Detecting and Quantifying Political Bias and Credibility in Written Communications using an LSTM-Based Deep Neural Network

Emergent Research Forum (ERF) Paper

Daniel S. Soper
Department of Information Systems & Decision Sciences
California State University, Fullerton
dsoper@fullerton.edu

Abstract

Virtually every organization uses writing in one form or another to communicate with its internal and external stakeholders and with the public at large. Although ideally all of an organization's written communications would be evaluated for political bias and credibility prior to public release, a notable constraint currently limits the feasibility of such activities. Specifically, the volume of written communications produced by many organizations is so large that accurately screening all of those communications for credibility and political bias would be a prohibitively time-consuming, expensive, and labor-intensive enterprise. In light of this constraint, the current project seeks to evaluate whether and to what extent a deep neural network can learn to automatically identify and accurately quantify political bias and credibility in written documents. The preliminary results reported in this emergent research forum paper are highly promising, and indicate that a deep neural network can achieve human-level performance in rating written documents for their degrees of credibility and political bias. The implications of these findings for companies and researchers are discussed, and next steps for the project are presented.

Keywords

Text analytics, political bias, credibility, artificial intelligence, deep neural networks, competitive advantage

Introduction

During the past few years, many questions have been raised about political bias and credibility in the news articles published by mainstream media companies (Barnidge et al. 2020; Elejalde et al. 2018; Lazaridou and Krestel 2016). Large media companies notwithstanding, virtually every organization uses writing in one form or another to communicate with its internal and external stakeholders and with the public at large. From websites to email to quarterly reports, and from press releases to advertising to social media, written communications permeate almost all aspects of modern business (Newman 2015). It is now well established that when communicating in writing, the use of certain words, phrases, clauses, abbreviations, or references can affect perceptions of the message's degree of credibility or truthfulness in the mind of the reader (Schmied and Dheskali 2019). Similarly, it is also now well established that the way in which a message is written can serve to connect or associate the message with a particular political ideology, and that such associations may provoke a negative response among readers who hold alternative political perspectives (Elejalde et al. 2018).

Judgments about credibility and political bias are, of course, not isolated to the messages generated by large media companies. Indeed, any organization's written communications (e.g., its marketing materials, press releases, website content, etc.) may, intentionally or unintentionally, be imbued with political bias or other cues that can affect readers' perceptions in terms of message credibility or impartiality. If the recipient of the message is a current or potential customer or other stakeholder, then such perceptions may, for better or worse, influence the recipient's view of the underlying organization itself, thus affecting the organization's prospects for success. This perspective is echoed in the literature by a recent survey in which consumers' perceptions and patronization patterns of many Fortune 500 companies were found to be linked to their political views (Matthews 2016), and by several highly cited studies that found strong relationships between customer loyalty and perceptions of credibility (Nguyen and Leblanc 2001; Sweeney and Swait 2008).

In light of the above, managers should be highly interested in knowing the extent to which their organization's written communications are viewed as credible or politically biased, thus allowing such communications to be evaluated in the context of the organization's strategic objectives and public image, and enabling corrective action to be taken if necessary. For example, if a CEO wants her company's latest quarterly report to be perceived as highly credible and politically neutral, but an analysis reveals that the report is written in a conservatively biased manner that is only moderately credible, then the CEO can intervene to correct the problem prior to releasing the report to the public. These concepts can be conveniently conceptualized using a geometric framework such as that illustrated in Figure 1. As shown in the figure, adopting such a framework allows the degrees of credibility and political bias in any written document to be readily described as a location within a two-dimensional Euclidean space.

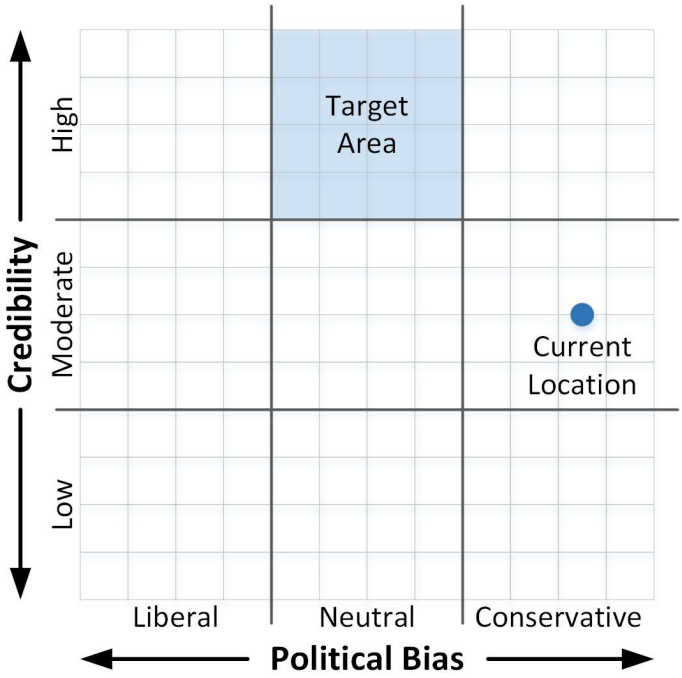

Figure 1. A geometric representation of political bias and credibility in written documents.

Although ideally all of an organization's written communications would be evaluated for political bias and credibility, a notable constraint currently limits the feasibility of such activities. Specifically, the volume of written communications produced by many organizations is so large that accurately screening all of those communications for credibility and political bias would be a prohibitively time-consuming, expensive, and labor-intensive enterprise. In light of this constraint and the broader considerations described above, an automated solution to this problem would be highly desirable. The primary goal of the current project is thus to gain insights into the following research question:

> *Can a deep neural network learn to automatically identify and accurately quantify political bias and credibility in written documents?*

This project currently relies on human evaluations of just 250 news articles as a basis for training the deep neural network, and is hence still very much a work in progress. Nevertheless, the preliminary results reported in this emergent research forum paper are highly promising, and stand independently as a proof of concept. As this project matures, it is anticipated that it will yield free and open-source methods and tools that can be used by organizations and researchers alike to evaluate and study political bias and credibility in large corpora of text. In so doing, the results of this project will contribute to advancing the ongoing debate about political bias and credibility in the age of ubiquitous information, and will ultimately help to create a more informed public.

Methodology

The data for the current project were 250 news articles written in English published by the 50 most popular news organizations in America (NBC, CBS, Fox News, CNN, Washington Post, Forbes, Huffington Post, NPR, Breitbart, etc.), with five articles originating from each news source. Articles were rated for political bias and credibility by workers participating in Amazon Mechanical Turk's Human Intelligence Tasks. Using the platform's qualifications technology, only adult, American subjects with a known U.S. political affiliation (i.e., liberal, independent, or conservative) were allowed to participate as judges. Each of the 250 news articles was rated by a politically balanced panel of six judges (two each of liberal, independent, and conservative), yielding 1,500 total ratings for each dimension. Each judge was tasked with rating five articles, and as such a total of 300 judges participated in the rating task (100 each of liberal, independent, and conservative). Political bias and credibility were rated using standard "slider" controls, which ranged from -1.0 to +1.0. With respect to political bias, the rating scale was anchored at *-1.0 = very liberal* and *+1.0 = very conservative*. With respect to credibility, the rating scale was anchored at *-1.0 = very low* and *+1.0 = very high*. The means of the judges' ratings for each article's degrees of political bias and credibility were taken as the true location coordinates for the article in the Euclidean space depicted in Figure 1. Average within-article standard error for the judges' ratings was observed to be 0.167 (or 8.33%).

After obtaining human ratings for the credibility and political bias of each article, the collection of articles was randomly shuffled, after which each article was assigned to one of 10 equally sized groups in support of later *k*-fold cross validation. The text of each article was then subjected to standard preprocessing to remove excess linefeeds, special characters, etc., after which each article was processed to identify and handle contractions, acronyms, and US/UK English language spelling variations. Next, each article was run through the Python NLTK natural language toolkit's sentence detector in order to subdivide articles into their constituent sentences (Bird et al. 2009). Each sentence was then processed using Google's Universal Sentence Encoder in order to transform it into a vector of 512 numbers (Cer et al. 2018), with each vector representing a location within high-dimension Euclidean space. These vector representations of the sentences within each article served as the primary input into the deep neural network described below.

After completing the preliminary text processing, a bidirectional long short-term memory (LSTM) deep neural network was implemented in Python using the Google TensorFlow library (Abadi et al. 2016). A deep bidirectional LSTM architecture was chosen because such models have been shown to achieve state-of-the-art performance on a variety of natural language processing tasks (Chiu and Nichols 2016; Huang et al. 2015). Rather than testing a single model, a self-avoiding random search algorithm was used to evaluate 1,000 different models, with each model having a unique combination of structural characteristics and hyperparameter settings (Soper 2018). Since an exhaustive search was computationally infeasible, this approach was adopted with a view toward identifying the parameters that would yield the best-performing model given a limited amount of time to search through the parameter space. Each model was trained using the Adam optimization algorithm (Kingma and Ba 2014), with the dropout regularization technique being used

to avoid overfitting and ensure model stability (Srivastava et al. 2014). Finally, each model's performance was evaluated using *k*-fold cross validation (k = 10), with the mean absolute error of the model's credibility and political bias predictions being used as the loss function. The overall methodological architecture for the project is depicted in Figure 2 below.

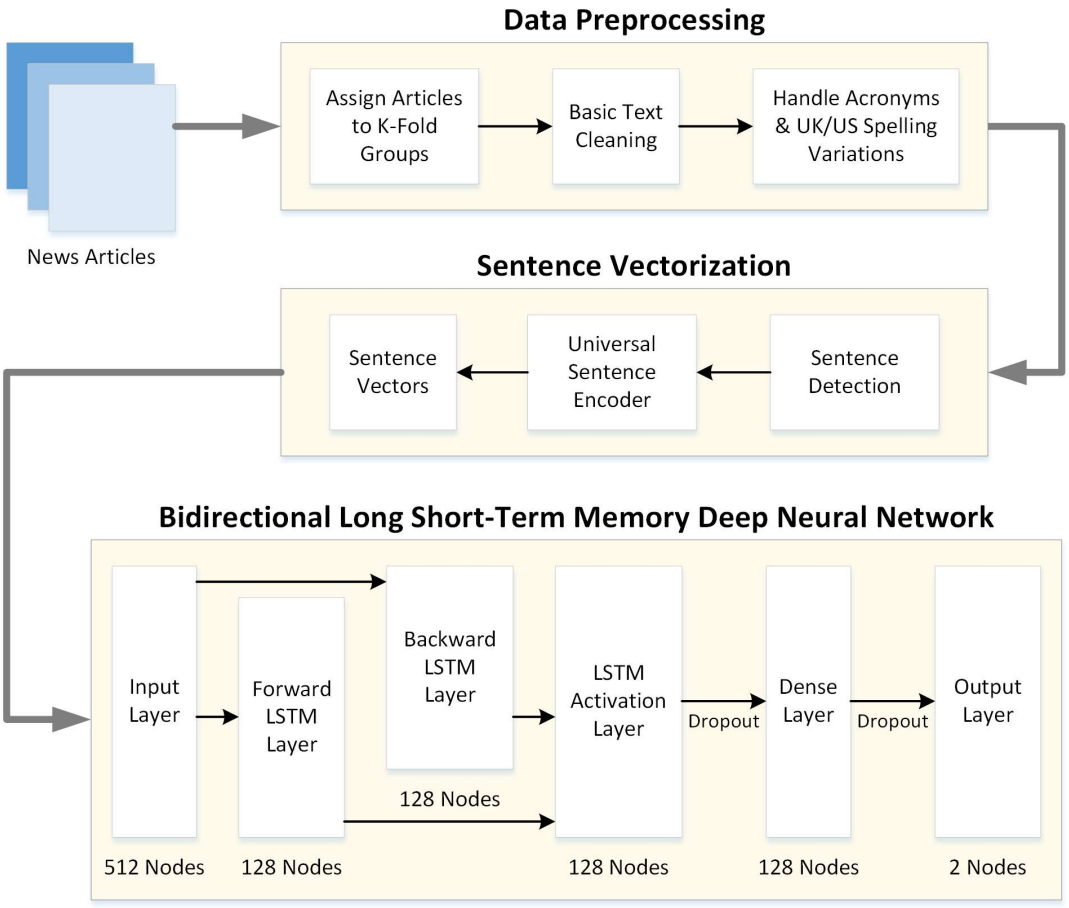

Figure 2. Overall methodological architecture.

Preliminary Results and Discussion

The minimum mean absolute validation error for ratings of credibility and political bias that was attained during testing and validation of the candidate models was 0.174. This result was achieved with the model structure shown in Figure 2 using an initial learning rate of 0.0003, a minibatch size of 16 articles, and a dropout rate of 0.50. Given that the scales for both credibility and political bias ranged from -1.0 to +1.0, this indicates that on average, predicted document locations generated by the bidirectional LSTM deep neural network were within 8.70% of each document's true location within the Euclidean space. Recalling that the overall standard error for human judges was 0.167 (or 8.33%), it can be concluded that on average, the ratings assigned by the deep neural network were statistically indistinguishable from those assigned by human judges

(Levene's test, p = n.s.). An illustrative example of the deep neural network's average predictive accuracy is shown in Figure 3.

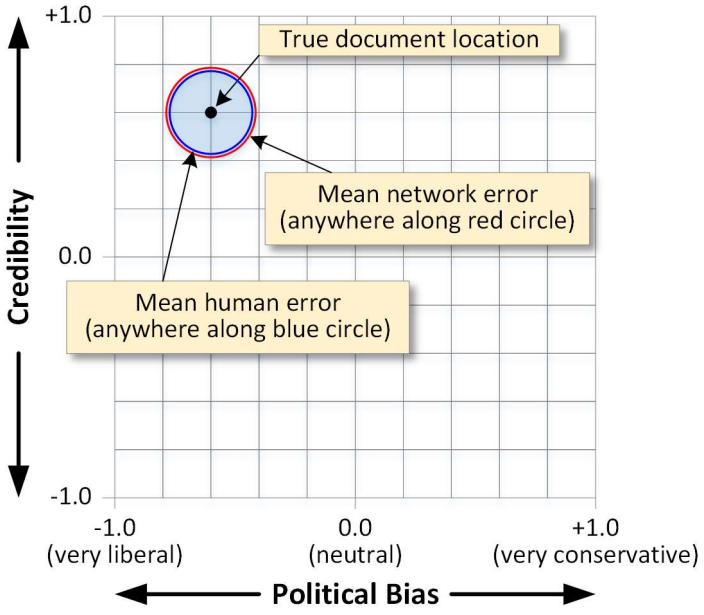

Figure 3. An illustrative example of average model predictive accuracy.

Despite being trained on a sample of just 250 news articles, the performance of the model described above is highly promising, insofar as it was able to automatically rate the degrees of credibility and political bias in written documents with a level of accuracy that is indistinguishable from a typical human judge. This is quite remarkable given that both credibility and political bias are human constructs that do not otherwise exist in the natural world. Although the project described in this emerging research forum paper is still in its infancy, and although the results reported above are hence preliminary, they nevertheless stand independently as evidence that deep neural networks can indeed be trained to accurately identify and quantify subtle and nuanced properties of written documents. Put differently, when we as humans read a document, we perceive characteristics such as humor, sarcasm, trustworthiness, political bias, etc. that exist *outside* of the document's primary content. It is highly encouraging to note that machines, too, can learn to perform this feat, and that projects such as this will soon yield tools that organizations and researchers can use to quickly and accurately identify and measure these subtle and nuanced characteristics in large corpora of text.

Limitations and Next Steps

The results observed in the current study clearly indicate that LSTM-based deep neural networks can be trained to accurately identify the degrees of credibility and political bias in written documents. Nevertheless, there are several limitations to the study that must be acknowledged. First, the input data for the current study were all news articles published by large media companies. Since the input data used to train a neural network play an important role in determining the generalizability of the network's predictions, this project will develop and

use a much larger corpus of documents of many different types and of from many additional sources as it matures. Including text from corporate reports, social media posts, press releases, advertising campaigns, etc. will undoubtedly yield a more robust, general-purpose tool for identifying and quantifying the degrees of credibility and political bias in written communications.

In its current state, this project is also limited insofar as each article that was used to train and evaluate the deep neural network was read and rated for its degrees of credibility and political bias by just three human judges. Although efforts were taken to ensure that each article was assessed by a politically balanced panel of judges, the mean scores assigned to each article are not as statistically stable as they otherwise would be if they had been derived from a larger pool of independent ratings. As such, the number of human judges who evaluate and rate each article will be increased by an order of magnitude as this project matures. With a larger sample of ratings for each article, the influence of any single judge will be vastly diminished, thus ensuring that the deep neural network learns to replicate the collective wisdom of a large group of human judges, rather than a smaller, comparatively unstable group.

Implications and Concluding Remarks

A deep neural network with the ability to accurately identify and quantify the degrees of credibility and political bias in written text has manifold implications for both organizations and researchers alike. Companies and governmental entities commonly convey information to their stakeholders by means of written communications, and based on these organizations' strategic objectives, it may or may not be desirable for such communications to espouse a particular political ideology or a particular level of trustworthiness. In either case, tools of the sort described in this paper could prove extremely useful in evaluating whether a company's written communications are aligned with its political ideology or goals. Just as sentiment analysis has proven to be of immense value in many different areas of inquiry, a tool that allows for objective, accurate measurement of the degrees of credibility and political bias in written text will open many fascinating new pathways for researchers, and may even shed some light on divisive social issues, such as the longstanding claim that there is a pervasive liberal bias in the mainstream media. These specific considerations notwithstanding, tools based on machine learning techniques clearly have great potential to extract new and interesting insights from written text that might otherwise remain hidden. As with the project described in this emergent research forum paper, many of these tools and techniques are still in their infancy. Nevertheless, they are evolving and maturing quickly, and can be expected to become a common part of both scientific research and organizational decision-making in the very near future.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., and Isard, M. 2016. "Tensorflow: A System for Large-Scale Machine Learning," *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265-283.

Barnidge, M., Gunther, A.C., Kim, J., Hong, Y., Perryman, M., Tay, S.K., and Knisely, S. 2020. "Politically Motivated Selective Exposure and Perceived Media Bias," *Communication Research* (47:1), pp. 82-103.

Bird, S., Klein, E., and Loper, E. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. " O'Reilly Media, Inc.".

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., and Tar, C. 2018. "Universal Sentence Encoder for English," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 169-174.

Chiu, J.P., and Nichols, E. 2016. "Named Entity Recognition with Bidirectional Lstm-Cnns," *Transactions of the Association for Computational Linguistics* (4), pp. 357-370.

Elejalde, E., Ferres, L., and Herder, E. 2018. "On the Nature of Real and Perceived Bias in the Mainstream Media," *PloS one* (13:3).

Huang, Z., Xu, W., and Yu, K. 2015. "Bidirectional Lstm-Crf Models for Sequence Tagging," *arXiv preprint arXiv:1508.01991*).

Kingma, D.P., and Ba, J. 2014. "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*).

Lazaridou, K., and Krestel, R. 2016. "Identifying Political Bias in News Articles," *Bulletin of the IEEE TCDL* (12).

Matthews, C. 2016. *Here Are the Fortune 500 Companies Liberals and Conservatives Hate the Most*. New York, NY: Fortune.

Newman, A. 2015. *Business Communication: In Person, in Print, Online*. Stamford, CT: Cengage Learning.

Nguyen, N., and Leblanc, G. 2001. "Corporate Image and Corporate Reputation in Customers' Retention Decisions in Services," *Journal of retailing and consumer services* (8:4), pp. 227-236.

Schmied, J., and Dheskali, J. (eds.). 2019. *Credibility, Honesty, Ethics & Politeness in Academic and Journalistic Writing*. Göttingen, Germany: Cuvillier Verlag.

Soper, D. 2018. "On the Need for Random Baseline Comparisons in Metaheuristic Search," *Proceedings of the 51st Hawaii International Conference on System Sciences*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *The journal of machine learning research* (15:1), pp. 1929-1958.

Sweeney, J., and Swait, J. 2008. "The Effects of Brand Credibility on Customer Loyalty," *Journal of retailing and consumer services* (15:3), pp. 179-193.