon-based approaches in sentiment analysis make use of a sentiment lexicon to estimate the overall sentiment polarity of a document as the aggregation of the sentiment polarities of the individual words within the document and, hence, do not require labelled data. Lexicon-based approaches can comprise (a) dictionary-based techniques, and (b) corpus-based techniques.

Dictionary-based techniques use a sentiment lexicon to label terms with sentiment polarity. A sentiment lexicon usually consists of words labeled with a sentiment polarity and its strength [35], such as Multi-Perspective Question Answering (MPQA) Subjectivity Lexicon [36], Bing Liu's Opinion Lexicon, NRC Valence, Arousal, and Dominance (VAD) lexicon [37], NRC Word-Emotion Association Lexicon (EmoLex) [38], NRC Emotion/Affect Intensity Lexicon [39], SentiWordNet [40], SenticNet [41], WordNet-Affect [42], General Inquirer, or Linguistic Inquiry, and Word Count (LIWC), which have also been summarized and explained in earlier work [30, 43].

Corpus-based techniques use co-occurrence statistics or syntactic patterns in a text corpus and a small set of paradigmatic positive and negative starting words and create a domain-, context-, or topic-specific lexicon [35]. The semantic orientation of the word can be assigned from the measure of its association with a set of predefined words with positive semantic orientation minus the measure of its association with a set of predefined words with negative semantic orientation [44]:

$$\mathrm{SO}-A(\mathrm{word}) = \sum_{\mathrm{pword}\in\mathrm{Pwords}} A(\mathrm{word},\mathrm{pword}) - \sum_{\mathrm{nword}\in\mathrm{Nwords}} A(\mathrm{word},\mathrm{nword}),$$

where

$\mathrm{Pwords} = \{\mathrm{good},\ \mathrm{nice},\ \mathrm{excellent},\ \mathrm{fortunate}\}$ and

$\mathrm{Nwords} = \{\mathrm{bad},\ \mathrm{nasty},\ \mathrm{poor},\ \mathrm{unfortunate}\}$.

When the value of $\mathrm{SO}-A(\mathrm{word})$ is positive, the word is marked with a positive semantic orientation, and with a negative semantic orientation otherwise. The higher the value of $\mathrm{SO}-A(\mathrm{word})$, the stronger the sentiment strength of the word. The measure of association can be exemplified by Pointwise Mutual Information (PMI):

$$A(\text{word}_1, \text{word}_2) = \text{PMI}(\text{word}_1, \text{word}_2) = \log_2\left(\frac{\frac{1}{N}\text{hits}(\text{word}_1 \text{ NEAR word}_2)}{\frac{1}{N}\text{hits}(\text{word}_1)\frac{1}{N}\text{hits}(\text{word}_2)}\right),$$

where N is the number of documents. The numerator of the PMI refers to the probability that $word_1$ and $word_2$ occur together and are thus semantically similar, while the denominator reflects the probability that these words occur independently.

Machine learning approaches in sentiment analysis make use of (a) traditional machine learning models, or (b) deep learning models to estimate the overall sentiment polarity of a document. *Traditional machine learning models* are related to machine learning techniques, such as the naïve Bayes classifier, maximum entropy classifier, or support vector machines (SVM). For traditional machine learning models, features are specified and extracted manually or by employing feature selection methods. Semantic, syntactic, stylistic, and Twitter-specific features can be used as the input to these algorithms [22]. In deep learning models, features are determined and extracted automatically.

Deep neural network (DNN) models are neural networks with multiple hidden layers. The most widely used learning algorithm to train a deep neural network model involves backpropagation based on gradient descent. In the first round, the weights are initialized on a random basis. Then, the weights are tuned to minimize the prediction error relying on gradient descent. The learning procedure consists of multiple consecutive forwards and backwards passes. In the forward pass, the input is forwarded through multiple nonlinear hidden layers and the computed output is compared with the actual output. Let X_i be the input and f_i be the nonlinear activation function for layer i, then the output of the layer I, which is also the input for layer $(i + 1)$, is given by

$$X_{i+1} = f_i(W_i X_i + b_i),$$

where W_i and b_i are the parameters between layers i and $(i - 1)$.

In the backward pass, the error derivatives with respect to the parameters are then back propagated, so that the parameters can be adjusted to minimize the prediction error:

$$W_{\text{new}} = W - \eta \partial E/\partial W, \text{and } b_{\text{new}} = b - \eta \partial E/\partial b,$$

where E is the cost function, and η is the learning rate. The overall process continues until a desired prediction improvement is reached [45].

In one of the recent surveys, analysis of 32 papers identified DNN, CNN, and hybrid approaches as the most commonly used models for sentiment analysis [24]. In a total of 112 deep learning-based papers on sentiment analysis published in 2020, the most commonly used deep learning algorithms were Long-Short Term Memory (LSTM) (36%), Convolutional Neural Networks (CNN) (33%), Gated Recurrent Units (GRU) (9%), and Recurrent Neural Networks (RNN) (8%) [14]. In comparison, CNN performed better than the other models in terms of both accuracy and CPU runtime. RNN mostly performed slightly better than CNN in terms of reliability, but required more computation time [24]. The deep neural network architecture of CNN usually consists of convolutional layers and pooling or subsampling layers, where convolutional layers extract features, while pooling or subsampling layers reduce their resolution. RNN's deep neural network architecture captures previous computations and reuses them in subsequent inputs. Long–short-term memory (LSTM) is a special type of RNN that uses long memory as input for activation features in the hidden layer [24].

Related Work on Sentiment Services

An earlier comparison of 15 free web services in terms of their accuracy on different text types [19] and three solutions—Alchemy, Text2data, and Semantria [16]—was completed in 2015. A comparison of 24 sentiment analysis methods based on 18 labeled data sets followed in 2016, evaluating several commercial sentiment analysis methods: LIWC (2007 and 2015), Semantria, SenticNet 3.0, Sentiment140, and SentiStrength [18]. Previously, eight sentiment analysis methods were compared in terms of coverage (i.e., the proportion of messages whose sentiment was identified) and agreement (i.e., the proportion of identified sentiments that agreed with the ground truth) [17]. Several (now older) analysis software solutions were tested on five different data sets in [15]. Independent and parallel studies to this research compare the accuracy of these services from four major cloud platforms—Amazon, Google, IBM, and Microsoft—with the bag-of-words approach [20] and explore the use of ensemble approaches based on the sentiment analysis services [21].

As far as we are aware, there are no other studies comparing recent developments and novel implementations of all these commercial services against a variety of established metrics, although they are used extensively in countless practical data science applications in industry.

Experimental Design

Data Set

In the experimental study we base on a real-world Twitter data set of 14,640 records related to the airline service quality retrieved from the publicly accessible kaggle.com platform,[1] also used in a comparative study of deep learning models in

[1] https://www.kaggle.com/crowdflower/twitter-airline-sentiment.

Table 1 Data set descriptions

Data set	Thematic category	Published	Number of tweets	Number of selected tweets	Positive, %	Negative, %	Neutral, %
Airlines data set	Airline	2019	14,640	13,519	16	63	21
Southwest Airline data set	Airline	2019	7064	7064	28	16	55
General tweets	General	2021	162,980	20,000	44	22	34

sentiment analysis [24] and comparable to the data sets used in other related studies [20, 21]. The data set included attributes, such as tweet ID, airline (the six largest U.S. airlines), polarity label, manually evaluated, i.e., positive, negative, neutral (see Table 1), confidence value for label, and publication date. When preparing the data set, the empty entries of each row were pre-processed for storage in the database. Afterwards, duplicates were removed based on the column of the tweet ID, the unique identifier of Twitter, what resulted in 14,639 left records. We further sorted out tweets that were annotated by humans with a confidence value of less than 0.65, annotated with the given class by almost more than two-thirds of the human classifiers. The final data represents the set of 13,519 tweets.

For further analysis, a similar data set with real-world Twitter data regarding airline service quality for a specific airline—Southwest Airlines[2] was selected, which consists of 7064 labelled tweets. The data set, consisting of airline service quality tweets for Southwest airlines, included following attributes: tweets, location, timestamp, sentiment (see Table 1), positive score, negative score, and neutral score.

In addition to this, we conduct an analysis on a data set of 162,980 general tweets[3] to compare the accuracy of services on specific and general data. The data set of general tweets included the least number of attributes: clean_text, category (see Table 1). 20,000 tweets were randomly selected with the same negative/neutral/positive ratio to ensure feasibility of evaluation.

Twitter data sets have been widely used in different sentiment analysis studies before [7, 18, 31, 46–49]. Tweets about service quality can provide valuable insights about consumer satisfaction and can be thus effective to infer firms' future earnings [7], their directional stock price movements [49], etc. The sentimental orientation of tweets requires special attention. Indeed, negative tweets enable more accurate forecasts than do positive tweets [7]. Neutral tweets are perceived as more helpful [50], lead to more neutral feedback [51], and also tend to be retweeted more often [46]. Sentimental reviews with positive sentiment polarity in their title receive more readership [50]. Sentiment-driven positive feedback generally leads to a superior level of online trust [52], knowledge reuse [53], willingness to share [54], and has substantial and sustainable impact [55].

Airlines are interested in using social media to establish online communities und involve their members into co-creating new solutions [56], however, hardly manage to respond even half of the tweets, as a relatively recent analysis of over three million complaining tweets related to seven major U.S. airlines on Twitter in the time period from September 2014 to May 2015 demonstrated [57].

Commercial Sentiment Analysis Solutions

The market for commercial sentiment analysis software includes many vendors of varying sizes. Our initial screening revealed Amazon Web Services Amazon Comprehend,[4] Dandelion Sentiment Analysis API,[5] Google Cloud Platform Natural Language API,[6] IBM Watson Natural Language Understanding,[7] Lexalytics Semantria API,[8] MeaningCloud Sentiment Analysis API,[9] Microsoft Azure Text Analytics,[10] ParallelDots Sentiment Analysis,[11] Repustate Sentiment Analysis,[12] Text2data Sentiment Analysis API,[13] TheySay PreCeive API,[14] and twinword Sentiment Analysis API.[15] Some sentiment analysis solutions such as AWS Amazon Comprehend, Google Cloud Platform Natural Language

[2] https://data.world/utmlfall2019/southwest-tweets/workspace/file?filename=southwest_tweet_sentiment.csv.

[3] https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset.

[4] https://aws.amazon.com/de/comprehend/.

[5] https://dandelion.eu/.

[6] https://cloud.google.com/natural-language.

[7] https://www.ibm.com/de-de/cloud/watson-natural-language-understanding.

[8] https://www.lexalytics.com/semantria.

[9] https://www.meaningcloud.com/products/sentiment-analysis.

[10] https://azure.microsoft.com/de-de/services/cognitive-services/text-analytics/.

[11] https://www.paralleldots.com/sentiment-analysis.

[12] https://www.repustate.com/sentiment-analysis/.

[13] https://text2data.com/sentiment-analysis-api.

[14] http://www.theysay.io/product/preceive/.

[15] https://www.twinword.com/api/.

API, and Microsoft Azure Text Analytics [20, 21], IBM Natural Language Understanding (NLU) [20, 21, 58], Lexalytics Semantria API [16, 18], and Text2data [16] were part of previous research.

Since the focus of this work is on commercial software, we first checked whether the solutions were fee-based. To enable this evaluation, we focused only on those that provided a free trial version with a sufficiently large quota. If no free contingent was offered or the volume of records exceeded the free contingent of a service, the total cost of a solution not exceeding the limit of 10 euros was still accepted. Therefore, the products of ParallelDots, Repustate, Text2data, Twinword and TheySay were excluded from further investigation in this study. Furthermore, Dandelion was excluded, because this solution only offers document-level analysis depth and does not enjoy higher visibility compared to Amazon Comprehend, which also only offers document-level sentiment analysis.

All solutions allow sentiment classification based on custom data sets and did not require configuration or training of models. They also offer a REST-compliant programming interface. This ensures that a company can integrate the product into its own applications as easily as possible. The programming interface can be run by the vendor in the cloud, so there is no need for the customer to have their own infrastructure. The functionality of the product, including the REST interface or client libraries, has been well-documented and publicly available. The solutions also enable communication via the encrypted HTTPS protocol, so that companies can also process personal or otherwise sensitive data.

Implementation

After selecting the six solutions mentioned above—Amazon Web Services (AWS) Amazon Comprehend, Google Cloud Platform Natural Language API, IBM Watson Natural Language Understanding (NLU), Microsoft Azure Text Analytics, Lexalytics Semantria API, and MeaningCloud Sentiment Analysis API—an analysis framework was designed and implemented in Python. First, a user account was created with each of the corresponding SaaS providers.

To store the JSON-like nested responses of the APIs, a document-oriented NoSQL MongoDB database was set up and hosted at the MongoDB Atlas cloud provider. For all database functions, the `DB_Manager` class, based on the pymongo library, was implemented to connect to the database at initialization and perform the necessary database queries to read, store, and modify data. For each of the sentiment analysis solutions, the functionality was implemented in separate modules using the client libraries. Each module included authentication and configuration of the service client, if required, as well as the `get_sentiment` method to request the respective service, get its response and extract the required information from the response object.

A `Benchmark` class was implemented to provide all the logic for querying each service, measuring the response time and associating each result with the data set using static methods. The data set to be processed was provided in the form of an object of the class `Tweet`. When passed to the `get_sentiment` method from the respective module, the response time was measured, and the result was assigned to the `Tweet` object. In the `Benchmark` module, the `get_tweet_sentiment` method also provided the ability to perform a per-service query for each tweet. This is then called for each tweet and stores the result in the database after getting each response from a service along with the response time.

However, only those services are requested for which there is not already a response in the `Tweet` object, e.g., from an earlier execution of the script. In the Tweet object, and thus also in the database, the complete response is stored with its respective nested structure. Although some providers also allow batch processing of a request, only one text per request is analyzed here for reasons of comparability of response times.

In all solutions with the synchronous programming interfaces, i.e., all except Lexalytics Semantria API, sequential processing of individual documents has been implemented. To reduce the processing time, parallel processing of multiple documents using multiprocessing was also implemented. However, since this also requires the pymongo client instance to be reinitialized for each process, as pymongo is not fork-safe, the maximum number of parallel processes was limited to four.

In the case of the Lexalytics Semantria API, asynchronous processing of the test data had to be performed. In the benchmark module, the `lexalytics_queue_tweets` method adds batches of five tweets to the Semantria API queue.

The batch size was set to five records for two reasons: on one hand, the processing time should be as close as possible to the time needed for one record to make the results comparable between services. On the other hand, testing revealed that the time required to receive the processed record is almost identical for a batch size of one record as it is for a batch size of five. Since this thread does not block the program flow, a polling thread can be started directly with the `lexalytics_polling` method. The `lexalytics_polling` method polls the API with four threads at random intervals between 0 and 100 ms for new processed documents until all documents added to the queue have been processed. If one or more batches have been returned in a query, they are processed further in batches of no more than 20 documents.

Table 2 Experimental settings

Solution	Target class			Version used	
	Positive	Negative	Neutral	API	Client library
Amazon Comprehend	Positive	Negative	Neutral, Mixed	September 28, 2020	1.16.1 BOTO3
Google Cloud Natural Language API	(0.25, 1]	[−1, −0,25)	[−0.25, 0.25]	1.2 (March 20, 2020)	2.0.0 { "google-cloud-language".)
IBM Watson NLU	Positive	Negative	Neutral	2020-08-01	4.7.1 (ibm-watson)
Microsoft Azure Text Analytics	Positive	Negative	Neutral, mixed	3.0	5.0.0 { “azure-ai-textanalytics”.)
Lexalytics Semantria API	Positive	Negative	Neutral	4.2 (6-4-2016)	4.2.92 { “semantria-sdk”.)
MeaningCloud Sentiment Analysis API	P+, P	N, N+	NEW, NONE	2.1 (10/September/2020)	2.0.0 (MeaningCloud-python)

Table 3 Experimental results in November 2020 (Study 1, Airlines data set)

Measure/provider	Amazon	Google	IBM	Microsoft	Lexalytics	Meaning cloud
Precision						
Positive	69.3	65.9	64.1	51.9	49.8	36.2
Neutral	39.9	41.4	65.3	34.2	29.2	32.5
Negative	94.4	89.5	87.1	91.1	91.6	90.1
Recall						
Positive	86.2	88.8	86.7	82	51.8	84.3
Neutral	77.4	48.2	44.9	59	77.6	50.3
Negative	61.4	77.3	87.9	57.7	43.9	45.5
F1						
Positive	76.9	75.7	73.7	63.6	50.8	50.6
Neutral	52.6	44.5	53.2	43.3	42.4	39.5
Negative	74.4	83	87.5	70.7	59.4	60.4
Macro F1	68	67.7	71.5	59.2	50.9	50.2
Accuracy	68.5	73.4	79.2	61.8	51.8	52.6
Response						
Mean	194	299	253	151	1321	1244
Median	165	194	243	139	1296	1200
Std. dev.	127	210	75	62	226	500
SLA	99.9131–99.9	99.9133–99.9	99.5134–99.9	99.9132–99.9	99.995	99.9136–99.9

This processing is done in separate threads—so as not to block the polling method—and involves calculating the response time and storing the results in the database. To ensure comparability of the solutions, the batch size was reduced.

The results of each solution were compared to the polarity labels of the annotated data sets (see Table 2). For IBM Watson NLU and Lexalytics Semantria API, the same classes were used as in the test data. For MeaningCloud Sentiment Analysis API, the labels for normal and strong positive and negative polarity were combined to positive and negative. In addition, absence of sentiment (NONE) and mixed sentiment (NEW) were combined to form the class neutral.

For Amazon and Azure, mixed sentiment was also translated to the neutral polarity class when there was no tendency for a positive or negative class. For Google, numeric values had to be translated into polarity classes. The class boundaries for the neutral class separating the positive from the negative class were chosen as −0.25 and +0.25, as recommended in the product demonstration.

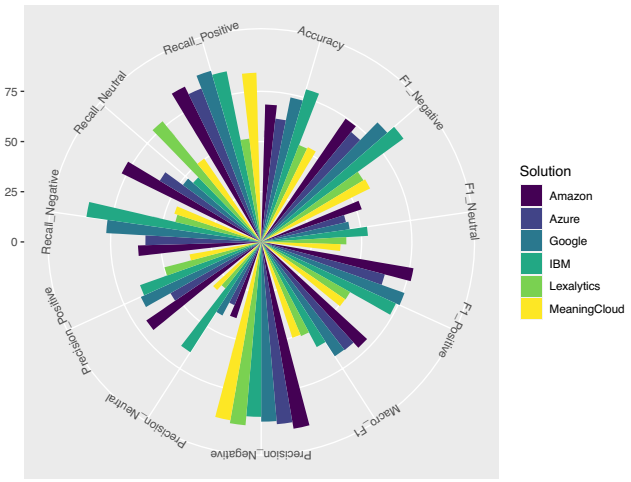

Fig. 1 Selected experimental results (polar coordinates)

Results

Sentiment analysis solutions were evaluated in terms of well-established measures, such as accuracy, precision, recall, (macro) *F*-score, [22, 23], SLAs, measured in percent, and time performance in milliseconds (ms).

With around 79% correctly classified samples, Watson NLU is the most accurate solution among the services tested (see Table 3 and Fig. 1). Only Google Cloud's service is close behind with 73.4% accurate classifications. Lexalytics Semantria API and MeaningCloud Sentiment Analysis API are the least accurate solutions, each classifying just over half of the texts correctly—51.8% and 52.6%, respectively.

For negative samples, all tested solutions showed quite high precision. The values range from 94.4% (Amazon Comprehend) to 87.1% (IBM Watson NLU). A more differentiated picture emerges for recall. With 88%, IBM Watson NLU has the highest recall. Only Google Cloud Natural Language API can also offer comparably high coverage with a recall of around 77%. AWS and Microsoft Azure services lag behind these solutions with 61.4% and 57.7% recall, respectively. Lexalytics Semantria API and MeaningCloud Sentiment Analysis API did not even achieve 50% recall. IBM Watson NLU achieved the best result among all solutions with an F1 score of 87.5%. Only Google Cloud Natural Language API could show a similarly high F1 value of 83%. The midfield is formed by AWS and Azure with F1 values of less than 75%. Lexalytics Semantria API and MeaningCloud Sentiment Analysis API are the least reliable solutions here.

Among the positive samples, the solutions from AWS, Google, and IBM are the most accurate solutions here, albeit with an accuracy of less than 70%. For Microsoft Azure Text Analytics and Lexalytics Semantria API, only every second positive classification was correct. MeaningCloud Sentiment Analysis API performed the worst, with an accuracy of only about 36%.

Still, almost all solutions correctly identified a similar proportion of texts as positive, with recall ranging from 89% (Google Cloud Natural Language API) to 82% (Microsoft Azure Text Analytics). Only Lexalytics Semantria API correctly classified just half of all positive texts with 52% recall. In terms of F1 score, Amazon Comprehend delivers the best result with 76.9%, closely followed by the solutions from Google and IBM with 75.7% and 73.7%, respectively. In the midfield is Microsoft Azure Text Analytics with 63.6%, while Lexalytics Semantria API and MeaningCloud Sentiment Analysis API close out the list with an F1 score of just over 50%.

For the neutral class, all solutions except IBM Watson NLU (65%) showed low precision values of below 40%. The worst precision of only 29% was shown by Lexalytics Semantria API. In terms of recall, only AWS and Lexalytics services achieved high coverage of around 77%. The next best result was achieved by Microsoft Azure Text Analytics with 59% recall. The remaining solutions have a recall of around 50% and below. In terms of F1 score, only AWS and IBM achieved F1 scores just above 50%. MeaningCloud Sentiment Analysis API remains below 40%.

While it took over 1200 ms on average to get a response from the solution, each of the major cloud providers only required an average response time of under 300 ms, with Microsoft Azure Text Analytics being the fastest solution in this study and Lexalytics Semantria API being the slowest. However, it should be noted that Lexalytics Semantria API provides an asynchronous programming interface and, therefore, requires two requests before the results of an analysis are available. Since many factors influence the API response time, including Internet connection and proximity to the server location, the evaluation of this criterion shows only a preliminary picture and is not necessarily representative. However, due to the large number of requests, the measurements of the individual solutions can be compared with each other, as they were all created under similar conditions. The response time is, therefore, only considered in relation to the other solutions and should not be regarded as an absolute value.

Moreover, the availability of IT systems and services is often contractually regulated in service level agreements (SLA). The agreed uptime is usually specified as a percentage and expresses the proportion of a period during which a system is to be available. In addition, when external services are used as building blocks for more advanced solutions, an analysis of the weakest links and mitigation of potentially cascading failures should be performed.

In the case of IBM Watson NLU, the (relatively) low uptime of 99.5134% is contractually guaranteed to customers on the standard tariff. This means that the solution can

Table 4 Experimental results in July 2022 (Study 2)

Measure/provider	Airlines data set		Southwest Airline data set		General tweets	
	Google	MeaningCloud	Google	MeaningCloud	Google	MeaningCloud
Precision						
Positive	65.9	37.5	68.2	57.5	68.5	60.3
Neutral	36.3	32	71.9	75.7	48.9	54
Negative	91.6	90.3	46.1	48.9	32.3	37.2
Recall						
Positive	88.7	84.2	81.6	84.0	27.5	55.1
Neutral	57.1	50.4	54.2	48.8	34.1	43.6
Negative	67.7	46.6	67.5	63.9	85.2	54.4
F1						
Positive	75.6	51.9	74.3	68.3	39.3	57.6
Neutral	44.4	39.1	61.8	59.4	40.2	48.3
Negative	77.9	61.5	55.8	55.4	46.9	44.2
Macro F1	66	50.8	63.6	61	42.1	50
Accuracy	68.9	53.3	65.2	62.6	42.6	51
Response						
Mean	209.78	1900.44	195.92	1902.61	218.74	1875
Median	178.61	1544.24	174.93	1562.05	185.17	1515.75
Std. Dev.	106.27	1283.95	109.6	1254.87	119.63	1265.87

be down for almost 44 h a year without contractual regulations taking effect. Only from the Premium tariff upwards is a higher monthly availability of 99.9% agreed in the SLAs. Customers of products from Amazon (99.9131%), Google (99.9133%), Microsoft Azure (99.9132%) and MeaningCloud (99.9136%) have to put up with around 9 h of downtime per year with an agreed uptime of 99.9%. Lexalytics promises an even higher monthly uptime of at least 99.995% at the time of this study.

Finally, to conduct a longitudinal study over time, the tests performed in November 2020 were rerun in July 2022 with two of the service providers on the same and two additional data sets. For the same data set as in Study 1 (Airlines data set), Google Cloud Natural Language API's accuracy interestingly decreased from 73.4% to 68.9%, while MeaningCloud Sentiment Analysis API's accuracy increased from 52.6 to 53.3% (see Table 4). We will provide more details on these results in the discussion section.

Google Cloud Natural Language API's results for positive samples did not change over time, while MeaningCloud Sentiment Analysis API showed higher precision of 37.5% (+1.3%) and lower recall of 84.2% (−0.1%) than in November 2020. Google Cloud Natural Language API shows large changes for neutral and negative samples. For neutral samples, precision decreased to 36.3% (−5.1%) and recall increased to 57.1% (+8.9%), and for negative samples, precision is 91.6 (+2.1%) and recall is 67.7 (−9.6%). For MeaningCloud Sentiment Analysis API, precision in recall improved only slightly on average: for neutral tweets, precision is 32% (−0.5%) and recall is 50.4% (+0.1%); for negative tweets, precision is 90.3 (+0.2%) and recall is 46.6% (+1.2%). The F1 score for neutral samples decreased slightly for both services: 44.4% (−0.1%) for Google Cloud Natural Language API and 39.1% (−0.4%) for MeaningCloud Sentiment Analysis API. However, for positive and negative samples, the F1 score for MeaningCloud Sentiment Analysis API increased to 51.9% (+1.3%) and 61.5% (+1.1%), respectively, while for Google Cloud Natural Language API it remained the same (75.6%) for positive samples and decreased to 77.9% (−5.1%) for negative samples.

For the second data set (Southwest Airline data set), which contained service quality records from only one airline, Google Cloud Natural Language API's accuracy was higher than MeaningCloud Sentiment Analysis API's at 65.2%, but lower than for the first data set. However, MeaningCloud Sentiment Analysis API's accuracy for this data set was significantly higher than for the previous data set: 62.6% compared to 53.3%. For some sample groups, MeaningCloud Sentiment Analysis API has higher accuracy than Google Cloud Natural Language API: 75.7% and 71.9% for neutral samples; 48.9% and 46.1% for negative samples. For positive samples, MeaningCloud Sentiment Analysis API also has a higher recall rate of 84%, compared to 81.6% for Google Cloud Natural Language API. However, Google Cloud Natural Language API has a higher F1 score in all cases.

For the third data set of general tweets, Google Cloud Natural Language API showed the worst accuracy among all the data sets analyzed. Its precision is 42.6%, while MeaningCloud Sentiment Analysis API's precision is 51%,

which is more in line with the results already obtained in other data sets. Google Cloud Natural Language API outperforms MeaningCloud Sentiment Analysis API in precision for positive samples (68.5% for Google Cloud Natural Language API and 60.3% for MeaningCloud Sentiment Analysis API), recall for negative samples (85.2% for Google Cloud Natural Language API and 54.4% for MeaningCloud Sentiment Analysis API), and F1 score for negative samples (46.9% for Google Cloud Natural Language API and 44.2% for MeaningCloud Sentiment Analysis API). In all other cases, MeaningCloud Sentiment Analysis API performs significantly better than Google Cloud Natural Language API in this data set.

Discussion

Watson NLU scored the highest for accuracy at 79%, followed closely by Google Cloud Natural Language API at 73%. Lexalytics Semantria API and MeaningCloud Sentiment Analysis API classified only slightly more than half of the texts correctly—52% and 53%, respectively, which is only slightly more accurate than guessing. Our results are consistent with previous measurements on a comparable data set [20], namely, Amazon Comprehend: 68.5% (overall: 72.7%, negative: 66.8%, neutral: 81.7%, positive: 92.2%); Google Cloud Natural Language API: 73.4% (overall: 74.1%, negative: 77.7%, neutral: 39.4%, positive: 91.8%); IBM Watson NLU: 79.2% (overall: 85.4%, negative: 91.2%, neutral: 52.0%, positive: 90.8%); Microsoft Azure Text Analytics: 61.8% (overall: 66.2%, negative: 68.6%, neutral: 31.3%, positive: 90.3%). On one hand, the results may point to still unresolved challenges in sentiment analysis technology, such as linguistic complications [59, 60], and in the case of social media content, the possible use of non-standard language (e.g., abbreviations, misspellings, emoticons, or multiple languages) [34, 61]. Nevertheless, researchers training different deep learning models on the same data set were able to achieve significantly higher accuracies, however, with only two classes—positive and negative [24]: based on TF-IDF DNN: 86%, CNN: 85%, and RNN: 83%; based on word embeddings DNN: 90%, CNN: 90%, and RNN: 90%.

For positive and neutral classifications, none of the solutions could achieve a precision value above 70%. However, for negative classifications, the results looked more favorable: Amazon Comprehend: 94%, Lexalytics Semantria API: 92%, Microsoft Azure Text Analytics: 91%, Google Cloud Natural Language API: 90%, MeaningCloud Sentiment Analysis API: 90%, and IBM Watson NLU: 87%. Researchers training different deep learning models on the same data set reduced to positive and negative classes [24] reported comparable accuracies as follows: based on TF-IDF DNN: 88%, CNN: 86%, and RNN: 84%; based on word embeddings DNN: 92%, CNN: 92%, and RNN: 93%.

All solutions except Lexalytics Semantria API showed high recognition rates for positive classifications at 82% and above. For neutral classifications, only AWS and Lexalytics achieved high recognition rates of about 77%. Watson NLU achieved the highest recall for negative classifications at 88%, followed closely by Google Cloud Natural Language API at 77%. Researchers training different deep learning models on the same data set with positive and negative classes [24] achieved significantly higher recalls: based on TF-IDF DNN: 96%, CNN: 97%, and RNN: 97%; based on word embeddings DNN: 96%, CNN: 96%, and RNN: 95%.

Compared to prior studies, Lexalytics Semantria API demonstrated quite mixed results, i.e., slightly lower, but still comparable accuracy of 51.8% (58.39% [16], and 61.54%, 68.89% [18]), rather strong precision of 91.6% (96.09% [16], and 39.57%, 49.82% [18]) and recall of 43.9% (37.31% [16], and 52.81%, 55.53% [18]) for negative classifications, rather weak precision of 49.8% (81.91% [16], and 67.28%, 48.86% [18]) and recall of 51.8% (82.23% [16], and 57.35%, 63.73% [18])) for positive classifications, rather weak precision of 29.2% (4.34% [16], and 65.98%, 82.02% [18]) and rather strong recall of 77.6% (43.28% [16], and 67.03%, 72.96% [18]) for neutral classifications.

Across all compared services, no solution could achieve an F1 score of more than 80% for all classes. In terms of the F metric, all models trained on the two class data set were more reliable [24]: based on TF-IDF DNN: 92%, CNN: 91%, and RNN: 90%; based on word embedding DNN: 94%, CNN: 94%, and RNN: 94%.

In terms of time performance, the major cloud providers required an average response time of less than 300 ms, with Microsoft Azure Text Analytics being the fastest: Amazon Comprehend: 0.194 s, Google Cloud Natural Language API: 0.299 s, IBM Watson NLU: 0.253 s, Microsoft Azure Text Analytics: 0.151 s, Lexalytics Semantria API: 1.321 s, and MeaningCloud Sentiment Analysis API: 1.244 s.

The response time of a solution can depend on a variety of factors, e.g., the distance and routing to the server used by an application programming interface, the bandwidth of the Internet connection. However, in the present study, they do not seem to explain the differences in time performance. Both Lexalytics Semantria API and MeaningCloud Sentiment Analysis API do not allow selection of server locations and do not appear to offer servers outside the US. AWS also only allows access to the "us-east-1" region in the U.S. in its academic version, but its solution is one of the best performing solutions in this study. The higher average response time for Lexalytics may also be due to the way it functions as an asynchronous interface. The previously mentioned experiments required more computation time: based on TF-IDF DNN: 1 min, CNN: 34.41 s, and RNN: 1 h 54 s; based on

Table 5 Number of samples with deviations of the new (July 2022) from the old (November 2020) sentiment

Provider	New sentiment is correct		Old sentiment was correct		Both options were incorrect		
	Old	New (correct)	Old (correct)	New	Old (incorrect)	New (incorrect)	Sentiment
Google							
Negative	239	0	835	0	16	0	0
Neutral	0	239	0	835	0	16	0
Positive	0	0	0	0	0	0	16
Total	239		835		16		
MeaningCloud							
Negative	123	2	23	3	2	2	179
Neutral	8	105	4	26	13	169	1
Positive	1	25	3	1	168	12	3
Total	132		30		183		

word embeddings DNN: 30.66 s, CNN: 1 min 22 s, and RNN: 2 min 41 s [24].

IBM Watson NLU and Google Cloud Natural Language API achieved the highest recall rates for negative classifications of 88% and 77%, respectively, and the highest F1 scores of 88% and 83%, respectively, and can, therefore, be preferred when the correct classification of negative text is the primary concern. Indeed, negative tweets allow more accurate predictions than positive tweets [7]. In addition, social media and rating websites in general are vulnerable to strategically driven abuse and manipulation, such as opinion spam and fake ratings [62]. Another possible strategy to mitigate reliability variability is the creation of ensemble models [21].

When re-enabled in July 2022 as part of our second study, the Google Cloud Natural Language API was still the clear winner in the airline data set compared to the MeaningCloud Sentiment Analysis API, but could no longer clearly compete in our other data sets. The MeaningCloud Sentiment Analysis API performed better than the Google Cloud Natural Language API in some cases, namely, precision for neutral and negative classifications and recall for positive classifications (including neutral in the general data set). Thus, in the general data set, Google Cloud Natural Language API actually achieved a lower F1 score for positive and neutral classifications than MeaningCloud Sentiment Analysis API. Nevertheless, Google Cloud Natural Language API remained a significantly better choice in terms of response times than MeaningCloud Sentiment Analysis API, measured on average.

In total, there were 1408 samples (negative: 80%; neutral: 18%; positive: 2%) for which a different assessment was determined after a longer period of time (see Table 5). Google Cloud Natural Language API has the largest number of such samples, and for most of them (77%) the previously determined sentiment was correct. Only for 22% of the entries did the change in sentiment lead to a correct result. However, there is only 1% of the samples, where both the new and the old tuning is wrong. MeaningCloud Sentiment Analysis API, on the other hand, failed to get the correct sentiment both times for 53% of the samples. In 38% of the samples, the new sentiment proved to be correct, and only in 9% of the samples was the sentiment changed to an incorrect result.

All examples where Google Cloud Natural Language API changed sentiment over time were originally classified as negative. Over the course of the further research, the sentiment for all of these examples was changed to neutral. The situation is similar for MeaningCloud Sentiment Analysis API: most of the samples that had negative or positive sentiment in November 2020 were classified as neutral after a period of time. However, 19% of the samples for which the correct sentiment was determined in the second test had a positive sentiment. It can also be noted that for the samples for which MeaningCloud Sentiment Analysis API determined an incorrect sentiment both times, the results are more accurate, since for most of them the sentiment was shifted from positive to neutral, although in reality it is negative.

Our study involves some limitations and could be continued in several dimensions to mitigate them: first, though we extended the scope in July 2022, even further and possibly much more heterogeneous data sets could be analyzed with the selected services to provide results for text corpora in English, but also languages other than English [24, 30, 63].

Second, the set of selected sentiment analysis services could be expanded to provide even broader market coverage, and other solutions that do not fit the current selection criteria [64], due to the present study's focus on commercial services could be considered, such as Dandelion, ParallelDots, Repustate, Text2data, TheySay, and twinword. The reasons for these differences should also be investigated. Indeed, experiments show that higher accuracies in sentiment classification can be achieved by selecting appropriate

features and representations [24, 31]. The study by Gao et al. [16] reports that the time efficiency of Text2data is too low for these purposes.

Third, this study only represents the development status of the solutions in November 2020 and July 2022 and may be updated in the future as the reliability of the solutions may change. The software scripts developed for this study, which form a modular open-source software framework that flexibly supports such analyses, could be further developed to allow easy expansion with new data sets and additional sentiment analysis services to support informed service selection.

Fourth, additional criteria can also be used to evaluate these solutions. For example, with 250,000 texts to be analyzed, IBM's use of sentiment recognition costs more than 2.5 times as much as Google's ($660 versus $249.5).

In addition, the offer and quality of further text analysis functions, e.g., availability and/or speech recognition, can also be taken into account. All solutions support at least ten different languages for sentiment recognition. However, not all of them recognize the language automatically.

Conclusion

In this paper, current commercial SaaS solutions for sentiment analysis of different market power were investigated and compared. The results show that IBM Watson NLU and Google Cloud Natural Language API solutions can be preferred when negative text detection is the main focus. For negative classifications, all of them demonstrate precision of around 90%; however, only IBM Watson NLU and Google Cloud Natural Language API achieve recall of over 70%. In other cases, all solutions might have some weaknesses, especially Lexalytics Semantria API and MeaningCloud Sentiment Analysis API. For positive and neutral classifications, none of the solutions showed precision of over 70%.

When tested in July 2022, the Google Cloud Natural Language API was still the clear winner in the Airline data set compared to the MeaningCloud Sentiment Analysis API. Based on our other data sets, this could no longer be clearly claimed. The MeaningCloud Sentiment Analysis API performed better than the Google Cloud Natural Language API in some cases, namely, precision for neutral and negative classifications and recall for positive classifications (including neutral in the general data set). In the general data set, Google Cloud Natural Language API achieved a lower F1 score for positive and neutral classifications than MeaningCloud Sentiment Analysis API. Overall, Google Cloud Natural Language API nevertheless responds significantly faster than MeaningCloud Sentiment Analysis API, as measured by the average.

The work envisages several further research avenues. Additional and heterogeneous data sets can be analyzed with the selected services. Other services can be considered that could not be included in this study. The measurements made refer to the status of the solutions as of November 2020 and July 2022 and may be updated again in the future. Other criteria for evaluating these solutions may also be used, such as cost and availability.

Overall, our study shows that an independent, critical experimental analysis of sentiment analysis services can provide interesting insights into their overall reliability and particular classification accuracy beyond marketing claims, and that it is possible to critically compare solutions based on actual data and analyze potential shortcomings and margins of error before making an investment.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability This study uses only publicly available datasets referenced in the article.

Declarations

Conflict of interest All authors declare that they have no conflict of interests. On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Liu B. Sentiment analysis: mining opinions, sentiments, and emotions (studies in natural language processing). Cambridge: Cambridge University Press; 2015.
2. Wiegand M, Balahur A, Roth B, Klakow D, Montoyo A. A survey on the role of negation in sentiment analysis. In: Proceedings of the workshop on negation and speculation in natural language processing, pp. 60–68. Uppsala: University of Antwerp (2010).
3. Lau R, Liao S, Wong KF, Chiu D. Web 2.0 environmental scanning and adaptive decision support for business mergers and acquisitions. Manag Inf Syst Quart. 2012;36:1239–68.

4. Hu T, Tripathi A. The effect of social media on market liquidity. ICIS 2015 Proceedings (2015).
5. Jiang C, Wang J, Tang Q, Lyu X. Investigating the effects of dimension-specific sentiments on product sales: the perspective of sentiment preferences. J Assoc Inf Syst. 2021. https://doi.org/10.17705/1jais.00668.
6. Lin Z, Goh K. Measuring the business value of online social media content for marketers. ICIS 2011 Proceedings (2011).
7. Ho SY, Choi K, Yang F. (Finn): harnessing aspect—based sentiment analysis: how are tweets associated with forecast accuracy? J Assoc Inf Syst. 2019. https://doi.org/10.17705/1jais.00564.
8. Luo X, Gu B, Zhang J, Phang CW. Expert blogs and consumer perceptions of competing brands. Manag Inf Syst Q. 2017;41:371–95.
9. Chung S, Animesh A, Han K. Customer attitude from social media, customer satisfaction index, and firm value. ICIS 2017 Proceedings (2017).
10. Kim K, Lee S-YT, Benyoucef M. The impact of social sentiment on firm performance similarity. ICIS 2017 Proceedings (2017).
11. Ermakova T, Blume J, Fabian B, Fomenko E, Berlin M, Hauswirth M. Beyond the hype: why do data-driven projects fail? In: Proceedings of the 54th Hawaii international conference on system sciences (2021).
12. Yadav A, Vishwakarma DK. Sentiment analysis using deep learning architectures: a review. Artif Intell Rev. 2020;53:4335–85. https://doi.org/10.1007/s10462-019-09794-5.
13. Mell P, Grance T. The NIST definition of cloud computing. Natl Inst Standards Technol. 2011. https://doi.org/10.6028/NIST.SP.800-145.
14. Ligthart A, Catal C, Tekinerdogan B. Systematic reviews in sentiment analysis: a tertiary study. Artif Intell Rev. 2021. https://doi.org/10.1007/s10462-021-09973-3.
15. Abbasi A, Hassan A, Dhar M. Benchmarking twitter sentiment analysis tools. In: Proceedings of the Ninth international conference on language resources and evaluation (LREC'14), pp. 823–829. European Language Resources Association (ELRA), Reykjavik, Iceland (2014).
16. Gao S, Jinxing H, Fu Y. The application and comparison of web services for sentiment analysis in tourism. In: 2015 12th international conference on service systems and service management (ICSSSM), pp. 1–6 (2015). https://doi.org/10.1109/ICSSSM.2015.7170341.
17. Gonçalves P, Araújo M, Benevenuto F, Cha M. Comparing and combining sentiment analysis methods. In: Proceedings of the first ACM conference on Online social networks, pp. 27–38. Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2512938.2512951.
18. Ribeiro FN, Araújo M, Gonçalves P, Benevenuto F, Gonçalves MA. SentiBench—a benchmark comparison of state-of-the-practice sentiment analysis methods. arXiv:1512.01818 [cs]. (2016).
19. Serrano-Guerrero J, Olivas JA, Romero FP, Herrera-Viedma E. Sentiment analysis: a review and comparative analysis of web services. Inf Sci. 2015;311:18–38. https://doi.org/10.1016/j.ins.2015.03.040.
20. Carvalho A, Harris L. Off-the-shelf technologies for sentiment analysis of social media data: two empirical studies. AMCIS 2020 proceedings (2020).
21. Carvalho A, Xu J. Studies on the accuracy of ensembles of cloud-based technologies for sentiment analysis. AMCIS 2021 proceedings (2021).
22. Giachanou A, Crestani F. Like it or not: a survey of twitter sentiment analysis methods. ACM Comput Surv. 2016;49:28:1-28:41. https://doi.org/10.1145/2938640.
23. Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D. Text classification algorithms: a survey. Information. 2019;10:150. https://doi.org/10.3390/info10040150.
24. Dang NC, Moreno-García MN, De la Prieta F. Sentiment analysis based on deep learning: a comparative study. Electronics. 2020;9:483. https://doi.org/10.3390/electronics9030483.
25. O'Connor B, Balasubramanyan R, Smith NA. From tweets to polls: Linking text sentiment to public opinion time series. In: Fourth international AAAI conference on weblogs and social media (2010).
26. Arunachalam R, Sarkar S. The new eye of government: citizen sentiment analysis in social media. In: Proceedings of the IJCNLP 2013 workshop on natural language processing for social media (SocialNLP). pp. 23–28. Asian Federation of Natural Language Processing, Nagoya, Japan (2013).
27. Kauffmann E, Peral J, Gil D, Ferrández A, Sellers R, Mora H. Managing marketing decision-making with sentiment analysis: an evaluation of the main product features using text data mining. Sustainability. 2019;11:4235. https://doi.org/10.3390/su11154235.
28. Ghiassi M, Skinner J, Zimbra D. Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network. Expert Syst Appl. 2013;40:6266–82. https://doi.org/10.1016/j.eswa.2013.05.057.
29. Wankhade M, Rao ACS, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev. 2022;55:5731–80. https://doi.org/10.1007/s10462-022-10144-1.
30. Yadollahi A, Shahraki AG, Zaiane OR. Current state of text sentiment analysis from opinion to emotion mining. ACM Comput Surv. 2017;50:25:1-25:33. https://doi.org/10.1145/3057270.
31. Krouska A, Troussas C, Virvou M. The effect of preprocessing techniques on Twitter sentiment analysis. In: 2016 7th international conference on information, intelligence, systems applications (IISA), pp. 1–5 (2016). https://doi.org/10.1109/IISA.2016.7785373.
32. Troussas C, Virvou M, Espinosa KJ, Llaguno K, Caro J. Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning. In: IISA 2013, pp. 1–6 (2013). https://doi.org/10.1109/IISA.2013.6623713.
33. Li G, Zheng Q, Zhang L, Guo S, Niu L. Sentiment information based model for Chinese text sentiment analysis. In: 2020 IEEE 3rd international conference on automation, electronics and electrical engineering (AUTEEE), pp. 366–371 (2020). https://doi.org/10.1109/AUTEEE50969.2020.9315668.
34. Silva NFFD, Coletta LFS, Hruschka ER. A survey and comparative study of tweet sentiment analysis via semi-supervised learning. ACM Comput Surv. 2016;49:151–1526. https://doi.org/10.1145/2932708.
35. Darwich M, Mohd Noah SA, Omar N, Osman N. Corpus-based techniques for sentiment lexicon generation: a review. J Digital Inf Manag. 2019;17:296. https://doi.org/10.6025/jdim/2019/17/5/296-305.
36. Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. Comput Linguist. 2009;35:399–433. https://doi.org/10.1162/coli.08-012-R1-06-90.
37. Mohammad S. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers). pp. 174–184. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1017.
38. Mohammad SM, Turney PD. Crowdsourcing a word-emotion association Lexicon. Comput Intell. 2013;29:436–65. https://doi.org/10.1111/j.1467-8640.2012.00460.x.
39. Mohammad S. Word affect intensities. In: Proceedings of the eleventh international conference on language resources and

evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (2018).
40. Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the seventh international conference on language resources and evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (2010).
41. Cambria E, Hussain A. SenticNet. In: Cambria E, Hussain A, editors. Sentic computing: a common-sense-based framework for concept-level sentiment analysis. Cham: Springer International Publishing; 2015. p. 23–71. https://doi.org/10.1007/978-3-319-23654-4_2.
42. Strapparava C, Valitutti A. WordNet affect: an Affective Extension of WordNet. In: Proceedings of the fourth international conference on language resources and evaluation (LREC'04). European Language Resources Association (ELRA), Lisbon, Portugal (2004).
43. Jurafsky D, Martin JH. Speech and language processing. International. Upper Saddle River: Prentice Hall; 2008.
44. Turney PD, Littman ML. Measuring praise and criticism: inference of semantic orientation from association. ACM Trans Inf Syst. 2003;21:315–46. https://doi.org/10.1145/944012.944013.
45. Sengupta S, Basak S, Saikia P, Paul S, Tsalavoutis V, Atiah F, Ravi V, Peters A. A review of deep learning with special emphasis on architectures, applications and recent trends. Knowl Based Syst. 2020;194: 105596. https://doi.org/10.1016/j.knosys.2020.105596.
46. Bachura E, Valecha R, Chen R, Rao HR. Data breaches and the individual: an exploratory study of the OPM hack. ICIS 2017 proceedings (2017).
47. Chung W, He S, Zeng D. eMood: modeling emotion for social media analytics on Ebola disease outbreak. ICIS 2015 proceedings (2015).
48. Li B, Chong A. What influences the dissemination of online Rumor messages: message features and topic-congruence. ICIS 2019 Proceedings (2019).
49. Zhang W, Lau R. The design of a network-based model for business performance prediction. ICIS 2013 proceedings (2013).
50. Salehan M, Kim D. Predicting the performance of online consumer reviews: a sentiment mining approach. ICIS 2014 proceedings (2014).
51. Deng Y, Khern-am-nuai W. The value of editorial reviews for UGC platform. ICIS 2019 Proceedings (2019).
52. Grigore M, Rosenkranz C. Increasing the willingness to collaborate online: an analysis of sentiment-driven interactions in peer content production. ICIS 2011 proceedings (2011).
53. Grigore M, Rosenkranz C, Sutanto J. The impact of sentiment-driven feedback on knowledge reuse in online communities. AIS Trans Hum Comput Interact. 2015;7:212–32.
54. Lin Y-W, Ahsen ME, Shaw M, Seshadri S. The impacts of patients' sentiment trajectory features on their willingness to share in online support groups. ICIS 2019 proceedings (2019).
55. Beduè P, Förster M, Klier M, Zepf K. Getting to the heart of groups—analyzing social support and sentiment in online peer groups. ICIS 2020 proceedings (2020).
56. Jarvenpaa S, Tuunainen V. How finnair socialized customers for service co-creation with social media. MIS quarterly executive. 2013;12.
57. Gunarathne P, Rui H, Seidmann A. Customer service on social media: the effect of customer popularity and sentiment on airline response. ICIS 2014 proceedings (2014).
58. Carvalho A, Levitt A, Levitt S, Khaddam E, Benamati J. Off-the-shelf artificial intelligence technologies for sentiment and emotion analysis: a tutorial on using IBM natural language processing. Commun Assoc Inf Syst. 2019. https://doi.org/10.17705/1CAIS.04443.
59. Do HH, Prasad PWC, Maag A, Alsadoon A. Deep learning for aspect-based sentiment analysis: a comparative review. Expert Syst Appl. 2019;118:272–99. https://doi.org/10.1016/j.eswa.2018.10.003.
60. Minaee S, Azimi E, Abdolrashidi A. Deep-sentiment: sentiment analysis using ensemble of CNN and Bi-LSTM Models. arXiv:1904.04206 [cs, stat]. (2019).
61. Fan S, Ilk N, Zhang K. Sentiment analysis in social media platforms: the contribution of social relationships. ICIS 2015 proceedings (2015).
62. Lee S-Y, Qiu L, Whinston A. Manipulation: online Platforms' inescapable fate. ICIS 2014 proceedings (2014).
63. Habimana O, Li Y, Li R, Gu X, Yu G. Sentiment analysis using deep learning approaches: an overview. Sci China Inf Sci. 2019;63: 111102. https://doi.org/10.1007/S11432-018-9941-6.
64. Geske F, Hofmann P, Lämmermann L, Schlatt V, Urbach N. Gateways to artificial intelligence: developing a taxonomy for AI service platforms. ECIS 2021 research papers (2021).