Off-the-Shelf Technologies for Sentiment Analysis of Social Media Data: Two Empirical Studies

Completed Research

Arthur Carvalho
Farmer School of Business
Miami University
arthur.carvalho@miamioh.edu

Lucas Harris
Farmer School of Business
Miami University
harrislr@miamioh.edu

Abstract

Off-the-shelf technologies provided by major cloud platforms promise to facilitate and democratize the use of artificial intelligence techniques. Organizations can now apply highly sophisticated, pre-trained models in a variety of situations, such as when analyzing the sentiment behind social media posts. Among other uses, this enables organizations to better understand their consumers' opinions regarding products and/or services. In this paper, we first review technologies for sentiment analysis provided by major cloud platforms. We then compare the accuracy of these technologies against a technique widely used in managerial and information systems studies, namely the bag-of-words approach. Our two empirical studies use social media data collected from Twitter (short posts) and Facebook (long posts). We find that all the studied off-the-shelf technologies for sentiment analysis are vastly more accurate than the bag-of-words approach. We conclude the paper by discussing our results in light of the recent rise of low/no-code development practices.

Keywords

Sentiment analysis, artificial intelligence, social media analytics, software as a service.

Introduction

Recent years have seen tremendous growth in interest in artificial intelligence (AI), with application domains ranging from playful settings, such as soccer robot (Oliveira et al., 2009; Carvalho and Oliveira, 2011) and artificial poker players (Brown and Sandholm, 2018), to energy trading (Babic et al., 2017) and life-saving scenarios, such as rescue robots (Davids, 2002) and the prediction of heart transplant outcomes (Dag et al., 2016). Although not consensual, AI can be defined as "the study of agents that receive percepts from the environment and perform actions" (Russell and Norvig, 2016). Agents can be from very basic scripts written in any computational language to fully autonomous robots. This agent-centric perspective brings together several research areas under a common framework. For example, machine learning is about how agents can learn in/from an environment, knowledge representation is about how agents represent their knowledge, and natural language processing (NLP) is about how agents can communicate and understand each other and human beings. It is this latter area that we focus on in this paper.

One of the primary focuses of NLP is to acquire information from written language (Russell and Norvig, 2016). This is a challenging task given the nature of the underlying data, i.e., textual data might be ambiguous, contain typos, grammar mistakes, jargon, slangs, etc. Using computational, statistical, and linguistic techniques, there are many types of textual analysis one can perform, e.g., text categorization, text clustering, automatic summarization, topic modeling, and sentiment analysis. We concentrate on sentiment analysis, with a particular focus on the analysis of social media data.

Sentiment analysis aims at understanding and quantifying the sentiment behind a text (Liu, 2012), e.g., it is used when one wants to determine whether a certain piece of text is positive, negative, or neutral. This,

in turn, enables organizations to understand, among other things, customer opinions regarding products or services. Posts on social media are a particularly interesting data source to apply sentiment analysis to given the close connection and communication channel organizations have with (potential) consumers. However, applying sentiment analysis and, broadly speaking, NLP techniques can be a challenging endeavor given the required expertise in several fields. As a consequence, small and midsize organizations might not have the resources to afford the use of those techniques. This obstacle has been gradually removed by modern tools that allow one to perform a range of different data analyses using pre-trained models that require no machine learning or NLP expertise. Instead, graphical tools and application programming interfaces (APIs) enable developers to create applications that rely on external AI technologies that require a few (or even no) lines of programming code. We call these external, ready-to-use technologies *off-the-shelf technologies*.

Several cloud platforms currently provide off-the-shelf AI technologies, e.g., IBM Cloud and its Watson family of services and Microsoft Azure and its collection of technologies called Cognitive Services. As we elaborate on later in this paper, some of these services charge a fraction of a penny to perform sentiment analysis on a single document, which effectively helps to democratize the use of AI. The rise of off-the-shelf AI technologies raises several managerial and technical research questions, e.g., how to integrate these technologies into different business processes, products, and services? How cost-effective are these technologies? And how accurate are these technologies? It is this last question that we address in this paper. In particular, our contributions are threefold. First, we review the sentiment analysis services provided by four major cloud platforms, namely IBM Cloud, Amazon Web Services, Microsoft Azure, and Google Cloud, in terms of offered features and pricing schemes. We next report two studies that measure the accuracy of the sentiment analysis services provided by the above cloud platforms. Our experiments are based on data collected from two social media platforms, namely Twitter, which represents scenarios involving short texts, and Facebook, which represents scenarios involving longer texts. Although there is no single off-the-shelf technology that always outperforms the others, we nonetheless find that all the studied off-the-shelf technologies outperform the bag-of-words approach, which is a rather basic, but still popular technique in information systems studies. After discussing many other findings, we conclude the paper by linking our results and research to a growing software development practice called low/no-code development.

Research Background

In what follows, we elaborate on a few sentiment analysis techniques and review the main features and pricing scheme of sentiment analysis services provided by major cloud platforms.

Sentiment Analysis

There are different ways to computationally estimate the sentiment behind a text. A very popular approach is to assign scores based on the polarity of individual words. For example, consider the following Facebook post, which is part of the data set we used in our experiments: "*teavana truly sucks*." Using a list of positive and negative words (also called a *dictionary* or a *lexicon*), one could then, for example, assign the value zero to "teavana" and "truly" since these words are not present in any list of words, and the value -1 to "sucks" since this word is present in the negative list. After scoring individual words, one way of assigning a single score to the whole document is by simply summing all the individual scores. If the aggregate score is less than zero (respectively, greater than zero), then the sentiment behind the document is estimated to be negative (respectively, positive). A final score equal to zero means that the sentiment behind the document is neutral. In our specific example, the sentiment score behind the above Facebook post is -1, meaning that the post is negative. We henceforth refer to this technique as the *bag-of-words* approach.

One can think of the bag-of-words approach as a technique that places all the words of a document in a bag, and then selects and scores one word at a time. The notion of order is lost is a sense that this approach returns the same aggregate score regardless of the word order in a document. In this paper, we use the bag-of-words approach as a baseline against which we compare off-the-shelf technologies for sentiment analysis. We argue that the bag-of-words approach is an acceptable baseline because it has been often used in managerial and information systems research (see for example the work by Fan et al., 2015; Mai et al., 2018; Cao and Rhue, 2019; Pentland et al., 2019), although we acknowledge that ignoring syntax and other textual cues make this technique rather naïve in nature. There are many other ways of calculating sentiment scores (e.g., see the work by Lowe et al., 2011). For instance, instead of descriptive techniques based on the

Technology	Features	Pricing Scheme (Per Month)
IBM Watson Natural Language Processing	Extract Entities Categorize text Extract Concepts Extract Keywords Provide Emotion Scores Provide Sentiment Scores Perform Syntax Analysis	• First 30,000 NLU items are free • Tier 1: US$0.003 per NLU item for the first 250,000 NLU items • Tier 2: US$0.001 per NLU item from 250,001 to 5,000,000 NLU items • Tier 3: US$0.0002 per NLU item for over 5,000,000 NLU items
Amazon Comprehend	Extract Entities Extract Keyphrases Provide Sentiment Scores Perform Syntax Analysis	• First 50,000 units are free • Tier 1: US$0.0001 per unit for the first 10 million units • Tier 2: US$0.00005 per unit from 10 million to 50 million units • Tier 3: US$0.000025 per unit for over 50 million units
Microsoft Text Analytics	Extract Entities Extract Keyphrases Provide Sentiment Scores	• Free tier: 5,000 free transactions • S tier: Up to 0.5M transactions: US$2 per 1,000 transactions 0.5M - 2.5M transactions: US$1 per 1,000 transactions 2.5M - 10M transactions: US$0.5 per 1,000 transactions Over 10M: $0.25 per 1,000 transactions • S0 tier: US$74.71 for up to 25,000 transactions • S1 tier: US$249.86 for up to 100,000 transactions • S2 tier: US$999.75 for up to 500,000 transactions • S3 tier: US$2,499.84 for up to 2,500,000 transactions • S4 tier: US$4,999.99 for up to 10,000,000 transactions
Google Natural Language	Extract Entities Categorize text Provide Sentiment Scores Perform Syntax Analysis	• No cost for the first 5,000 units • US$1.00 per 1,000 units from 5,000 to 1 million units • US$0.50 per 1,000 units from 1 million to 5 million units • US$0.25 per 1,000 units from 5 million to 20 million units

Table 1. Summary of the Off-the-Shelf Technologies for Sentiment Analysis.

polarity of words, another approach is to build statistical models capable of estimating the sentiment behind a document. In detail, after preprocessing a text corpus, one can then train a statistical model on the underlying data as long as all the documents are appropriately labeled. The many design choices, such as different preprocessing techniques and/or the machine learning algorithms used for model training, make this modeling-based approach a poor choice for a baseline method in our experiments. For a discussion on other sentiment analysis techniques, we refer the interested reader to the work by Medhat et al. (2014).

Off-the-Shelf Technologies for Sentiment Analysis

We describe next the off-the-shelf technologies for sentiment analysis we study in this paper. For each technology, we elaborate on the offered features and pricing scheme. Table 1 summarizes the technologies. It is noteworthy that we do not attempt to explain how these technologies work since they are proprietary software. In other words, they are called *black boxes* since neither information about their inner workings nor the data they were trained on are available to the public. We return to this point later in the paper.

IBM Watson Natural Language Understanding

The first sentiment analysis technology we describe in this section is called *Natural Language Understanding* [1] (NLU), which is part of the IBM Watson family of services (Ferrucci et al., 2013;

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

Vroegindeweij and Carvalho, 2019). NLU has been successfully used in a variety of applications, ranging from analyzing social media posts regarding cryptocurrencies (Jerdack et al., 2018) to consumer opinions on corporate-social-responsibility policies (Carvalho et al., 2019). Besides providing a sentiment score ranging from -1 (negative) to +1 (positive), NLU is also able to detect entities (e.g., people, places, and events) and the relationships among them in a text. Moreover, it can identify concepts (high-level themes), extract emotions, keywords, and parse semantic roles. NLU fully or partially supports a total of 13 languages. NLU's pricing scheme is defined in terms of NLU units. Each NLU unit is defined as the number of data units times the number of requested features, i.e., sentiment scores, emotion scores, etc. One data unit is equal to 10,000 characters or less. For illustration, extracting emotion and sentiment scores from 15,000 characters requires 2 data units times 2 features, which is equal to 4 NLU items. Currently, NLU offers the first 30,000 NLU items per month for free. Thereafter, the following tiered model is applied: US$0.003 per NLU item for the first 250,000 NLU items in a month; US$0.001 per NLU item after reaching 250,001 until 5,000,000 NLU items in a month; US$0.0002 per NLU item after reaching 5,000,001 items in a month.

Amazon Comprehend

Amazon Comprehend[2] is a natural language processing service provided by Amazon as part of its cloud platform Amazon Web Services (AWS). Besides returning the sentiment behind a text ("positive", "neutral", "negative", or "mixed") alongside a confidence score, Amazon Comprehend can also provide syntax analysis and extract key entities and phrases. Amazon Comprehend supports a total of 6 languages. Requests for the above features are measured in units of 100 characters. On a monthly basis, the analysis of the first 50,000 units (5 million characters) using any of the above features is free. Thereafter, the following tiered model is used: US$0.0001 per unit for up to 10 million units in a month; US$0.00005 per unit from 10 million to 50 million units a month; $0.000025 per unit when over 50 million units a month.

Microsoft Text Analytics

Microsoft Text Analytics[3] is an off-the-shelf technology for natural language processing that is part of the Cognitive Services family inside the Microsoft Azure cloud platform. The sentiment analysis feature returns scores between 0 and 1, where scores close to 0 (respectively, 1) indicate a negative (respectively, positive) sentiment. Besides sentiment analysis, Text Analytics is also capable of keyphrase extraction and entity detection, and it supports 5 different languages. Similar to NLU, usage in Text Analytics is defined in terms of documents (up to 5,120 characters each) and requested features. In particular, each requested feature for a single document determines one transaction. In terms of pricing, Text Analytics defines a series of tiers users can choose from at a fixed monthly price. The free tier allows for 5,000 transactions per month. The S0, S1, S2, S3, and S4 tiers cost, respectively, US$74.71, US$249.86, US$999.75, US$2,499.84, and US$4,999.99 per month, and they allow for up to, respectively, 25,000, 100,000, 500,000, 2,500,000, and 10,000,000 transactions per month. If the usage on any tier other than the free tier is exceeded, the user then starts to accrue overages. Users can also sign for the S tier, where one is billed only for the number of transactions submitted to the service. In this tier, a transaction can have up to 1,000 characters, and users are charged as follows: US$2 per 1,000 transactions for the first 500,000 transactions; US$1 per 1,000 transactions from 500,000 to 2.5 million transactions; US$0.5 per 1,000 transactions from 2.5 million to 10 million transactions; and US$0.25 per 1,000 transactions when over 10 million transactions. It is noteworthy that we used version 2.1 of Microsoft Text Analytics in our experiments since the latest version 3.0 was still in the testing phase at the time when we conducted our experiments.

Google Natural Language

Google Natural Language[4] is Google Cloud platform's natural language processing service. The Natural Language's sentiment analysis feature associates a score from -1 (negative) to +1 (positive) with each analyzed text. Other features include syntax analysis, the extraction of entities, content classification, and support to 10 different languages. When it comes to pricing, Natural Language charges a user per unit,

[2] https://aws.amazon.com/comprehend
[3] https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/
[4] https://cloud.google.com/natural-language

which is equal to 1,000 characters. Different features have different pricing schemes. Focusing only on the sentiment analysis feature, the monthly price per 1,000 units is defined as follows: no cost for the first 5,000 units; US$1.00 from 5,000 to 1 million units; US$0.50 from 1 million to 5 million units; and US$0.25 from 5 million to 20 million units.

Empirical Studies

To test the accuracy of off-the-shelf technologies for sentiment analysis, we conducted two different studies using social media data. Social media enables organizations to effectively engage in organization-customer dialog. Moreover, it creates mechanisms for customer-customer dialog. Perhaps more importantly, social media allows organizations to monitor and mediate those dialogues. The new communication channels between companies and consumers generate precious data that are often highly unstructured in nature, such as comments in natural language, images, and videos. Analyzing these data, a practice now called *social media analytics* (Fan and Gordon, 2014), empowers organizations to better understand (potential) consumers' opinions towards past, current, and even future products and services. It also enables one to understand in real-time the current sentiment of a relevant population towards an organization and its managerial practices, such as corporate-social-responsibility policies (Carvalho et al., 2019).

In our studies, we collected data from two different social media platforms, namely Twitter and Facebook. An independent set of individuals categorized each post in our data sets as "positive", "neutral", or "negative". We elaborate on this labeling process in future subsections. Given that some off-the-shelf technologies do not always return similar sentiment labels, we had to preprocess and standardize the outputs from each of those technologies before analyzing their accuracy. Specifically, for IBM NLU and Google Natural Language, we considered all documents with sentiment scores less than, equal to, and greater than zero to be, respectively, "negative", "neutral", and "positive". We adapted a similar scheme to Microsoft Text Analytics' scoring range where sentiment scores less than, equal to, and greater than 0.5 were classified as, respectively, "negative", "neutral", and "positive". Besides the "negative", "neutral", and "positive" labels, Amazon Comprehend can also return a "mixed" label when it is deemed that no sentiment dominates an entire text. Given that this label is not assigned by the other technologies, we removed from our data sets all the documents classified as "mixed" by Amazon Comprehend. Besides those, we also removed documents that received no score/label from at least one of the studied technologies due to, for example, being too short. Finally, for the bag-of-words approach, we used the dictionary of positive and negative words provided by Hu and Liu (2014). To calculate the accuracy of each method, we calculated how often the proposed labels matched the sentiments assigned by a group of humans. In other words, for each data set, each method received an accuracy value between 0% and 100% representing how accurate (i.e., close to human judgment) that method is. Besides reporting the overall accuracy, we also report the accuracy of each approach when determining specific sentiments.

Study 1: Twitter

In our first study, we investigate the accuracy of off-the-shelf technologies for sentiment analysis of *tweets*, i.e., posts on the Twitter social media platform. Tweets are short in nature; at the time of writing, the maximum length of a tweet is 280 characters. Arguably, the short length makes the task of estimating the sentiment behind a text easier since social media users have to be less verbose and use a straight-to-the-point writing style. Our Twitter data set concerns a study about the problems major U.S. airlines face[5]. A total of 14,640 tweets were scraped from February 16, 2015, to February 24, 2015. Each tweet was then classified as "positive", "negative", or "neutral" by at least two crowd workers working on the crowdsourcing platform formerly known as CrowdFlower, and currently named Figure Eight. In our analysis, we decided to only consider tweets labeled by at least three crowd workers. Although there has been some interesting research on the ideal number of labelers as well as novel relabeling schemes (Ipeirotis et al., 2014; Carvalho, Dimitrov, and Larson, 2015; Carvalho, Dimitrov, and Larson, 2016; Geva, Saar-Tsechansky, and Lustiger, 2019), we do not consider the number of labelers in our analysis beyond the aforementioned cutoff point. Besides removing tweets labeled by less than three workers, we also removed tweets that did not receive sentiment scores or labels from all off-the-shelf technologies due to, for example, being too short. Moreover, we also removed tweets classified as "mixed" by Amazon Comprehend. In total, we ended up with 10,380

[5] At the time of writing, the data set is available at https://www.figure-eight.com/data-for-everyone.

Technique/ Technology	Accuracy (Negative)	Accuracy (Neutral)	Accuracy (Positive)	Accuracy (Overall)
IBM NLU	6,696/7,341 [91.2%]	796/1,531 [52.0%]	1,369/1,508 [90.8%]	8,861/10,380 [85.4%]
Amazon Comprehend	4,902/7,341 [66.8%]	1,251/1,531 [81.7%]	1,391/1,508 [92.2%]	7,544/10,380 [72.7%]
Microsoft Text Analytics	5,036/7,341 [68.6%]	479/1,531 [31.3%]	1,361/1,508 [90.3%]	6,876/10,380 [66.2%]
Google Natural Language	5,705/7,341 [77.7%]	603/1,531 [39.4%]	1,384/1,508 [91.8%]	7,692/10,380 [74.1%]
Bag of Words	3,413/7,341 [46.5%]	1,073/1,531 [70.1%]	1,128/1,508 [74.8%]	5,614/10,380 [54.1%]

Table 2. Accuracy Results Concerning the First Study.

tweets, where 7,341 of them were classified by the crowd workers as negative, 1,531 as neutral, and 1,508 as positive. The average tweet length in our data set is approximately 107.5 characters, the standard deviation being equal to 34.82. The maximum and minimum length are, respectively, 176 and 12 characters.

Results

Table 2 shows the results of our first study. Starting with overall accuracy values, it is interesting to note how all off-the-shelf technologies are more accurate than the bag-of-words approach. In particular, the difference between IBM NLU, the most accurate technology, and the bag-of-words approach is more than 30 percentage points. Even when considering the least accurate off-the-shelf technology, namely Microsoft Text Analytics, the difference is still statistically significant (two-proportions Z-test, $\chi^2 = 319.59$, df = 1, p-value < 2.2e-16). Qualitatively, the above result is also true when considering accuracy for only negative and positive posts. Specifically, Amazon Comprehend, the least accurate off-the-shelf technology for negative posts, is still statistically more accurate than the bag-of-words approach for negative posts (two-proportions Z-test, $\chi^2 = 614.04$, df = 1, p-value < 2.2e-16). In fact, even an unbiased coin is more accurate than the bag-of-words approach when classifying a document as either negative or nonnegative. For positive posts, Microsoft Text Analytics, the least accurate off-the-shelf technology for positive posts, is statistically more accurate than the bag-of-words approach (two-proportions Z-test, $\chi^2 = 123.76$, df = 1, p-value < 2.2e-16).

It is rather interesting to observe that the above results are no longer true when considering only neutral posts. In particular, the bag-of-words approach is the second most accurate method when classifying neutral posts, and the difference to the third most accurate method is an impressive 18.1 percentage points. A possible explanation for this result is that "neutral" is the default sentiment assigned by the bag-of-words approach. In detail, unless there is at least one single word deemed as positive or negative, the score returned by the bag-of-words approach is always equal to zero, meaning a neutral sentiment. As such, researchers and practitioners should expect sentiment classifications to have an inherent bias towards neutral when using the bag-of-words approach, which naturally makes this technique less appealing.

Another interesting point to observe is that there is no single technology that is consistently more accurate than the others across different sentiments. For example, Amazon Comprehend is highly accurate when classifying neutral and positive posts, but drastically less so when handling negative posts. IBM NLU, on the other hand, is highly accurate when classifying positive and negative posts, but considerably less accurate when dealing with neutral posts. These results suggest an interesting research direction, *i.e.*, how to effectively combine many off-the-shelf technologies in order to create a powerful ensemble model for sentiment analysis. We return to this point later in the paper.

Technique/ Technology	Accuracy (Negative)	Accuracy (Neutral)	Accuracy (Positive)	Accuracy (Overall)
IBM NLU	1174/1770 [66.3%]	39/133 [29.3%]	1128/1264 [89.2%]	2341/3167 [73.9%]
Amazon Comprehend	1258/1770 [71.1%]	63/133 [47.4%]	1096/1264 [86.7%]	2417/3167 [76.3%]
Microsoft Text Analytics	994/1770 [56.2%]	32/133 [24.1%]	1058/1264 [83.7%]	2084/3167 [65.8%]
Google Natural Language	1188/1770 [67.1%]	38/133 [28.6%]	894/1264 [70.7%]	2120/3167 [66.9%]
Bag of Words	612/1770 [34.6%]	57/133 [42.9%]	1005/1264 [79.5%]	1674/3167 [52.9%]

Table 3. Accuracy Results Concerning the Second Study.

Study 2: Facebook

In our second study, we investigate the performance of different off-the-shelf technologies for sentiment analysis on social media posts that, unlike in our first study, have less severe length constraints. Our data set comes from the study by Carvalho *et al.* (2019). In detail, the authors collected comments from Starbucks' public page on Facebook from a few days before to a few days after Starbucks pledged to hire thousands of refugees. The research goal was to analyze how the overall sentiment changed from before to after the announcement and, hence, how effective the underlying corporate-social-responsibility policy was. A total of 3,240 posts were analyzed and scored by five random individuals on the crowdsourcing platform Amazon Mechanical Turk. The average sentiment score received by a Facebook post was then used to calculate the post's label. In our experiments, we removed the posts that did not receive sentiment scores or labels from all off-the-shelf technologies or were classified as "mixed" by Amazon Comprehend. We ended up with a total of 3,167 posts, where 1,264 of them were labeled as positive, 133 as neutral, and 1,770 as negative. The average post length in our data set is approximately 205 characters, the standard deviation being approximately 238. The maximum and minimum length are, respectively, 2,711 and 8.

Results

Table 3 shows the results of our second study. Similar to the results from our first study, the overall accuracy and the accuracy on negative posts for all off-the-shelf technologies are higher than the accuracy of the bag-of-words approach. In particular, even the least accurate off-the-shelf technology is still statistically more accurate than the bag-of-words approach for all the posts (two-proportions Z-test, χ^2= 109.45, df = 1, p-value < 2.2e-16) and for negative posts only (two-proportions Z-test, χ^2= 165.44, df = 1, p-value < 2.2e-16). Surprisingly, the bag-of-words approach is only the second least accurate method for positive posts, a case in which all off-the-shelf technologies perform reasonably well.

An interesting finding is that, except for Amazon Comprehend, the overall accuracy of all methods in our second study is lower than in our first study. We previously mentioned a possible explanation for this result, in that shorter texts are less verbose and might have a more salient sentiment, whereas longer posts allow for mixed sentiments in the same text, which might make an overall estimation more complex for all methods. This drop in performance is particularly noticeable for neutral posts, where no method can beat a simple unbiased coin when determining whether a Facebook post is neutral or not. Finally, we note that similar to our first study, no method is consistently more accurate than the others across all sentiments.

Summary & Discussion

Off-the-shelf technologies provided by cloud platforms have the potential to democratize the use of artificial intelligence techniques. Focusing on sentiment analysis tasks, we started this paper by reviewing high-profile off-the-shelf technologies in terms of features and pricing schemes. It is particularly relevant to highlight the cost aspect of using such technologies, *e.g.*, the estimation of sentiments behind thousands of documents might cost a fraction of a penny. We next measured the accuracy of the studied technologies in two different studies involving social media data. It is important to mention that no data preprocessing was performed when analyzing social media posts. Nevertheless, the studied off-the-shelf technologies were still able to produce accurate results. This point underlines the potential of such technologies, in that application developers are no longer required to understand complex machine learning and/or natural language processing techniques in order to estimate the sentiments behind texts. Instead, an application can simply access off-the-shelf technologies when needed via APIs.

In our experiments, we used the still popular bag-of-words approach as a baseline against which we compared four off-the-shelf technologies for sentiment analysis. We believe our findings are of great value to both practitioners and researchers. In particular, we first found that off-the-shelf technologies greatly outperform the bag-of-words approach in terms of overall accuracy. Although we also found that those technologies tend to struggle with neutral posts, we nonetheless believe that the simplicity, cost-effectiveness, and overall accuracy of the studied off-the-shelf technologies imply that any of them should be a candidate to replace the bag-of-words approach in any future research study or practical application.

Another interesting finding is that all but one of the studied off-the-shelf technologies underperformed in terms of overall accuracy when analyzing longer texts in comparison to shorter texts. We believe this result might be general, but we acknowledge that more research is needed to (dis)confirm whether it holds beyond the data sets we analyzed in this paper. Finally, we found that no single technology consistently outperforms all the others when determining different types of sentiments or across the two studies. This observation creates a great opportunity for the development of *ensemble models*. In detail, ensemble methods are machine learning algorithms that combine classifications or predictions made by individual methods when classifying or predicting new data points. Theoretically, an ensemble method is more accurate than any of its members when the individual classifiers or predictors are accurate and diverse (Hansen and Salamon, 1990). In this context, accuracy means that the error rate of each technique is lower than random guessing, and diversity means that the errors made by the individual techniques are uncorrelated. The question that arises is then: how to combine and appropriately weigh the classifications by each off-the-shelf technology? There are several different ways of combining predictions made by different techniques (*e.g.,* see the work by Winkler, 1981; Hora, 2004; Carvalho and Larson, 2013). In ongoing work, we are currently investigating the performance of different ensembles created by combining off-the-shelf technologies for sentiment analysis. Moreover, we plan to perform different analyses with the collected data, e.g., by taking into account the distance between sentiments. Currently, we consider an estimated negative sentiment to be as wrong as an estimated neutral sentiment when the true label is positive. Finally, we plan to investigate richer sentiment levels beyond the three (negative, neutral, and positive) studied in this paper.

Low-Code/No-Code Application Development

As we suggested above, off-the-shelf technologies provided by cloud platforms are meant to be easily integrated with other applications. We see this practice as part of a bigger trend that goes by the name of *low-code development*. In fast-paced business environments, organizations are constantly looking for quicker and cheaper ways to meet their information technology needs, including software development. In response, low-code development platforms have emerged with the promise that organizations can rely on nontechnical professionals to develop enterprise-level applications. Hopefully, by adding in extra parties to the development process, more business ideas can be quickly explored, thus leading to greater value generation. Several platforms, such as Mendix App Platform[6] and Microsoft Power Apps[7], are now offering graphical tools that enable one to rapidly develop model-driven applications with little to no programming

[6] https://www.mendix.com/
[7] https://powerapps.microsoft.com/

experience. This is also happening with specific emerging technologies, such as blockchain, where newer tools facilitate the development of blockchain network prototypes and smart contracts (Carvalho, 2020).

The above said, off-the-shelf AI technologies promote the development of applications powered by state-of-the-art AI techniques, and all of that requires just a few lines of programming code. When delivered by cloud platforms, those technologies can be seen as part of the software as a service (SaaS) paradigm, meaning that organizations can use the technologies when they need, pay a relatively low fee for what they use, and quit using whenever they want to. Naturally, organizations relying on SaaS must place their trust in the service provider. For example, when using off-the-shelf technologies for sentiment analysis, organizations implicitly trust that the underlying models were trained and evaluated appropriately and that the data used in the training process are relevant to the application at hand.

Regarding the above point, we mentioned before that off-the-shelf technologies are black boxes, meaning that the precise techniques to preprocess the data and create models are not publicly available. This is rather understandable since those technologies are proprietary software. But an equally important point regards the data the models were trained on. For example, a model trained to estimate the sentiment behind lengthy book reviews written in a formal style might not be appropriate to classify tweets, which are considerably shorter and oftentimes written in a more casual manner. To tackle this issue, many cloud platforms are now allowing users to train models by simply uploading a data set containing two attributes, namely the text to be analyzed and the corresponding label. A model will then be trained using the cloud platform's proprietary algorithms. This "no-code" solution is currently offered by all the four cloud platforms we have studied in this paper. For example, for sentiment analysis, IBM offers Watson Natural Language Classifier[8]; Amazon Comprehend allows for the creation of custom models; Google offers AutoML Natural Language[9]; and Microsoft offers AI Builder[10]. All these services allow developers to quickly and easily build custom text classification models without the need for a text mining or machine learning background. The custom models can then be accessed via APIs in a manner similar to the off-the-shelf technologies we discussed in this paper. An interesting research question, which we leave for future work, is then: how accurate are these tailored models in comparison to off-the-shelf technologies? When combined, the no-code approach for model training together with the low-code approach for model usage have the potential to not only democratize AI, but also to revolutionize different products, services, and research endeavors by allowing them all to use powerful, state-of-the-art AI technologies in a very accessible and cost-effective way.

REFERENCES

Babic, J., Carvalho, A., Ketter, W., and Podobnik, V. 2017. "Electricity Trading Agent for EV-Enabled Parking Lots," in *Agent-Mediated Electronic Commerce. Designing Trading Strategies and Mechanisms for Electronic Markets*, Ceppi S., David E., Hajaj C., Robu V., and Vetsikas I. (eds.), pp. 35-49.

Babic, J., Carvalho, A., Ketter, W., and Podobnik, V. 2017. "Evaluating Policies for Parking Lots Handling Electric Vehicles," *IEEE Access* (6), pp. 944-961.

Brown, N. and Sandholm, T. 2018. "Superhuman AI for Heads-Up No-Limit Poker: Libratus Beats Top Professionals," *Science* (359:6374), pp. 418-424.

Cao, M. and Rhue, L. 2019. "Disposed of Bitcoin? Using the Disposition Effect to Understand Financial News Sentiment and Bitcoin Returns," in *Proceedings of the 25th Americas Conference on Information Systems*.

Carvalho, A. 2020. "A Permissioned Blockchain-Based Implementation of LMSR Prediction markets," *Decision Support Systems* (130), pp. 1-15.

Carvalho, A., Dimitrov, S., and Larson, K. 2015. "A Study on the Influence of the Number of MTurkers on the Quality of the Aggregate Output," in *Multi-Agent Systems. Lecture Notes in Computer Science*, vol. 8953, N. Bulling (ed.), Springer International Publishing, pp. 285–300.

Carvalho, A., Dimitrov, S., and Larson, K. 2016. "How Many Crowdsourced Workers Should a Requester Hire?" *Annals of Mathematics and Artificial Intelligence* (78:1), pp. 45-72.

[8] https://www.ibm.com/cloud/watson-natural-language-classifier

[9] https://cloud.google.com/automl

[10] https://powerapps.microsoft.com/en-us/ai-builder/

Carvalho, A. and Larson, K. 2013. "A Consensual Linear Opinion Pool," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pp. 2518–2524.

Carvalho, A., Levitt, A., Levitt, S., Khaddam, E., and Benamati, J. 2019. "Off-the-Shelf Artificial Intelligence Technologies for Sentiment and Emotion Analysis: A Tutorial on Using IBM Natural Language Processing," *Communications of the Association for Information Systems* (44), pp. 918–943.

Carvalho, A. and Oliveira, R. 2011. "Reinforcement Learning for the Soccer Dribbling Task," in *Proceedings of the 2011 IEEE Conference on Computational Intelligence and Games*, pp. 95–101.

Dag, A., Topuz, K., Oztekin, A., Bulur, S., and Megahed, F. M. 2016. "A Probabilistic Data-Driven Framework for Scoring the Preoperative Recipient-Donor Heart Transplant Survival," *Decision Support Systems* (86), pp. 1-12.

Davids, A. 2002. "Urban Search and Rescue Robots: From Tragedy to Technology," *IEEE Intelligent Systems*, (17:2), pp. 81-83.

Fan, W. and Gordon, M. D. 2014. "The Power of Social Media Analytics," *Communications of the ACM* (57:6), pp. 74–81.

Fan, S., Ilk, N., and Zhang, K. 2015. "Sentiment Analysis in Social Media Platforms: The Contribution of Social Relationships," in *Proceedings of the 2015 International Conference on Information Systems*.

Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., and Mueller, E. T. 2013. "Watson: Beyond Jeopardy!" *Artificial Intelligence* (199), pp. 93–105.

Geva, T., Saar-Tsechansky, M., and Lustiger, H. 2019. "More for Less: Adaptive Labeling Payments in Online Labor Markets," *Data Mining and Knowledge Discovery*, (33:6), pp. 1625-1673.

Hansen, L. K. and Salamon, P. 1990. "Neural Network Ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (12:10), pp. 993-1001.

Hora, S. C. 2004. "Probability Judgments for Continuous Quantities: Linear Combinations and Calibration," *Management Science* (50:5), pp. 597–604.

Hu, M. and Liu, B. 2004. "Mining and Summarizing Customer Reviews," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168-177.

Ipeirotis, P. G., Provost, F., Sheng, V. S. and Wang, J. 2014. "Repeated Labeling Using Multiple Noisy Labelers," *Data Mining and Knowledge Discovery* (28:2), pp. 402-441.

Jerdack, N., Dauletbek, A., Divine, M., Hult, M., and Carvalho, A. 2018. "Understanding What Drives Bitcoin Trading Activities," in *Proceedings of the 2018 Annual Meeting of the Decision Sciences Institute*, pp. 1864–1872.

Liu, B. 2012. *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers.

Lowe, W., Benoit, K., Mikhaylov, S. and Laver, M. 2011. "Scaling Policy Preferences from Coded Political Texts," *Legislative Studies Quarterly*, (36:1), pp. 123-155.

Mai, F., Shan, Z., Bai, Q., Wang, X., and Chiang, R. H. L. 2018. "How Does Social Media Impact Bitcoin Value? A Test of the Silent Majority Hypothesis," *Journal of Management Information Systems* (35:1), pp. 19–52.

Medhat, W., Hassan, A. and Korashy, H. 2014. "Sentiment Analysis Algorithms and Applications: A Survey," *Ain Shams Engineering Journal*, (5:4), pp. 1093-1113.

Oliveira, R., Adeodato, P., Carvalho, A., Viegas, I., Diego, C., Ing-Ren, T. 2009. "A Data Mining Approach to Solve the Goal Scoring Problem," in *Proceedings of the 2009 International Joint Conference on Neural Networks*, pp. 2347–2352.

Pentland, S., Spitzley, L., Fuller, C. and Twitchell, D. 2019. "Data Quality Relevance in Linguistic Analysis: The Impact of Transcription Errors on Multiple Methods of Linguistic Analysis," in in *Proceedings of the 25th Americas Conference on Information Systems*.

Russell, S. J. and Norvig, P. 2016. *Artificial Intelligence: A Modern Approach*, 3rd edition, Pearson Hall.

Vroegindeweij, R. and Carvalho, A. 2019. "Do Healthcare Workers Need Cognitive Computing Technologies? A Qualitative Study Involving IBM Watson and Dutch Professionals," *Journal of the Midwest Association for Information Systems* (1), pp. 51–68.

Winkler, R. L. 1981. "Combining Probability Distributions from Dependent Information Sources," *Management Science* (27:4), pp. 479–488.