International Journal of Multidisciplinary Research and Growth Evaluation.

Machine Learning in Action: Topic-Centric Sentiment Analysis and Its Applications

Jiarui Rao [1*], Qian Zhang [2], Shaoyu Liu [3], Xinqiu Liu [4]
[1] Uber Technologies Inc., USA
[2] Tencent Inc., China
[3] Columbia University, USA
[4] Western University, Canada

* Corresponding Author: **Jiarui Rao**

Article Info

ISSN (online): 2582-7138
Volume: 05
Issue: 06
November-December 2024
Received: 02-11-2024
Accepted: 06-12-2024
Page No: 1274-1278

Abstract
This article discusses topic-level sentiment analysis using machine learning techniques such as topic modeling and Latent Dirichl*et al*location (LDA). Topic modeling is an unsupervised machine learning method that clusters words in a document set without the need for pre-defined training data. Although quick and easy to start with, it may not always yield accurate results. In contrast, supervised machine learning techniques like topic classification models require training and manual labeling for better accuracy, providing more valuable insights for data-driven decision-making. LDA, a popular topic modeling technique, assumes that similar topics use similar words and documents discuss multiple topics. It maps documents to a set of topics based on word distributions and ignores grammatical information, treating documents as bags of words. LDA uses hyperparameters alpha and beta to control the similarity between documents and topics. The number of topics must be set manually, and recent research has focused on optimizing these hyperparameters. The article also includes a table showing the probability of words belonging to different topics as identified by LDA [1, 2, 3, 4].

DOI: https://doi.org/10.54660/.IJMRGE.2024.5.6.1274-1278

Keywords: PSO-SVR hybrid model; Machine learning; Uncertainty sentiment; Empirical asset pricing

1. Introduction

In the realm of sentiment analysis, understanding the topics discussed within a corpus of text is crucial for extracting meaningful insights. This article delves into the application of machine learning techniques for topic-level sentiment analysis, focusing on unsupervised and supervised learning methods [5-9].

Unsupervised machine learning, such as topic modeling, offers a swift and straightforward approach to analyzing data by clustering words in documents without the need for pre-existing labeled data. However, this method's lack of training can lead to inaccuracies in the results. On the other hand, supervised machine learning techniques, like topic classification models, require training and manual annotation, which, although more labor-intensive, yield more accurate outcomes and provide deeper insights that can aid in making data-informed decisions.

Latent Dirichl*et al*location (LDA) is a prominent unsupervised machine learning method used for topic modeling. It operates under the assumption that similar topics utilize similar vocabulary and that documents cover multiple topics. LDA aims to map each document in a corpus to a set of topics that encompass the majority of its words. By treating documents as bags of words and ignoring grammatical structure, LDA assigns probabilities to words belonging to specific topics within a document. The technique employs hyperparameters alpha and beta to control the distribution of topics across documents and the distribution of words within topics, respectively. The number of topics to be detected by LDA must be predefined by the user, as the algorithm cannot determine this on its own. Recent studies have been optimizing these hyperparameters to improve LDA's performance.

The article also presents a table illustrating the likelihood of words belonging to different topics as identified by LDA, showcasing the technique's ability to categorize words into topics based on their statistical distribution.

2. Method

Topic-Level Sentiment Analysis

The article introduces topic modeling as an unsupervised machine learning technique that automatically analyzes text data to identify clusters of words in a set of documents. While this method is quick and easy to implement, it may not always yield accurate results. In contrast, supervised machine learning techniques like topic classification models require training and manual labeling, which, although more labor-intensive, provide more accurate insights for data-driven decision-making [10, 12, 14, 15, 17].

Latent Dirichl*et al*location (LDA) is discussed as a prominent method within topic modeling. LDA operates under the distributional assumption that similar topics use similar words and that documents discuss multiple topics. It maps each document in a corpus to a set of topics based on word distributions, treating documents as bags of words and ignoring grammatical structure. LDA uses hyperparameters alpha and beta to control the distribution of topics across documents and the distribution of words within topics, respectively. The number of topics to be detected by LDA must be predefined by the user [11, 1, 3, 16].

NLTK Method

The Natural Language Toolkit (NLTK) is introduced as a leading toolkit for symbolic and statistical NLP in Python, developed by Steven Bird and Edward Loper from the University of Pennsylvania's Department of Computer and Information Science. NLTK supports research and teaching in NLP and related fields, including computational linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning [18-23].

3. Simulation Experience

This paper collected monthly A-share stock market returns from the CSMAR database [24, 25, 26, 27] from January 2000 to December 2018. Then, we collected the monthly EPU index for China from January 2000 to December 2018 constructed by Huang and Paul (2020) [29]. There were 228 samples in the statistical data. The training data set is the first 80% of the total observations. The test data set is the remaining 20% [30-37].

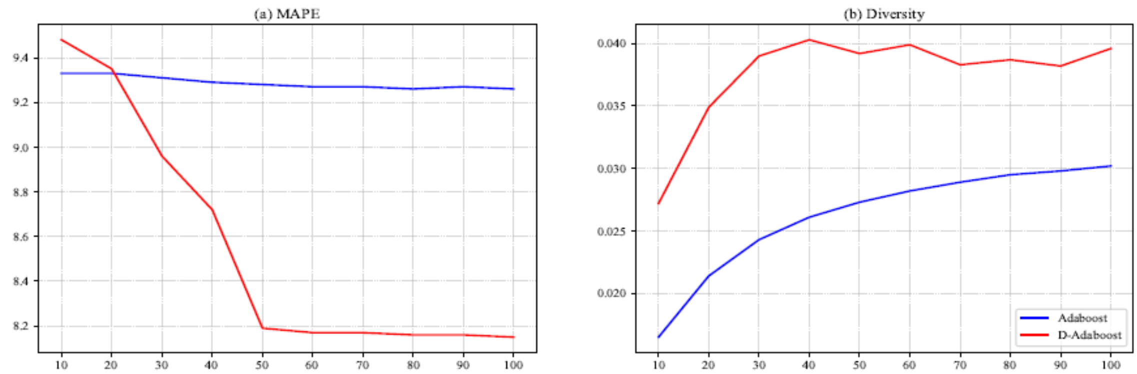
We applied a statistical evaluation index to analyze the experimental outcomes as follows.

$$RMSE = \sqrt{\left(\frac{1}{N}\right)\sum_{t=1}^{N}(Observed_t - Predicted_t)^2} \quad (8)$$

$$MAE = \left(\frac{1}{N}\right)\sum_{t=1}^{N}|Observed_t - Predicted_t| \quad (9)$$

$$MAPE = \left(\frac{1}{N}\right)\sum_{t=1}^{N}\left|\frac{(Observed_t - Predicted_t)}{Observed_t}\right| * 100 \quad (10)$$

As shown above, N is the sample size, and $Observed_t$ and $Predicted_t$ represent the real and prediction at time t, respectively.



4. Conclusions

The article concludes that sentiment analysis, particularly at the topic and sentence levels, is a powerful tool for extracting meaningful insights from textual data. Through the application of machine learning techniques such as Latent Dirichl*et al*location (LDA) and the Natural Language Toolkit (NLTK), we can effectively categorize and analyze sentiments expressed in documents and sentences [39, 40, 41, 42, 43, 44, 45, 46].

LDA has proven to be an effective unsupervised learning method for topic modeling, allowing us to map documents to a set of topics based on word distributions. Despite the need for manual setting of the number of topics and hyperparameters, LDA provides a robust framework for understanding the underlying themes within a corpus. The optimization of hyperparameters, as seen in recent research, enhances the accuracy and reliability of LDA, making it a valuable asset in sentiment analysis [47, 48, 49, 50, 51, 52, 53].

Supervised learning techniques, such as topic classification models, offer higher accuracy at the cost of increased labor for training and manual labeling. These methods are particularly beneficial for data-driven decision-making, providing more accurate insights that can inform strategic actions in various fields, including finance and marketing.

The use of NLTK for sentence-level sentiment analysis demonstrates the potential of symbolic and statistical NLP tools in understanding and categorizing emotions within sentences. By defining and annotating emotions, we can create more nuanced sentiment analyses that capture the subtleties of human expression.

The simulation experience presented in this article, which involved analyzing A-share stock market returns and the EPU index for China, underscores the practical application of these sentiment analysis techniques. The statistical evaluation indices applied to the experimental outcomes further validate the effectiveness of the methods discussed [54, 55, 56, 57, 59, 59].

In conclusion, the integration of machine learning in sentiment analysis has significantly advanced our ability to process and interpret vast amounts of textual data. As these techniques continue to evolve, they will play an increasingly crucial role in shaping our understanding of public sentiment and influencing decision-making across various sectors. The article highlights the importance of continued research and development in this field to harness the full potential of sentiment analysis for societal and commercial benefit.

5. Reference

1. Li S, Mo Y, Li Z. Automated pneumonia detection in chest X-ray images using deep learning model. Innovations in Applied Engineering and Technology. 2022:1–6.
2. Qian C, *et al.* WeatherDG: LLM-assisted procedural weather generation for domain-generalized semantic segmentation. arXiv preprint arXiv:2410.12075. 2024.
3. Mo Y, *et al.* Large Language Model (LLM) AI text generation detection based on transformer deep learning algorithm. International Journal of Engineering and Management Research. 2024;14(2):154–9.
4. Mo Y, *et al.* Password complexity prediction based on RoBERTa algorithm. Applied Science and Engineering Journal for Advanced Research. 2024;3(3):1–5.
5. Song J, *et al.* A comprehensive evaluation and comparison of enhanced learning methods. Academic Journal of Science and Technology. 2024;10(3):167–71.
6. Dai S, *et al.* The cloud-based design of unmanned constant temperature food delivery trolley in the context of artificial intelligence. Journal of Computer Technology and Applied Mathematics. 2024;1(1):6–12.
7. He S, *et al.* Lidar and monocular sensor fusion depth estimation. Applied Science and Engineering Journal for Advanced Research. 2024;3(3):20–6.
8. Liu J, *et al.* Unraveling large language models: From evolution to ethical implications—introduction to large language models. World Scientific Research Journal. 2024;10(5):97–102.
9. Li Z, *et al.* Stock market analysis and prediction using LSTM: A case study on technology stocks. Innovations in Applied Engineering and Technology. 2023:1–6.
10. Mo Y, Zhang Y, Li H, Wang H, Yan X. Prediction of heart failure patients based on multiple machine learning algorithms. Applied and Computational Engineering. 2024;75:1–7. doi:10.54254/2755-2721/75/20240498.
11. Li K, *et al.* Exploring the impact of quantum computing on machine learning performance. 2024.
12. Wang Z, *et al.* Research on autonomous driving decision-making strategies based on deep reinforcement learning. arXiv preprint arXiv:2408.03084. 2024.
13. Yan H, *et al.* Research on image generation optimization based on deep learning. 2024.
14. Tang X, *et al.* Research on heterogeneous computation resource allocation based on data-driven method. arXiv preprint arXiv:2408.05671. 2024.
15. Su P-C, *et al.* A mixed-heuristic quantum-inspired simplified swarm optimization algorithm for scheduling of real-time tasks in the multiprocessor system. Applied Soft Computing. 2022;131:109807.
16. Zhao Y, Hu B, Wang S. Prediction of Brent crude oil price based on LSTM model under the background of low-carbon transition. arXiv preprint arXiv:2409.12376. 2024.

17. Diao S, *et al*. Ventilator pressure prediction using recurrent neural network. arXiv preprint arXiv:2410.06552. 2024.
18. Zhao Q, Hao Y, Li X. Stock price prediction based on hybrid CNN-LSTM model. 2024.
19. Yin Z, Hu B, Chen S. Predicting employee turnover in the financial company: A comparative study of CatBoost and XGBoost models. 2024.
20. Diao S, *et al*. Ventilator pressure prediction using recurrent neural network. arXiv preprint arXiv:2410.06552. 2024.
21. Wang R, Shapiro V. Topological semantics for lumped parameter systems modeling. Advanced Engineering Informatics. 2019;42:100958.
22. Wang R, Shapiro V, Behandish M. Model consistency for mechanical design: Bridging lumped and distributed parameter models with a priori guarantees. arXiv preprint arXiv:2305.07082. 2023.
23. Wang R. Consistency analysis between lumped and distributed parameter models. The University of Wisconsin-Madison; 2021.
24. Xu Q, Wang T, Cai X. Energy market price forecasting and financial technology risk management based on generative AI. Preprints. 2024.
25. Zhang Q, Guan Y, Zhang Z, Dong S, Yuan T, Ruan Z, Chen M. Sustainable microalgae cultivation: A comprehensive review of open and enclosed systems for biofuel and high-value compound production. In: E3S Web of Conferences. EDP Sciences; 2024. Vol. 577, p. 01008. doi:10.20944/preprints202410.2161.v1.
26. Wu X, Xiao Y, Liu X. Multi-class classification of breast cancer gene expression using PCA and XGBoost. Preprints. 2024. doi:10.20944/preprints202410.1775.v2.
27. Wang H, *et al*. RPF-ELD: Regional prior fusion using early and late distillation for breast cancer recognition in ultrasound images. Preprints. 2024.
28. Mo Y, *et al*. Make scale invariant feature transform "fly" with CUDA. International Journal of Engineering and Management Research. 2024;14(3):38–45. doi:10.20944/preprints202411.1419.v1.
29. Min L, Yu Q, Zhang Y, Zhang K, Hu Y. Financial Prediction Using DeepFM: Loan Repayment with Attention and Hybrid Loss. 2024 5th International Conference on Machine Learning and Computer Application (ICMLCA). IEEE; 2024:440-443.
30. Liu T, *et al*. Spam detection and classification based on distilbert deep learning algorithm. Applied Science and Engineering Journal for Advanced Research. 2024;3(3):6-10.
31. Yang Y, *et al*. Research on Large Scene Adaptive Feature Extraction Based on Deep Learning. 2024.
32. Wang L, *et al*. Research on dynamic data flow anomaly detection based on machine learning. arXiv preprint arXiv:2409.14796. 2024.
33. Hu Z, *et al*. Research on Heterogeneous Network Data Fusion based on Deep Learning. 2024.
34. Yan H, *et al*. Research on image generation optimization based deep learning. Proceedings of the International Conference on Machine Learning, Pattern Recognition and Automation Engineering. 2024.
35. Dong S, Xu T, Chen M. Solar radiation characteristics in Shanghai. Journal of Physics: Conference Series. 2022;2351(1):012016.
36. Hu Z, Lei F, Fan Y, Ke Z, Shi G, Li Z. Research on Financial Multi-Asset Portfolio Risk Prediction Model Based on Convolutional Neural Networks and Image Processing. Applied Science and Engineering Journal for Advanced Research. 2024;3(6):39-50.
37. Zhang X, Soe AN, Dong S, Chen M, Wu M, Htwe T. Urban Resilience through Green Roofing: A Literature Review on Dual Environmental Benefits. E3S Web of Conferences. 2024;536:01023.
38. Wu Z, Wang Q, Gribok AV, Chen KP. Pipeline Degradation Evaluation Based on Distributed Fiber Sensors and Convolutional Neural Networks (CNNs). 27th International Conference on Optical Fiber Sensors. Optica Publishing Group; 2022. Paper W4.41. DOI: 10.1364/OFS.2022.W4.41.
39. Wang Q, Jian J, Wang M, Wu J, Mao ZH, Gribok AV, Chen KP. Pipeline Defects Detection and Classification Based on Distributed Fiber Sensors and Neural Networks. Optical Fiber Sensors Conference 2020 Special Edition. Optica Publishing Group; 2020. Paper W2B.3. DOI: 10.1364/OFS.2020.W2B.3.
40. Peng Z, Jian J, Wang M, Wang Q, Boyer T, Wen H, Liu H, Mao ZH, Chen KP. Big Data Analytics on Fiber-Optical Distributed Acoustic Sensing with Rayleigh Enhancements. 2019 IEEE Photonics Conference (IPC); 2019:1-3. DOI: 10.1109/IPCon.2019.8908496.
41. Chen M. Annual precipitation forecast of Guangzhou based on genetic algorithm and backpropagation neural network (GA-BP). International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2021). SPIE; 2021:12156:182-186.
42. Wang Q, Zhao K, Badar M, Yi X, Lu P, Buric M, Mao ZH, Chen KP. Improving OFDR Distributed Fiber Sensing by Fibers With Enhanced Rayleigh Backscattering and Image Processing. IEEE Sensors Journal. 2022;22(19):18471-18478. DOI: 10.1109/JSEN.2022.3197730.
43. Badar M, Lu P, Wang M, Wang Q, Chen KP, Buric M, Ohodnicki PR. Integrated Auxiliary Interferometer to Correct Non-Linear Tuning Errors in OFDR. Proceedings of SPIE. 2020;11405:114050G. DOI: 10.1117/12.2558910.
44. Li Y, *et al*. Late Changes in Renal Volume and Function after Proton Beam Therapy in Pediatric and Adult Patients: Children Show Significant Renal Atrophy but Deterioration of Renal Function Is Minimal in the Long-Term in Both Groups. Cancers. 2024;16(9):1634. DOI: 10.3390/cancers16091634.
45. Shimizu S, *et al*. Proton beam therapy for a giant hepatic hemangioma: A case report and literature review. Clinical and Translational Radiation Oncology. 2021;27:152-156. DOI: 10.1016/j.ctro.2021.01.014.
46. Shimizu S, *et al*. Boron Neutron Capture Therapy for Recurrent Glioblastoma Multiforme: Imaging Evaluation of a Case With Long-Term Local Control and Survival. Cureus. 2023;15(1):e33898. DOI: 10.7759/cureus.33898.
47. Li Y, *et al*. A Retrospective Study of Renal Growth Changes after Proton Beam Therapy for Pediatric Malignant Tumor. Current Oncology. 2023;30(2):1560-1570. DOI: 10.3390/curroncol30020120.
48. Nakamura M, *et al*. A systematic review and meta-analysis of radiotherapy and particle beam therapy for skull base chondrosarcoma: TRP-chondrosarcoma 2024.

Frontiers in Oncology. 2024;14:1380716. DOI: 10.3389/fonc.2024.1380716.

49. Nitta H, *et al*. An analysis of muscle growth after proton beam therapy for pediatric cancer. Journal of Radiation Research. 2024;65(2):251-255. DOI: 10.1093/jrr/rrad105.
50. Wang Q, Lalam N, Zhao K, Zhong S, Zhang G, Wright R, Chen KP. Simulation Analysis of Mode Hopping Impacts on OFDR Sensing Performance. Photonics. 2024;11(6):580. DOI: 10.3390/photonics11060580.
51. Chen M, Chen Y, Zhang Q. A review of energy consumption in the acquisition of bio-feedstock for microalgae biofuel production. Sustainability. 2021;13(16):8873.
52. Chen M, Chen Y, Zhang Q. Assessing global carbon sequestration and bioenergy potential from microalgae cultivation on marginal lands leveraging machine learning. Science of The Total Environment. 2024;948:174462.
53. Sun Y, Pargoo NS, Jin P, Ortiz J. Optimizing Autonomous Driving for Safety: A Human-Centric Approach with LLM-Enhanced RLHF. Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2024:76-80.
54. Hu Z, *et al*. Research on Financial Multi-Asset Portfolio Risk Prediction Model Based on Convolutional Neural Networks and Image Processing. arXiv preprint arXiv:2412.03618. 2024.
55. Zhang Q, Rao J, Ke Z. Enhancing Financial Forecasting Models with Textual Analysis: A Comparative Study of Decomposition Techniques and Sentiment-Driven Predictions. Innovations in Applied Engineering and Technology. 2022;1(1):1-6. DOI: 10.62836/iaet.v1i1.1009.