Studies on the Accuracy of Ensembles of Cloud-Based Technologies for Sentiment Analysis

Completed Research

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

Jiaozhe Xu
Farmer School of Business
Miami University
xuj21@miamioh.edu

Abstract

The accuracy of different cloud-based technologies for sentiment analysis may vary based on attributes such as the length of the analyzed texts and the dominant sentiment in a corpus. A potential strategy to reduce the variability in accuracy is to create ensemble models formed by individual technologies. Our goal in this paper is to study the performance of different ensembles of cloud-based technologies for sentiment analysis. Overall, we find that ensemble models perform better on long texts, a scenario where individual technologies tend to struggle. We also find that score-based ensembles perform better than label-based ensembles. Besides being of value to practitioners, we discuss how our results might increase the reliability of research findings that rely on sentiment analysis. In particular, we argue that ensemble models may reduce the chances of sentiment-dependent results being driven by a single technology.

Keywords

Sentiment analysis, ensemble modeling, artificial intelligence, software as a service.

Introduction

Artificial intelligence (AI) is quickly becoming ubiquitous. For example, in healthcare, AI-based technologies are used to collect and process non-uniform data from patients, which can be used to efficiently search for a cure for different diseases (Vroegindeweij and Carvalho, 2019). In human resources, AI techniques are often used to aid in filtering and selecting job applicants (Tambe, Cappelli, and Yakubovich, 2019). In the energy domain, AI models are vital to enable demand-side response by looking for suitable solutions that can help consumers to reduce or shift electricity consumption away from periods of low generation (Antonopoulos *et al.*, 2020; da Silva *et al.*, 2020) as well as when strategically determining the placement of new electric vehicle charging stations (Pevec *et al.*, 2017; Pevec *et al.*, 2018). *Natural language processing* (NPL) is one specific subfield of AI that is consistently rising in popularity (Manning and Schutze, 1999). NLP covers various tasks and applications such as automatic text summarization (Gambhir and Gupta, 2017), topic modeling (Wallach, 2006), and sentiment analysis (Feldman, 2013), the latter being our focus in this work.

Sentiment analysis consists of techniques to estimate a writer's sentiment from written texts. Since those techniques come from several fields, such as computational linguistics and statistics, developing and deploying effective sentiment-analysis models can be a daunting task that requires tremendous expertise. An immediate implication is that small- and even mid-sized organizations might miss out on the opportunities brought about by sentiment analysis and, more broadly, AI for not having the resources to hire experts. Fortunately, recent years have seen the rise of cloud platforms that offer off-the-shelf AI technologies that follow the low/no-code application development paradigm. For example, as we elaborate later in this paper, prominent cloud platforms such as Amazon Web Services, Google Cloud, Microsoft Azure, and IBM Cloud all offer cloud-based sentiment-analysis services that can be easily accessed through application programming interfaces (APIs).

A recently discussed issue regarding cloud-based technologies for sentiment analysis is that their accuracy non-uniformly varies based on a series of factors, *e.g.*, their overall accuracy depends on the length of the texts in a corpus and on the predominant sentiment (Carvalho and Harris, 2020). To mitigate the issue of variance in individual accuracy, we suggest in this paper to combine the outputs from many sentiment-analysis technologies to create *ensemble models*. Our overarching research questions (RQ) are:

RQ #1: under which circumstances are ensembles of sentiment-analysis technologies more accurate?

RQ #2: which ensemble models are more accurate when estimating sentiments?

A successful answer to the above questions shall guide practitioners and potentially cause research findings that rely on sentiment analysis to be more robust. In particular, we argue that, for example, findings that rely on sentiment scores/labels might be driven by the underlying technology/technique rather than being actual results. One can mitigate this problem by having a basket of technologies assigning sentiments to individual texts. But to do so effectively, one needs answers to our research questions. That said, this paper represents a first attempt at answering the above research questions. In particular, we study three different ensemble models that combine outputs from individual technologies in different ways. We investigate the performance of the ensemble models in two studies representing scenarios of short and long texts. We find that score-based ensemble models are more accurate on average than label-based ensemble models, and the former tend to be more accurate than individual technologies on long texts. Throughout the paper, we provide explanations for the above results, *e.g.*, outputs from individual technologies are more correlated in short than in long texts.

Besides this introductory section, the rest of the paper is organized as follows. In the following section, we briefly introduce ensemble modeling and sentiment analysis. This section is followed by a discussion on the data sets we use in our studies as well as the performed data analysis. Finally, in the last section, we summarize our findings and conclude our work.

Research Background

We next introduce the concepts of ensemble modeling and sentiment analysis.

Ensemble Modeling

For our purposes, ensemble modeling is the process of aggregating multiple outputs from many models to create a single output. This task can be achieved by using different models and/or by training models on different data sets (Winkler and Clemen, 2004). The motivation for using ensemble models is often to reduce the expected error when making estimations/predictions. This result happens because of the expectation that extreme and potentially wrong outputs will cancel each other out. More formally, an ensemble method is more accurate than any of its members when the individual models are accurate and diverse (Hansen and Salamon, 1990), where accuracy means that the error rate is lower than random guessing, and diversity means that the errors made by the individual models are uncorrelated.

Ensemble modeling has been applied in a plethora of scenarios, *e.g.*, to predict stock prices (Weng *et al.*, 2018), in bioinformatics (Yang *et al.*, 2010), when analyzing data streams (Bifet *et al.*, 2009), among many others. As we suggested above, the idea of combining outputs from many individual models perfectly fits the context of cloud-based sentiment analysis because it has been suggested that different technologies perform better under different circumstances (Carvalho and Harris, 2020).

Sentiment Analysis

The process of estimating sentiments from written texts is traditionally referred to as sentiment analysis. There are many ways of computationally estimating the sentiment behind a text. For example, the bag-of-words approach assigns polarity scores to individual words and, subsequently, aggregates the individual scores into a single score (Hu and Liu, 2004). Recently, more powerful sentiment-analysis methods have been proposed using deep-learning models (Zhang, Wang, and Liu, 2018). In terms of applications, sentiment-analysis techniques have been applied in several domains, ranging from understanding how sentiments behind social-media posts influence the value of cryptocurrencies (Jerdack *et al.*, 2018) to hotspot detection based on comments on online forums (Li and Wu, 2010) as well as online reviews

(Valdivia, Luzón, and Herrera, 2017). Although our focus in this paper is on textual data, sentiment analysis techniques can also be applied to other unstructured data, such as audio and videos (Wöllmer *et al.*, 2013).

Our interest in this paper is on cloud-based sentiment-analysis technologies that are ready to be used. Features such as accessible prices, high accuracy, and ease of use make these technologies highly attractive, and they effectively democratize access to state-of-the-art AI models. We note that some previous work has compared the performance of such technologies (Qaisi and Aljarah, 2016; Koneru *et al.*, 2018; Carvalho and Harris, 2020). However, to the best of our knowledge, our work is the first to study ensembles of off-the-shelf, cloud-based technologies for sentiment analysis.

Data Source & Preprocessing

We use two different data sets in our studies to measure the accuracy of ensembles of cloud-based technologies for sentiment analysis. We explain the nature of such data sets in the following two subsections. These are followed by a description of the technologies we use in our experiments and the data transformations we perform before our analysis.

Twitter Data Set

The first data set represents short texts as the underlying observations are posts on the social media platform Twitter. Specifically, this data set concerns the opinions of Twitter users on prominent U.S. airlines. The raw data set consists of 14,640 tweets collected in February 2015. On average, each collected tweet has approximately 107 characters, the minimum and maximum values being, respectively, 12 and 176. Each tweet was classified as "positive", "neutral", or "negative" by crowd workers from the crowdsourcing platform CrowdFlower. To decrease the number of non-consensual classifications, we only consider tweets labeled by at least three crowd workers, which reduces the total number of tweets to 14,621. Beyond this reason, there is no attempt to determine the optimal number of labelers in our study, as it is done in other research (Ipeirotis, Provost, Sheng, and Wang, 2014; Carvalho, Dimitrov, and Larson, 2015; Carvalho, Dimitrov, and Larson, 2016; Geva, Saar-Tsechansky, and Lustiger, 2019).

Facebook Data Set

The second data set represents long texts as it concerns posts on the social media platform Facebook. In particular, the data come from Starbucks' public page on Facebook. The same is originally part of a study to understand changes in sentiment towards Starbucks after its pledge to hire refugees (Carvalho, Levitt, Levitt, Khaddam, and Benamati, 2019). The raw data set consists of 3,240 posts collected in 2017. On average, each collected post has approximately 238 characters, the minimum and maximum values being, respectively, 2,711 and 8. Each Facebook post was classified as "very positive", "positive", "neutral", "negative", or "very negative" by five crowd workers from the crowdsourcing platform Amazon Mechanical Turk. Similar to our Twitter data set, we do not attempt to determine an optimal number of labelers for the Facebook data set.

Sentiment Analysis Technologies

Having defined the data sets we use in our study, we now explain the cloud-based technologies we combine to create ensembles. These technologies follow the software-as-a-service paradigm inside some of the best-known cloud platforms. We apply each technology to every one of the texts in our data sets without any *ex-ante* processing of the data. From IBM Cloud, we use the *Natural Language Understanding* (NLU) service to obtain sentiment scores within the interval [-1, 1], where -1 represents an entirely negative post, 0 (zero) represents a neutral post, and 1 represents an entirely positive post. From Microsoft Azure, we use the *Text Analytics* cognitive service to obtain scores inside the interval [0, 1]. Finally, from Google Cloud, we use the Natural Language service to obtain sentiment scores within the interval [-1, 1]. Table 1 summarizes the above technologies. Carvalho and Harris (2020) provide an in-depth analysis of each technology, including usage costs and data request constraints. The *Amazon Comprehend* natural language processing service from Amazon Web Services is one high-profile sentiment-analysis technology missing in our study. The reason for its absence is that, unlike the studied technologies, Amazon Comprehend returns a probability

Cloud Platform	Technology	Output	Sample Output
IBM Cloud	Natural Language Understanding	Score inside the interval [-1,1]	0.94
Microsoft Azure	Text Analytics	Score inside the interval [0,1]	0.94
Google Cloud	Natural Language	Score inside the interval [-1,1]	0.9

Table 1. Summary of the Studied Technologies. The Sample Output Column Concerns the Text "*I applaud Starbucks for its strong stance*."

distribution over the sentiments "positive", "neutral", "negative", and "mixed", instead of a single sentiment score. Deriving a single score from a distribution may introduce confounding factors in our studies.

Data Transformation & Cleaning

Since some of our analysis consists of aggregating scores produced by different technologies for sentiment analysis, we must first ensure that the individual scores are within the same interval. For Microsoft Azure's Text Analytics technology, we calculate new scores by defining a positive affine transformation of the previous scores. Specifically, we set the new score as two times the old score minus one. Consequently, the new score is within the interval [-1, 1]. After this transformation, we find 16 tweets and 31 Facebook posts with scores assigned by NLU that are considerably greater than 1. This issue is assumed to be a glitch and, consequently, we remove the underlying tweets/posts from our data sets. The final data sets now have 14,605 tweets and 3,209 Facebook posts. Next, we remove tweets/posts that do not receive a score from at least one of the three technologies. Cases like this happen when, for example, the underlying text is too short, which makes the estimation of sentiments infeasible. All Facebook posts have sentiment scores from all technologies, but 4,194 tweets do not have at least one sentiment score. We thus remove the underlying texts, and the final number of observations in our Twitter data set is 10,411, whereas the number of Facebook posts remains unchanged at 3,209.

With regards to the labels provided by humans, we note that each tweet has a gold-standard label defined by CrowdFlower based on its internal aggregation algorithm and input from its crowd workers. To create a gold-standard label for each Facebook post, we first assign a number to each crowd worker's reported label, *i.e.*, the values -2, -1, 0, 1, and 2 represent the labels "very negative", "negative", "neutral", "positive", and "very positive". Thereafter, a Facebook post's gold-standard label is defined as the median value reported by the crowd workers. If the median value is greater than 0 (zero), then the gold-standard label is defined as "positive". Otherwise, if the median value is less than 0 (zero), then the gold-standard label is defined as "negative". Finally, if the median value is equal to 0 (zero), then the gold-standard label is defined as "neutral". Although there are other ways of combining subjective opinions that may better represent consensus (Carvalho and Larson, 2013), we nonetheless believe median values offer an effective solution that mitigates the impact of outlier labels.

At this point, both data sets are unbalanced with respect to gold-standard labels. In particular, the Twitter data set has, respectively, 7,368, 1,532, and 1,511 negative, neutral, and positive tweets. The Facebook data set has 1,618 negative, 463 neutral, and 1,128 positive posts. To balance our data sets, we perform stratified sampling. In particular, we sample without replacement an equal number of observations for each sentiment. Each sample size is equal to the absolute frequency of the least popular sentiment in the data set. This sampling approach implies that our final Twitter data set has 3 * 1,511 = 4,533 observations, whereas the Facebook data set has 3 * 463 = 1,389 observations.

Ensemble Modeling & Data Analysis

In what follows, we explain how we combine sentiment scores and labels from sentiment-analysis technologies to create ensemble models. This description is followed by an analysis of the ensemble models' performance under the two data sets we previously described.

Ensemble Models

We create ensemble models by combining outputs produced by the three sentiment-analysis technologies we study in this work. The first model — henceforth referred to as the "average model" — aggregates the three sentiment scores received by a text by simply averaging them. To calculate the accuracy of this model on different data sets, we transform each average score into one of the following labels: "positive", "neutral", or "negative". To do so, we define score intervals of roughly equal length to map sentiment scores onto labels. In particular, if the average score is within the interval [-1, -0.33), then the assigned label is "negative". Alternatively, if the average score is within the interval [-0.33, 0.33], then the corresponding label is "neutral". Finally, the label is "positive" when the average score is within the interval (0.33, 1].

The second model — henceforth referred to as the "weighted average model" — aggregates the three sentiment scores received a text by calculating the weighted average. In detail, we assign the weight 0.5 to the most accurate individual technology and 0.25 to each remaining technology. After combining the individual scores, we assign sentiment labels similar to how we do with the average model.

The third ensemble model we consider in this paper — henceforth called the "majority model" — aggregates sentiment labels instead of sentiment scores. To do so, we first assign a sentiment label to every single score produced by a sentiment-analysis technology. Specifically, if the sentiment score is less than (respectively, greater than) zero, then the assigned sentiment is "negative" (respectively, "positive"). For a sentiment score equal to zero, the assigned sentiment is "neutral". After having labels associated with each sentiment score produced by an individual technology, we next aggregate the resulting labels by looking at the most popular label for a given text, *i.e.*, we take a majority voting approach. We pick the neutral sentiment as the tiebreaker when there is no clear winner, *i.e.,* when all the individual technologies produce a different label.

In the following subsections, we compare the performance of the above-mentioned ensemble models against the individual technologies on different data sets.

Study #1: Twitter Data Set

In our first study, we investigate the accuracy of ensemble models on the Twitter data set, which in turn represents a collection of short texts.

Correlation Analysis

We start by analyzing the (linear) correlation between the scores resulting from individual technologies. To a certain degree, this analysis enables us to understand how effective ensembles can be. Table 2 presents all the correlation coefficients. The resulting p-values are all less than 10^{-4} even after adjusting for multiple comparisons using Bonferroni correction. From Table 2, one can see that the sentiment-analysis technologies produce scores that are relatively highly correlated. Ideally, the correlation between the *errors* produced by different technologies/models should be small to make an ensemble effective. But since we do not have ground-truth sentiment scores, estimating errors without relying on labels is not feasible. The correlation coefficients nonetheless indicate that an ensemble model might not be so effective in this setting because the scores from the individual technologies tend to move strongly in the same positive direction.

	Natural Language	**NLU**	**Text Analytics**
Natural Language	1.00	0.75	0.67
NLU	0.75	1.00	0.75
Text Analytics	0.67	0.75	1.00

Table 2. Linear Correlation Between the Individual Scores on the Twitter Data Set.

We repeat the above analysis, but now for sentiment labels, as opposed to sentiment scores. In particular, we calculate the Spearman correlation between the labels resulting from individual technologies. To do so, we assign the value -1 to negative labels, the value 0 (zero) to neutral labels, and the value 1 to positive

labels. Clearly, as long as the values are correctly ordered, the specific values assigned to labels do not matter to calculate the Spearman correlation since this measure works on an ordinal scale. Table 3 shows the results from our second correlation analysis. The resulting p-values are all less than 10^{-4} even after adjusting for multiple comparisons using Bonferroni correction. Table 3 shows a relatively strong monotonic relationship between the sentiments produced by the individual technologies. Once again, although we are not estimating the correlation between the models' errors, the results in Table 3 indicate that ensemble models based on labels might not be as effective for this data set.

	Natural Language	**NLU**	**Text Analytics**
Natural Language	1.00	0.71	0.60
NLU	0.71	1.00	0.66
Text Analytics	0.60	0.66	1.00

Table 3. Spearman Correlation Between the Individual Labels on the Twitter Data Set.

The above correlation analyses indicate that ensembles might not be as effective in this first study. We formalize this conclusion by analyzing the accuracy of models and technologies in the following subsection.

Accuracy Analysis

We calculate the overall accuracy of each model and technology on the Twitter data set by comparing how often the sentiment label reported by the model/technology agrees with the aggregate label reported by crowd workers. Table 4 shows the results. Since NLU is the most accurate individual technology, we define the weights in the weighted average model as 0.5 for NLU and as 0.25 for the other technologies.

Natural Language	**NLU**	**Text Analytics**	**Average Model**	**Weighted Avg. Model**	**Majority Model**
69.4%	77.5%	63.4%	76.2%	78.0%	76.0%

Table 4. Overall Accuracy of Different Models/Technologies on the Twitter Data Set.

As expected from the correlation analysis, the ensemble models are not too effective. For example, NLU is slightly more accurate than both the average model and the majority model, and it is considerably more accurate than the other two individual technologies. The weighted average model leverages this information to build the most accurate model among all. A possible explanation for the above result is the vast disparity between the accuracy of NLU and that of the other technologies. In particular, the former is so much more accurate than the others that combining the three of them together causes the overall accuracy to decline, *i.e.*, it brings NLU's accuracy down. The only case when this result does not happen is when NLU is weighted higher in the ensemble. We further discuss this result in the last section of the paper.

Study #2 Facebook Data Set

We next move to our second study, whose data set represents a collection of long texts. Similar to what we do in our first study, we start by performing a correlation analysis before moving to an accuracy analysis.

Correlation Analysis

We show in Table 5 the correlation coefficients resulting from our linear correlation analysis. The resulting p-values are all less than 10^{-4} even after adjusting for multiple comparisons using Bonferroni correction. Table 5 shows that the sentiment-analysis technologies produce scores that are relatively highly correlated.

Interesting, all correlation coefficients are less than the respective coefficients from the previous study (Table 2). This preliminary result indicates that score-based ensemble models might perform better in this setting with longer texts than in the previous setting with shorter texts.

	Natural Language	**NLU**	**Text Analytics**
Natural Language	1.00	0.67	0.58
NLU	0.67	1.00	0.68
Text Analytics	0.58	0.68	1.00

Table 5. Linear Correlation Between the Individual Scores on the Facebook Data Set.

Moving to the analysis of sentiment labels, as opposed to sentiment scores, we calculate the Spearman correlation between the labels resulting from individual technologies similar to how we do it in our first study. Table 6 shows the results from our second correlation analysis, where the resulting *p*-values are all less than 10^{-4} after adjusting for multiple comparisons using Bonferroni correction. Table 6 shows a relatively strong monotonic relationship between the sentiments produced by the individual technologies. But again, all the individual correlation coefficients are less than the respective coefficients from our previous study (Table 3).

	Natural Language	**NLU**	**Text Analytics**
Natural Language	1.00	0.60	0.50
NLU	0.60	1.00	0.57
Text Analytics	0.50	0.57	1.00

Table 6. Spearman Correlation Between the Individual Labels on the Facebook Data Set.

The above correlation analyses indicate that although the correlations between outputs from individual technologies are still high, they are nonetheless less than in our previous study. This result indicates that ensembles might be more effective in this study with long texts than when working with short texts. We formalize this conclusion by performing an accuracy analysis in the following subsection.

Accuracy Analysis

We calculate the overall accuracy of each model and technology on the Facebook data set in a way similar to what we do with the Twitter data set in our first study. Table 7 shows the overall accuracy results. Given that NLU is the most accurate individual technology, we define the weights in the weighted average model as 0.5 for NLU and as 0.25 for the other technologies.

Natural Language	**NLU**	**Text Analytics**	**Average Model**	**Weighted Avg. Model**	**Majority Model**
56.9%	62.0%	55.1%	62.3%	63.1%	61.6%

Table 7. Overall Accuracy of Different Models/Technologies on the Facebook Data Set.

Compared to the other technologies, Table 7 shows that the ensemble models are considerably more effective now than in our first study. For example, although NLU is still the most accurate individual

technology, it is only marginally more accurate than the majority model while being less accurate than the average and weighted average models. Furthermore, the difference in accuracy between NLU and the other technologies is now smaller than in our first study. It has been suggested that cloud-based technologies struggle to accurately estimate the sentiment behind long texts (Carvalho and Harris, 2020). This finding may explain why the individual accuracy goes down in our second study and why the overall accuracy is more uniform across all the technologies. This result, in turn, favors the (weighted) average model in that individual technologies now better and more equally contribute to the success of the ensemble.

Summary & Discussion

Different technologies for sentiment analysis can provide different outputs when analyzing the same text. The quality of such outputs may vary based on the nature of the underlying corpus. Thus, instead of choosing a single technology, it seems reasonable to consider aggregating outputs from many technologies. In this manner, the ultimate aggregate output has the potential to be more accurate in expectation than that produced by any individual technology. This approach can then provide better estimates of sentiments and, thus, improve decisions based on those estimates.

Given the above motivation, we study in this paper different ensemble models based on cloud-based technologies for sentiment analysis. In two studies involving short and long texts, we find a high degree of correlation between the sentiment scores as well as between the sentiment labels produced by those technologies. Moreover, the degree of correlation is higher for short, as opposed to long texts. In terms of accuracy, we find that ensemble models that average the scores from individual technologies perform better than models that aggregate sentiment labels. Moreover, relative to individual technologies, ensembles tend to be more accurate when estimating the sentiments behind long texts. Finally, the weighted-average approach is the only ensemble model we study that is consistently more accurate than the individual technologies and other ensembles.

The above results lead to interesting preliminary conclusions. As we mentioned before, a condition for ensembles to be more accurate than individual technologies is a low correlation between the latter's errors. But estimating errors can be infeasible since one might not have access to gold-standard outputs in practice. Our results, in turn, suggest that ensembles are more accurate when the outputs produced by the underlying technologies are not highly correlated among themselves. This scenario tends to happen when the underlying tasks are more challenging since there may be no consensus on the right output. In the context of sentiment analysis, we argue that estimating the sentiments behind long instead of short texts is a more challenging task. In particular, short texts are less verbose and tend to have a more salient sentiment. In contrast, longer texts allow for mixed sentiments in the same text, making an overall sentiment estimation more complicated. In other words, we believe ensemble models are more accurate relative to the individual sentiment-analysis technologies when estimating the sentiment behind long texts. This observation effectively answers our first research question, namely, "*under which circumstances are ensembles of sentiment-analysis technologies more accurate*?".

Another interesting observation concerns model composition. Our results show that averaging sentiment scores produces more accurate results than aggregating labels. This result effectively answers our second research question, namely, "*which ensemble models are more accurate when estimating sentiments*?". At the same time, this result produces several new research questions. For example, how can one assign weights to different sentiment-analysis technologies in an ensemble model? The approach we follow in our experiments is simply to assign more weight to the most accurate technology, namely NLU. In practice, this approach is only feasible if one acquires gold-standard labels to measure the accuracy of individual technologies. Subsequently, one can derive weights based on the accuracy results. With or without access to gold-standard labels, there is literally an infinite number of ways to define weights. Since we find that the outputs from individual technologies are correlated, a promising approach to explore in the future is to borrow from the expert forecasting literature and define aggregation methods for dependent sources (Winkler, 1981; Clemen, 1989; Winkler *et al.*, 2019). Along similar lines, an exciting avenue for future research is to explore different ways of aggregating labels. Technically, one can use any voting mechanism to combine the outputs produced by individual technologies (Brown, 2010).

Continuing with possible future research directions, it is clear that our results come from a limited number of studies. More experiments are required to generalize the findings we report here, ideally using data sets

beyond social media (*e.g.,* online reviews). Another reason for replications is that cloud-based technologies for sentiment analysis tend to be black-boxes, *i.e.*, their source code is not publicly available. Consequently, the way these models assign scores to texts might change at any time, which calls for continuous experimentation with different versions of the technologies. Such experiments can involve not only off-the-shelf technologies but also models tailored to one's data set. Specifically, the cloud platforms studied in this paper all allow a user to upload a data set to get a model trained using proprietary algorithms. Hence, it is interesting to investigate how much more accurate these models are than off-the-shelf models. Similarly, it is worth investigating how much more accurate ensembles of custom models are in relation to ensembles of off-the-shelf technologies. Clearly, cost-considerations must be considered when building tailored models since they are more costly to use (financially and timewise) than pre-built models. The cost aspect is not currently addressed in our work, but it is worth performing a cost-benefit analysis regarding the use of ensembles of cloud-based AI technologies. Such a cost analysis can help one solve another relevant question: how many technologies should one add to an ensemble? From the forecasting literature, it is well-known that there is a marginal gain in accuracy when adding one extra expert/technology beyond three (Winkler and Clemen, 2004), the number we use in this paper. But understating whether this result also holds in the context of sentiment analysis is an open question.

Answers to the above open questions will not only help practitioners produce more accurate estimates of sentiment analysis, but they will also potentially help research findings that depend on sentiment analysis to be more robust and reliable. For example, sentiment scores are often used as predictors in regression-based models, where the associated coefficients measure the influence of sentiments on the dependent variable when holding all the other variables constant. But any found statistical association is only as reliable as the underlying sentiment-analysis technology/technique. By studying ensemble models, we hope to shed light on that issue and suggest a way of removing the dependence on a single technology/technique, which can then improve the true knowledge generation process.

REFERENCES

Antonopoulos, I., Robu, V., Couraud, B., Kirli, D., Norbu, S., Kiprakis, A., Flynn, D., Elizondo-Gonzalez, S., and Wattam, S. 2020. "Artificial Intelligence and Machine Learning Approaches to Energy Demand-Side Response: A Systematic Review," *Renewable and Sustainable Energy Reviews* (130), 109899.

Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., and Gavalda, R. 2009. "New Ensemble Methods for Evolving Data Streams," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 139-148.

Brown, G. 2010. "Ensemble Learning," in *Encyclopedia of Machine Learning*, Sammut, C. and Webb, G. I. (eds.), Springer, pp. 15-19.

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.

Carvalho, A. and Harris, L. 2020. "Off-the-Shelf Technologies for Sentiment Analysis of Social Media Data: Two Empirical Studies," in *Proceedings of the 26th American Conference on Information Systems*.

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.

Clemen, R. T. 1989. "Combining Forecast: A Review and Annotated Bibliography," *International Journal of Forecasting* (5), pp. 559–583.

da Silva, I. R. S., Rabêlo, R. A. L., Rodrigues, J. J. P. C., and Carvalho, A. 2020. "A Multi-Objective Approach for Energy Management in a Microgrid Scenario," in Proceedings *of the 5th International Conference on Smart and Sustainable Technologies*.

da Silva, I. R. S., Rabêlo, R. A. L, Rodrigues, J. J. P. C., Solic, P., and Carvalho, A. 2020. "A Preference-Based Demand Response Mechanism for Energy Management in a Microgrid," *Journal of Cleaner Production* (255), 120034.

Feldman, R. 2013. "Techniques and Applications for Sentiment Analysis," *Communications of the ACM* (56:4), pp. 82-89.

Gambhir, M. and Gupta, V. 2017. "Recent Automatic Text Summarization Techniques: A Survey," *Artificial Intelligence Review*, (47:1), pp. 1-66.

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.

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.

Koneru, A., Bhavani, N. B. N. S. R., Rao, K. P., Prakash, G. S., Kumar, I. P., and Kumar, V. V. 2018. "Sentiment Analysis on Top Five Cloud Service Providers in the Market," in *Proceedings of the 2nd International Conference on Trends in Electronics and Informatics*, pp. 293-297.

Li, N. and Wu, D. D. 2010. "Using Text Mining and Sentiment Analysis for Online Forums Hotspot Detection and Forecast," *Decision Support Systems* (48:2), pp. 354-368.

Manning, C. and Schutze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT press.

Pevec, D., Kayser, M., Babic, J., Carvalho, A., Ghiassi-Farrokhfal, Y., and Podobnik, V. 2017. "A Computational Framework for Managing Electric Vehicle Charging Infrastructure," in *Proceedings of the 9th International Exergy, Energy and Environment Symposium*, pp. 14-17.

Pevec, D., Babic, J., Kayser, M. A., Carvalho, A., Ghiassi-Farrokhfal, Y., and Podobnik, V. 2018. "A Data-Driven Statistical Approach for Extending Electric Vehicle Charging Infrastructure," *International Journal of Energy Research* (42:9), 3102-3120.

Qaisi, L. M. and Aljarah, I. 2016. "A Twitter Sentiment Analysis for Cloud Providers: A Case Study of Azure vs. AWS," in *Proceedings of the 7th International Conference on Computer Science and Information Technology*, pp. 1-6.

Tambe, P., Cappelli, P., and Yakubovich, V. 2019. "Artificial Intelligence in Human Resources Management: Challenges and a Path Forward," *California Management Review* (61:4), pp. 15-42.

Valdivia, A., Luzón, M. V., and Herrera, F. 2017. "Sentiment Analysis in Tripadvisor," *IEEE Intelligent Systems* (32:4), pp. 72-77.

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.

Wallach, H. M. 2006. "Topic Modeling: Beyond Bag-of-Words," In *Proceedings of the 23rd International Conference on Machine Learning,* pp. 977-984.

Weng, B., Lu, L., Wang, X., Megahed, F. M., and Martinez, W. 2018. "Predicting Short-Term Stock Prices Using Ensemble Methods and Online Data Sources," *Expert Systems with Applications* (112), pp. 258-273.

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

Winkler, R. L. and Clemen, R. T. 2004. "Multiple Experts vs. Multiple Methods: Combining Correlation Assessments," *Decision Analysis* (1:3), pp. 167–176.

Winkler, R. L., Grushka-Cockayne, Y., Lichtendahl Jr, K. C., and Jose, V. R. R. 2019. "Probability Forecasts and Their Combination: A Research Perspective," *Decision Analysis* (16:4), pp. 239-260.

Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., and Morency, L. P. 2013. "YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context," *IEEE Intelligent Systems* (28:3), pp. 46-53.

Yang, P., Hwa Yang, Y., B. Zhou, B., and Y. Zomaya, A. 2010. "A Review of Ensemble Methods in Bioinformatics," *Current Bioinformatics* (5:4), pp. 296-308.

Zhang, L., Wang, S., and Liu, B. 2018. "Deep Learning for Sentiment Analysis: A Survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (8:4), e1253.