Proceedings of the 2023 Pre-ICIS SIGDSA Symposium

Special Interest Group on Decision Support and Analytics (SIGDSA)

12-10-2023

Extracting Business Intelligence with Data- Centric Neural Network Language Models

Wingyan Chung
University of Texas at Tyler

Extracting Business Intelligence with Data-Centric Neural Network Language Models

Research-in-Progress – Extended Abstract

Wingyan Chung
Department of Computer Science, Soules College of Business
The University of Texas at Tyler, Tyler, Texas, U.S.A.
wchung@uttyler.edu

Abstract

Artificial intelligence (AI) technologies offer promising opportunities for detecting market movements from news and textual data. However, many organizations fail to create or meet any data quality standards that are key to AI development. In this research, we developed and applied data-centric neural network (NN) language models to extracting business intelligence (BI) factors automatically from textual news articles of high-tech companies. Our methodology iteratively improves the datasets for use in building NN language models. Experimental results confirm that our approach helped to increase the predictive performance across different BI categories. The approach makes the NN language models more transparent to managers, BI specialists, and users through the iterative refinement and query search process. The research produces new information systems artifacts in the form of new methodology, new NN models, and new empirical findings of using the models to extract BI from textual news.

Keywords

Data-centric AI, machine learning, neural network language models, business intelligence.

Introduction

Artificial intelligence (AI) technologies offer novel distinctive opportunities to gather timely and relevant business intelligence (Chen et al. 2012) due to their versatilities in pattern recognition and natural language understanding (Benbya et al. 2021). AI technologies are estimated to produce $13 trillion of GDP growth globally by 2030 (Bughin et al. 2018). Among many applications of AI, automatically detecting company movements and potential threats from textual news articles can provide managers with business intelligence (BI) (Norris 2020), defined as the result of "acquisition, collation, analysis and exploitation of information in business" (Chung et al. 2005). These articles report recent trends and incidents posing risks and challenges related to firms' operations, economic and strategic environments, legal matters, and technologies. For example, in the fast-changing industry of consumer electric vehicle sales, news about rivalry movements can change the competitive landscape dramatically. The news statement "*Tesla has denied that the latest shutdown of its Model 3 production line is related to safety fears or the need to move away from automation*" (Bradshaw 2018) indicates Tesla's operational risk due to product safety and over-reliance on automation.

Although managers and developers understand intuitively that data quality is important in AI development, most organizations fail to create or meet any data quality standards (Sambasivan et al. 2021). Effort to improve data quality is viewed as "operational" relative to the more preferred work of model development. This lack of attention to data can lead to problematic AI model deployment in high-stake domains such as fraud detection, autonomous driving, and cancer treatment. The models may miss valuable business opportunities, fail to detect crimes, and even cause a loss of lives (Strickland 2019). Different from traditional approaches, data-centric AI (DCAI) is a movement that improves AI models by increasing data quality to enable the models to perform more accurately (Ng et al. 2021). While much work on DCAI has focused on image recognition, relatively little work has been done in text and natural language processing that could benefit BI extraction from news articles. Neural-network (NN) language models, a promising machine learning (ML) approach, have been shown to be robust and accurate in text classification (Buddana et al. 2021; Schmidhuber 2015). Their application to extracting BI factors has not been available.

In this research, we aim to develop a DCAI approach to business intelligence extraction and categorization. Specifically, we developed data-centric neural network language models to extract BI factors automatically from textual news articles, and applied the models to discovering BI factors of high-tech companies whose activities are widely reported in the news. The research seeks to answer these questions: (1) How can a DCAI approach to improving data quality of news articles having imbalanced distribution of BI categories be developed? (2) How can the approach enhance the performance of neural-network language models in extracting BI factors from news articles about high-tech companies? (3) What is the performance of the approach as measured by ML evaluation metrics?

Data-Centric Neural Network Language Models

From our literature review, we find scarce prior research using DCAI in BI extraction. Existing benchmarks and annotation efforts use non-domain-specific data as target, thus not generalizable to many BI applications. To address the gaps, we propose a novel DCAI methodology for iteratively improving datasets for use in building neural network language models for BI extraction from textual news articles. Based on a design science research paradigm (Hevner et al. 2004), our methodology consists of three generic, iterative steps of BI extraction and validation using NN language modeling:

(1) First, the sources of data are identified and used to develop suitable labeled datasets for BI extraction. Company-specific textual news articles (on Apple and Tesla) are the target data source due to their widespread usage in business to understand and assess risk and business intelligence. Manual annotation based on a five-category taxonomy of company BI factors (Chung 2014) is used to label selected (e.g., first five) sentences as BI factors to provide input to train ML models. To minimize subjectivity and bias, the annotator followed a systematic procedure (as documented in (Chung 2014)) to categorize the BI factors and achieved a 54% agreement ratio in a validation study with 25 graduate students who annotated BI factors selected randomly from textual news reports. This agreement ratio is significantly higher than that of a random 5-category matching (that would result in an average of 20% agreement only).

(2) Second, shallow neural network language models are built to learn from the input datasets (training sets) and to predict on new datasets (testing sets). Each word in an input sentence is modeled as an embedding vector using the GloVe method (Pennington et al. 2014). Many possible architectural options are excluded intentionally (e.g., attention mechanism, dropout, word ordering, additional layers, convolution and recurrent network, etc.) because this study is not focused on exploiting model capabilities.

(3) Third, the evaluation of the ML models followed an iterative process analyzing the outcomes of the performance and then identifying categories and models that need improvement. Then, new queries are used to collect additional data to improve existing datasets to improve the model performance. The new collection allows human intervention in the improvement of datasets to be used for training the algorithm.

Experimental Findings

The training and testing accuracies found in the experiments range from 33.33% to 87.17%, which are all significantly higher than 20% that would be the average accuracy if a random approach is used in a five-class categorization. In general, both training and testing accuracies obtained in datasets that use a higher-dimension representation (GloVe 42B, dim=300) are higher than those in datasets using a lower-dimension representation (GloVe 6B, dim=50). *The better results from a higher dimension of word representation confirm the hypotheses that these richer and more expressive representation used in the datasets improved the predictive accuracies.*

Two rounds of data collections and model building were done in our experiments. Round 1 used 500 articles on Apple and 498 articles on Tesla. Round 2 added 649 articles on Apple and 634 articles on Tesla. A total of 8 models were built (2 dimensions × 2 rounds × 2 companies). To examine the per-class performance of Round 1's models using different dimensional representation of words, we calculated the precision, recall, and F-score and their average values. Both companies' high-dimensional models achieved higher average recalls and F-scores. *The results confirm the hypothesis that high-dimensional models generally enable higher predictive performance across most BI categories.* When using the models to predict only on Round

2 test sets, we find that *models built in Round 2 generally outperformed models built in Round 1 when evaluated using the same test sets, thus confirming the second hypothesis*. Therefore, we believe that the methodology is effective in improving ML models predictive performance by increasing data quality and by allowing critical human judgement to support the ML training process.

The encouraging results highlight the promise of using the iterative, data-centric approach to improve NN language models for BI extraction. Managers and developers can use the higher dimension representation of word embedding to increase predictive accuracies. Unlike traditional NN models (often considered a “black box”), our data-centric approach allows human-in-the-loop judgement and intervention to improve labeling and to increase data quality that lead to better predictive performance.

Conclusion

In this research, we developed and applied data-centric neural network (NN) language models to extracting business intelligence (BI) factors automatically from textual news articles of high-tech companies. The research produces new information systems artifacts in the form of new DCAI methodology, new ML models, and new empirical findings of using ML models to extract BI from textual news. Future directions include applying the methodology to more complex NN language modeling using deep learning and larger datasets, reducing algorithmic biases in ML processes (labeling, modeling training, validation, deployment), and designing human-centric tools to support both explainable AI and modifiable AI in NN model development.

References

Benbya, H., Pachidi, S., and Jarvenpaa, S. 2021. "Artificial Intelligence in Organizations: Implications for Information Systems Research," *Journal of the Association for Information Systems* (22:2).

Bradshaw, T. 2018. "Tesla Denies Model 3 Production Line Shutdown Is Safety-Related," in: *Financial Times*. https://www.ft.com/content/509c934a-41e1-11e8-803a-295c97e6fd0b: Financial Times.

Buddana, H. V. K. S., Kaushik, S. S., Manogna, P. V. S., and Ps, S. K. 2021. "Word Level LSTM and Recurrent Neural Network for Automatic Text Generation," *2021 International Conference on Computer Communication and Informatics (ICCCI), 27-29 Jan. 2021*, Piscataway, NJ, USA: IEEE, pp. 1-4.

Bughin, J., Seong, J., Manyika, J., Chui, M., and Joshi, R. 2018. "Notes from the AI Frontier: Modeling the Impact of AI on the World Economy," McKinsey Global Institute.

Chen, H., Chiang, R., and Storey, V. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* (36:4), pp. 1165-1188.

Chung, W. 2014. "BizPro: Extracting and Categorizing Business Intelligence Factors from Textual News Articles," *International Journal of Information Management* (34:2), pp. 272-284.

Chung, W., Chen, H., and Nunamaker, J. F. 2005. "A Visual Framework for Knowledge Discovery on the Web: An Empirical Study on Business Intelligence Exploration," *Journal of MIS* (21:4), pp. 57-84.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *Management Information Systems Quarterly* (28:1), pp. 75-105.

Ng, A., Aroyo, L., Diamos, G., Coleman, C., Reddi, V. J., Vanschoren, J., Wu, C.-J., Zhou, S., and He, L. 2021. "Workshop on Data Centric AI," *Thirty-fifth Conf. on Neural Information Processing Systems*

Norris, M. 2020. *The Value of AI-Powered Business Intelligence*. O'Reilly Media, Inc.

Pennington, J., Socher, R., and Manning, C. D. 2014. "Glove: Global Vectors for Word Representation," *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar: Association for Computational Linguistics, pp. 1532-1543.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. 2021. "“Everyone Wants to Do the Model Work, Not the Data Work”: Data Cascades in High-Stakes AI," in *Proceedings of the 2021 Chi Conference on Human Factors in Computing Systems*. ACM, p. Article 39.

Schmidhuber, J. 2015. "Deep Learning in Neural Networks: An Overview," *Neural Networks* (61), pp. 85-117.

Strickland, E. 2019. " IBM Watson, Heal Thyself: How IBM Overpromised and Underdelivered on AI Health Care," *IEEE Spectrum* (56:4), pp. 24-31.