Journal of Systemics, Cybernetics and Informatics (2024) 22(7), 77-84
https://doi.org/10.54808/JSCI.22.07.77

Bridging the Gap: Harnessing the Power of Machine Learning and Big Data for Media Research

Li-jing Arthur CHANG, PhD (CDS&E[1]), PhD (Journalism)
Department of Journalism and Media Studies, Jackson State University
Jackson, MS 39204, USA

ABSTRACT

This paper explores the use of machine learning and big data to enhance mass media research. It covers topics such as principles of machine learning relevant to media studies, integration of computational methods with media research, data collection and preprocessing techniques, visualization of research findings, machine learning research tools, data quality and bias, ethical considerations, cross-disciplinary skills and knowledge, and best practices in data-driven research. Additionally, the paper addresses the status of media research with machine learning and big data, discussing its impact and contributions to academia and society, as well as the future challenges it may face.

Keywords: Machine Learning, Big Data, Media Research.

1. INTRODUCTION

Media research has seen the potential to increase speed, quantity, and depth by leveraging ever-improving computing power and constant-growing digital data[1][2]. Breakthroughs in computer hardware and algorithms have improved the possibilities of understanding mass media phenomena and societal issues[2][3][4][5]. This paper aims to cover the status of the intersection of machine learning, big data, and media research, along with its methodologies, data collection, information preprocessing, research tools, visualization of findings, contributions, best practices in data-driven research, ethical considerations, and future challenges.

Machine learning can be considered as a kind of artificial intelligence (AI) that allows algorithms to learn a model from data without specific programming from computer software engineers[6][7]. Two of the most common types of machine learning models are supervised and unsupervised[8][9]. In media text research, the supervised method helps build text classification models for automatic media content analysis[10]. In contrast, the unsupervised approach can uncover the latent topics of media messages[11]. The obvious advantage of applying machine learning models in media research is that it reduces the need for manual content analysis and human reading of large amounts of text to discover patterns[12][13][14].

The training of machine learning models depends on the availability of data[15]. Larger data sizes have increased the accuracy of machine learning models[16][17]. Over the past decade, research data has grown exponentially[18][19]. The arrival of big data has helped researchers with machine-learning tasks[20]. For example, with a vast amount of social media feeds, researchers can perform sentiment analyses to detect the moods of the tweets [21]. Another aspect of applying big data to machine learning is the reliance on computing facilities[22]. Complex machine learning procedures such as deep learning are only achievable through sufficient computing power[23]. This issue has been resolved over past decades when the computing capacities of personal computers has drastically improved in terms of the processing speeds of both CPUs (Central Processing Units) and GPUs (Graphical Processing Units)[24][25].

Before machine learning tasks on media text data, the first steps are gathering media data through social media platforms or acquiring news data through data banks[26][27]. Following the collection of the text data, the text itself needs to go through preprocessing steps to eliminate unnecessary noise and undergo a transformation process so that the raw text can become columns of features for analysis[28][29][30]. These preprocessing steps are part of a technique called natural language process (NLP)[31]. In short, the NLP transforms text into data ready for machine learning analysis[30].

The two most common research tools for machine learning of media text are Python and R[32]. Both are open-source programming languages[33][34]. Among the data scientists who conduct machine learning jobs on media text, Python is more popular due to its ease of model building and testing[32]. On the other hand, R is often seen as user-friendly[35][36]. Python and R can both produce visualization for research findings [37].

Machine learning technologies have contributed to automatic content analysis where the amount of text data makes manual coding unfeasible[10][38]. Past research has recommended using neural networks and transfer learning to enhance the accuracy of the supervised models used for automatic content analysis[39][40]. The quality of research samples should be a significant consideration for the generalizability of the automatic content analysis[41]. Supervised learning can also help test well-established media theories, such as agenda settings, by automatically labeling media agenda topics[42][43]. Besides supervised learning, unsupervised machine learning algorithms can see their uses in detecting latent media content topics[44][45][46][47].

Ethics concerns are among the challenges facing the application of machine learning and big data in media research[48][49]. One of the ethical concerns is data representation. For example, the data from social media tweets merely reflects people who are willing to voice their views, not mirroring public opinions[50].

[1] CDS&E stands for Computational and Data-Enabled Science and Engineering.

Another ethical consideration is the inherent bias that may be present in some media content[51]. Generalizing using biased information may result in incorrect conclusions[52][53][54]. Beyond ethics, other challenges facing media researchers include interdisciplinary collaborations between machine learning and media researchers due to their differing research priorities[55][56]. Sometimes, media researchers have to overcome their unfamiliarity with machine learning's algorithmic process, strengths, and limitations[56].

2. THEORETICAL FOUNDATIONS

Machine learning technology can enhance the testing of traditional media theories like agenda settings, framing, and uses and gratifications[43][46][57][58][59][60]. The agenda settings theory posits that the issues deemed important in media coverage will also become those the public thinks are essential [61]. The framing theory hypothesizes that the media presents issues as frames in its coverage, and these frames may impact how audiences feel about the topics[62]. The uses and gratifications approach assumes that audiences have individual motivations and selectively consume media content to gratify their needs[63]. Several machine learning-related approaches are applicable in testing these theories. These approaches include natural language processing, supervised and unsupervised learning, and predictive modeling[30][35][64][65].

For example, media articles must undergo a preprocessing phase for quantitative analysis using natural language processing (NLP) techniques. In this stage, researchers will remove unnecessary noise from the text data[30]. After that, the text will undergo a tokenization process to transform it into columns of digital data[28][66].

Following the preprocessing steps, the text becomes columns of data ready for analysis. The data analysis steps include supervised and unsupervised learning[64]. Supervised learning is suitable for data with a target variable (equivalent to the dependent variable in journalism quantitative research)[67][68]. For example, in a sentiment analysis, a machine-learning technique to detect text sentiment, the target variable could have a positive or negative value[69]. Text data is split into training and testing data as part of supervised learning[70]. Researchers use training data to train a predictive model, which will undergo evaluation with the test data to see model accuracy[71]. A theory-based model with acceptable accuracy rates can help make predictions, thereby offering theoretical insights[72].

The procedure is called supervised learning because the model building comes under the guidance of the target variable[73]. Supervised learning models can be used for classification or regression, depending on the type of information in the target variables[68]. If the target variable is categorical (binary or multi-class), the model is a classification problem (for example, a sentiment analysis)[68]. Otherwise, when the target variable is continuous, the model is a regression problem[68].

Unlike supervised learning, unsupervised learning works on data without a target variable[8]. In the example of COVID-19-related tweets, it is hard to determine the underlying major topics or frames for the tweets, thereby lacking a target variable[45]. In this case, an unsupervised learning technique called cluster analysis can be applicable to detect the tweets' underlying clusters (i.e., topics)[45].

The machine learning-related techniques—natural language processing (NLP), supervised and unsupervised learning, and predictive modeling—are suitable for testing and applying traditional media theories such as agenda settings and framing. These techniques can automate content analysis, which testers of agenda-setting and framing theories can rely on to detect major issues covered in the media[74]. Unlike the conventional content analysis that requires humans to code each article for the presence of a media topic (i.e., the target variable), machine learning-assisted content analysis only requires a small portion of the text data to be manually coded[57]. The rest of the articles can be labeled automatically with a supervised machine learning model derived from the initial human-coded sample[57]. This automatic procedure saves human and time costs and makes the analysis of large quantities of media text feasible in a much shorter time[75]. The quality of automatic content analysis can be checked by the inter-coder reliability for the initial human-coded sample and later via model evaluation metrics like accuracy and precision for the automatically coded content[71]. For harder-to-code (i.e., classify) media content, researchers can use unsupervised clustered analysis to detect underlying topics for the text[45].

The uses and gratifications approach, another media theory, is suitable for the use of supervised machine learning models to improve the model accuracy[59][60]. One of the popular methods used to test the theory is the survey, where the collected data often undergo the statistical procedure of conventional regression analysis[63]. For this type of analysis, machine learning researchers can use a type of algorithmic optimization called hyperparameters tuning to enhance linear regression model accuracy[76][77][78]. This enhancement is achievable without undermining the model's interpretability[76][77][78].

For both text and survey data, the accuracy rates of supervised learning models can be further enhanced through algorithmic comparison to pick the algorithm with the highest accuracy rate[79]. Once validated with acceptable accuracy levels, the machine-learning models built by testing these theories can make predictions and provide insights for the public[35]. For example, a model built for framing analysis can give the scholars top topics in media coverage. Election campaign managers can detect the likely trendy concerns from social media analyses. Policymakers can learn from tweet analyses about the sentiments toward public health issues. Consumers can be aware whether the topics of media content are relevant to their lives.

3. MACHINE LEARNING IN MEDIA RESEARCH

As covered in the previous section, machine learning saves human and time costs for media theory testing. These savings may come from automatic content analysis powered by supervised learning or the detection of underlying media topics aided by unsupervised learning[57]. In addition, predictive models built from classification and regression algorithms can offer insights for theoretical testing[35][72]. This section covers machine learning techniques for media researchers, such as data collection and preprocessing, data analysis, visualization of data and findings, and research tools like Python and R.

The research data for testing media theories in general have two major types: (1) media text and (2) survey data[80]. Sources of media text typically come from social media feeds and newspaper articles[81][82]. Collecting social media feeds is achievable through Application Programming Interfaces (APIs) that let researchers access data from a social media platform, such as X (formerly Twitter)[83][84]. In addition, newspaper articles can be gathered through public archives[27]. Besides text data, survey data could come from mail and online surveys[85]. Regardless of the data types, researchers need to describe data sources so that readers of the research reports may have some idea about the sample's representativeness compared to the research population.

Following the collection of media research data, the next phase is the data preprocessing. The preprocessing could have the following steps: (1) data cleaning, (2) data transformation, and (3) feature engineering. For both text and survey data, the data cleaning phase involves handling missing values[86]. Also, at this phase, text data requires additional procedures, including noise reduction (such as lowercasing of each word and removal of punctuations, stop words, and symbols) and stemming or lemmatization (stripping words to their linguistic roots)[30]. During the next phase of data transformation, text must first undergo tokenization to convert it to columns of data through techniques such as bag-of-words (i.e., a simple tokenization technique that converts text to word count) and TF-IDF (a method that picks frequent words in each text document but excludes the most common ones in the entire dataset)[28][66]. Afterward, the tokenized columns of text data could undergo normalization transformation to eliminate scales' impact on data analysis[87]. The same normalization could apply to survey data [87]. Text and survey data could also undergo different routes for the following feature engineering phase. At this stage, text data often needs dimensionality reduction due to their large number of features[88]. On the other hand, survey data usually needs feature selection to pick the most critical features (a term equivalent to independent variables in journalism research) by removing redundant and irrelevant information[89].

After the data preprocessing, the media data are ready for analysis. Depending on the nature of the data, the analysis often involves one of the two types of supervised learning: (1) classification and (2) regression[68]. Classification is usually suitable for media text analysis[90]. For example, researchers could use automatic content analysis to classify media text to test framing theory[91]. Regression, on the other hand, is often applicable to survey data[92]. Besides supervised learning, unsupervised learning could be useful in media research. For example, Latent Dirichlet Allocation (LDA) could help detect underlying topics in social media posts[47].

Data and findings visualization could be a crucial part of media research because visualization brings trends and patterns to a more digestible format for quicker and easier understanding of the otherwise hard-to-grasp statistical information[93][94]. Visualizations could be simple pattern diagrams such as bar charts, histograms, trendlines, and pie charts[95]. A little more complex examples include a heatmap that shows a correlation matrix in colors with darker colors demonstrating stronger correlations[96]. Another example could be a word cloud showing text frequency patterns and significant themes in a corpus[97]. An even more complex example could be interactive plots such as Plotly-powered dashboards where users can interact with the data to see machine learning research findings under different scenarios in real time[37]. Another interesting case could be a network presentation of a social media network, which can show a network's connection frequencies and node intensity (i.e., the presence of hot spots)[98].

Except for data collection, all the steps involved in the machine learning application of media research (i.e., data preprocessing, analysis, and visualization) are achievable through Python and R [35][99]. Python and R have dedicated libraries for machine learning and deep learning, the latter of which is a branch of machine learning that involves learning through multiple layers of neural networks[35][99][100][101]. R, designed for statistical analysis, has the pre-built capacity to view data in an Excel-like format[102]. However, Python has a Pandas library that performs a similar function[103]. Although the two languages are effective in machine learning modeling and visualization, one drawback is that both require coding and have a comparable learning curve for people without previous coding experience[36].

4. CURRENT STATUS AND CHALLENGES

As mentioned, machine learning applications in media research could have a few benefits: (1) saving human and time costs through automatic content analysis, (2) enhancing predicting accuracy, and (3) shedding insights through predictive models. Because of the advantages, a few researchers have begun to test media theories using machine learning.

For example, studies have used supervised machine learning for automatic content analysis to detect agenda topics in media articles[43]. In such automatic content analysis, researchers first manually coded a small sub-sample from the dataset and used the coded sub-sample to build a machine learning classifier to label the rest of the dataset[57]. Another way to automatically label media articles for agendas involves using unsupervised machine learning techniques such as Latent Dirichlet Allocation (LDA) to detect the underlying topics of the media text[104]. For automatic content analysis, machine learning techniques (both supervised and unsupervised) are considered better approaches than the conventional dictionary-based method (e.g., the keyword(s) counting), the latter of which is often seen as inflexible because the use of single words or groups of words frequently falls short of detecting the rich context in the text[57][68]. Between the supervised and unsupervised learning techniques, supervised learning performs better than the unsupervised method because the topic detection patterns learned from the former can undergo validation from labeled text[57].

Similarly, researchers have used machine learning (both supervised and unsupervised) for automatic content analysis to spot media frames[91]. Likewise, the automatic approach has been the methodology for two popular types of media research: social media sentiment analysis and fake news detection[105][106]. Regarding the uses and gratifications theory, researchers have started to use supervised machine learning to enhance the accuracy of prediction models by comparing algorithms and picking top performers[59]. Although the predictive models built from these media studies can undergo deployment to build user interfaces to access real-time theoretical insights[35], academic researchers have yet to explore this application, making it a future research possibility.

While the media research applying machine learning has shown the viability of such an application, the novel approach also faces challenges, including data quality and bias, ethical concerns, and interdisciplinary collaborations[50][51][55][56][107][108].

These challenges call for adopting best practices in a machine learning-driven research paradigm.

An obvious challenge in machine learning-aided media research is data quality and bias. For example, tweets from the X platform (formerly Twitter) have information that would be considered research noise, such as hashtags, emojis, capitalizations, and HTML links[109]. Some blank posts could represent missing values in research data[110]. Additionally, social media feeds do not fully represent public opinions because only a portion of the population are active on these media [50]. In another example, news stories could contain biases as the framing research showed that the media report news events using preexisting frames, and these frames could also promote stereotypes[51][54]. The data issues pose challenges because a machine learning model can only be as effective as the data on which it is trained.

A related challenge is the ethical concerns arising from data usage and research reporting. For instance, although posts on the X platform are data available to the public, the tweets should be de-identified to keep anonymity and avoid privacy issues[111]. Another ethical concern is the possible misuse of study results due to a lack of knowledge about data sources and model training processes[112][113]. To prevent this, researchers applying machine learning technology in media studies should focus on the interpretability of the machine learning models[114][115].

A third challenge is related to the interdisciplinary nature of applying machine learning in media research[6][55]. The collaborating parties, namely data scientists from the machine learning field and media researchers from the social sciences and humanities side, may have different research priorities. Data scientists are more interested in model accuracy, scalability, and efficiency, while media researchers are more concerned about model interpretability and theory building[68]. In a joint venture, data scientists and media researchers must learn to compromise to succeed. The trade-offs between model accuracy and interpretability depend on the nature of the research [114]. For example, while a deep learning algorithm may bring higher accuracy in testing the uses and gratifications theory, researchers may opt for the slightly less accurate linear regression algorithm because the latter can ensure interpretability[114].

To tackle the challenges from data quality and bias, ethical concerns, and interdisciplinary collaborations, researchers could observe a few practices: (1) data management, (2) validation and evaluation, (3) documentation and reproducibility, (4) ethic review, and (5) stakeholder engagement. Data management should focus on data cleaning to remove noise and data selection to enhance diversity[113][116]. Validation and evaluation could ensure that algorithms are adequately trained and tested[117]. Documentation could help clearly describe and explain data and methodology to allow reproducibility[118]. Ethical review aims to prevent privacy issues and research finding misrepresentation[111][113]. Finally, stakeholder engagement fosters compromise and teamwork to promote theory-testing synergy during interdisciplinary collaborations[68][115].

5. CONCLUSION

The advances in machine learning algorithms, the availability of big media data, and the ever-improving computing power provide new opportunities for media researchers to test media theories such as agenda settings, framing, and uses and gratifications[42][58][59]. The benefits of machine learning in testing media theories include time and labor savings from automatic content analysis, enhancement of research model accuracy, and potential deployment of predictive models to offer theoretical insights[35][57][77].

For example, supervised learning-aided content analysis saves costs in testing agenda settings and framing theories by using a small sample of human-coded text documents to create a classifier for labeling the rest of the text data [42][58]. This automated approach saves time and human labor and provides the advantage of the scalability of big media data, often available in social media[57][105]. Besides the supervised approach, researchers can opt for unsupervised topic modeling to classify large quantities of text documents without labels[64]. As another example, supervised learning can enhance model accuracy for testing the uses and gratifications theory through feature selection, algorithmic optimization, and model comparison[59][77][79][89].

The capacity to automatically analyze large amounts of data with greater accuracy provides potentially more robust research results[77][79]. Compared to traditional manual content analysis, machine learning-aided content analysis offers the opportunity to analyze much larger and more representative samples, as the volume of text data increases by hundreds or even thousands of times or more[119][120]. Although yet to be studied for media research, deploying the predictive models would provide opportunities for media scholars to test theories in real-time. This possibility could become a topic of further studies.

To leverage machine learning in media studies, researchers need to pay attention to techniques in data collection and preprocessing, supervised learning (including classification and regression), unsupervised learning (e.g., LDA or Latent Dirichlet Allocation), and data and research findings visualization (e.g., bar graph, interactive chart, or dashboard)[30][37][47][64][85][95]. Although the adoption of machine learning in media research is still in its early stage, the synergy coming out of the intersection between machine learning and media research brings the advantages of cost and time saving, scalability (with big data), and enhanced model accuracy[57][77][79][119].

Despite these benefits, interdisciplinary research poses challenges because scholars from different domains may have differing research priorities. Machine learning researchers may focus more on model accuracy, while media scholars may emphasize model interpretability more[68]. To compromise, both parties may need to weigh the costs and benefits of focusing on accuracy or interpretability[114].

Adopters of machine learning in media research may face other challenges like data quality and ethical concerns. For example, data from social media feeds may have quality issues because of noise such as symbols and HTML links[30]. Also, media articles may have biased news frames that could promote stereotypes[51][53][54]. To improve data quality, researchers should remove the noise and attempt to diversify data sources to minimize bias[113]. To help readers judge the findings, researchers could mention the details about data preprocessing steps, data sources, and sampling processes in their reports [118]. Research reports that have this information along with data testing and analysis steps could facilitate research reproducibility and prevent misuse of findings[114][115][118].

The intersection of machine learning, big data, and media research holds great potential. The synergy from the joint venture can accelerate media theory testing, increase sample representativeness, optimize model accuracy, and offer real-time theoretical insights. To adopt machine learning in media studies, researchers need to heed data quality and bias, ethical concerns, research reproducibility, and resolution of interdisciplinary differences.

6. REFERENCES

[1] W. Van Zoonen and G. L. A. Toni, "Social media research: The application of supervised machine learning in organizational communication research.," *Comput Human Behav*, vol. 63, pp. 132–141, 2016.

[2] L. Ma and B. Sun, "Machine learning and AI in marketing–Connecting computing power to human insights," *International Journal of Research in Marketing*, vol. 37, no. 3, pp. 481–504, 2020.

[3] I. Lundberg, J. E. Brand, and N. Jeon, "Researcher reasoning meets computational capacity: Machine learning for social science," *Soc Sci Res*, vol. 108, p. 102807, 2022.

[4] D. M. J. Lazer *et al.*, "Computational social science: Obstacles and opportunities," *Science (1979)*, vol. 369, no. 6507, pp. 1060–1062, 2020.

[5] M. Felt, "Social media and the social sciences: How researchers employ Big Data analytics," *Big Data Soc*, vol. 3, no. 1, p. 2053951716645828, 2016.

[6] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science (1979)*, vol. 349, no. 6245, pp. 255–260, 2015.

[7] D. Dhall, R. Kaur, and M. Juneja, "Machine learning: a review of the algorithms and its applications," *Proceedings of ICRIC 2019: Recent innovations in computing*, pp. 47–63, 2020.

[8] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4, no. 4. Springer, 2006.

[9] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[10] B. Burscher, D. Odijk, R. Vliegenthart, M. De Rijke, and C. H. De Vreese, "Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis," *Commun Methods Meas*, vol. 8, no. 3, pp. 190–206, 2014.

[11] L. Guo, C. J. Vargo, Z. Pan, W. Ding, and P. Ishwar, "Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling," *Journal Mass Commun Q*, vol. 93, no. 2, pp. 332–359, 2016.

[12] M. Scharkow, "Content analysis, automatic," *The international encyclopedia of communication research methods*, pp. 1–14, 2017.

[13] G. C. Nunez-Mir, B. V Iannone III, B. C. Pijanowski, N. Kong, and S. Fei, "Automated content analysis: addressing the big literature challenge in ecology and evolution," *Methods Ecol Evol*, vol. 7, no. 11, pp. 1262–1272, 2016.

[14] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP J Adv Signal Process*, vol. 2016, pp. 1–16, 2016.

[15] S. Suthaharan, "Machine learning models and algorithms for big data classification," *Integr. Ser. Inf. Syst*, vol. 36, pp. 1–12, 2016.

[16] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intell Syst*, vol. 24, no. 2, pp. 8–12, 2009.

[17] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[18] M. Phiri, "Exponential growth of data," Medium. Accessed: Sep. 10, 2024. [Online]. Available: https://medium.com/@mwaliph/exponential-growth-of-data-2f53df89124

[19] B. Marr, "How much data do we create every day? The mind-blowing stats everyone should read," *Forbes*, May 21, 2018.

[20] L. Zhou, S. Pan, J. Wang, and A. V Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, 2017.

[21] K. Al-Barznji and A. Atanassov, "Big data sentiment analysis using machine learning algorithms," in *Proceedings of 26th International Symposium" Control of Energy, Industrial and Ecological Systems, Bankia, Bulgaria*, 2018.

[22] V. Sze, Y.-H. Chen, J. Emer, A. Suleiman, and Z. Zhang, "Hardware for machine learning: Challenges and opportunities," in *2017 IEEE custom integrated circuits conference (CICC)*, 2017, pp. 1–8.

[23] B. Y. Kasula, "Machine Learning Unleashed: Innovations, Applications, and Impact Across Industries," *International Transactions in Artificial Intelligence*, vol. 1, no. 1, pp. 1–7, 2017.

[24] F. H. Khan, M. A. Pasha, and S. Masud, "Advancements in microprocessor architecture for ubiquitous AI—An overview on history, evolution, and upcoming challenges in AI implementation," *Micromachines (Basel)*, vol. 12, no. 6, p. 665, 2021.

[25] E. Buber and D. Banu, "Performance analysis and CPU vs GPU comparison for deep learning," in *2018 6th International Conference on Control Engineering & Information Technology (CEIT)*, 2018, pp. 1–6.

[26] I. Lasri, A. Riadsolh, and M. Elbelkacemi, "Real-time Twitter Sentiment Analysis for Moroccan Universities using Machine Learning and Big Data Technologies.," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 5, 2023.

[27] M. Broersma and F. Harbers, "Exploring machine learning to study the long-term transformation of news: Digital newspaper archives, journalism history, and algorithmic transparency," in *Journalism History and Digital Archives*, Routledge, 2020, pp. 38–52.

[28] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PLoS One*, vol. 15, no. 5, p. e0232525, 2020.

[29] A. I. Kadhim, "An evaluation of preprocessing techniques for text classification," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 16, no. 6, pp. 22–32, 2018.

[30] D. Sarkar, *Text analytics with python*, vol. 2. Springer, 2016.

[31] F. Sun, A. Belatreche, S. Coleman, T. M. McGinnity, and Y. Li, "Pre-processing online financial text for sentiment classification: A natural language processing

approach," in *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, 2014, pp. 122–129.

[32] S. Anitha Elavarasi and J. Jayanthi, "Programming Language Support for Implementing Machine Learning Algorithms," 2020.

[33] K. R. Srinath, "Python–the fastest growing programming language," *International Research Journal of Engineering and Technology*, vol. 4, no. 12, pp. 354–357, 2017.

[34] P. Mair, E. Hofmann, K. Gruber, R. Hatzinger, A. Zeileis, and K. Hornik, "Motivation, values, and work design as drivers of participation in the R open source project for statistical computing," *Proceedings of the National Academy of Sciences*, vol. 112, no. 48, pp. 14788–14792, 2015.

[35] B. Lantz, *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd, 2019.

[36] C. Ozgur, T. Colliau, G. Rogers, Z. Hughes, and others, "MatLab vs. Python vs. R," *Journal of data Science*, vol. 15, no. 3, pp. 355–371, 2017.

[37] S. K. A. Fahad and A. E. Yahya, "Big data visualization: allotting by R and python with GUI tools," in *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, 2018, pp. 1–8.

[38] G. C. Nunez-Mir, B. V Iannone III, B. C. Pijanowski, N. Kong, and S. Fei, "Automated content analysis: addressing the big literature challenge in ecology and evolution," *Methods Ecol Evol*, vol. 7, no. 11, pp. 1262–1272, 2016.

[39] M. Aydoğan and A. Karci, "Improving the accuracy using pre-trained word embeddings on deep neural networks for Turkish text classification," *Physica A: Statistical Mechanics and its Applications*, vol. 541, p. 123288, 2020.

[40] L. Chang, "Comparison of machine learning and deep learning algorithms in detecting fake news," in *Proceedings of the WMSCI2024 conference*, 2024, pp. 0–7.

[41] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[42] F. Gilardi, T. Gessler, M. Kubli, and S. Müller, "Social media and political agenda setting," *Polit Commun*, vol. 39, no. 1, pp. 39–60, 2022.

[43] H. Wang and J. Shi, "Intermedia agenda setting amid the pandemic: A computational analysis of China's online news," *Comput Intell Neurosci*, vol. 2022, no. 1, p. 2471681, 2022.

[44] K. Wu, "Agenda-setting in cross-national coverage of COVID-19: an analysis of elite newspapers in US and China with topic modeling," *Online J Commun Media Technol*, vol. 11, no. 4, p. e202116, 2021.

[45] K. Liu, X. Geng, and X. Liu, "The application of network agenda setting model during the COVID-19 pandemic based on latent dirichlet allocation topic modeling," *Front Psychol*, vol. 13, p. 954576, 2022.

[46] C. Han, M. Yang, and A. Piterou, "Do news media and citizens have the same agenda on COVID-19? an empirical comparison of twitter posts," *Technol Forecast Soc Change*, vol. 169, p. 120849, 2021.

[47] X. Wang, L. Chen, J. Shi, and H. Tang, "Who sets the agenda? The dynamic agenda setting of the wildlife issue on social media," *Environ Commun*, vol. 17, no. 3, pp. 245–262, 2023.

[48] W. Chen and A. Quan-Haase, "Big Data Ethics and Politics: Toward New Understandings," *Soc Sci Comput Rev*, vol. 38, no. 1, pp. 3–9, 2020.

[49] R. Matheus and M. Janssen, "Transparency dimensions of big and open linked data: Transparency as being synonymous with accountability and openness," in *Open and Big Data Management and Innovation: 14th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2015, Delft, The Netherlands, October 13-15, 2015, Proceedings 14*, 2015, pp. 236–246.

[50] T. O'Brien, "Are Twitter users an accurate reflection of real public sentiment?," Muck Rack. Accessed: Sep. 13, 2024. [Online]. Available: https://muckrack.com/blog/2020/11/17/twitter-and-public-sentiment

[51] D. P. Baron, "Persistent media bias," *J Public Econ*, vol. 90, no. 1–2, pp. 1–36, 2006.

[52] T. L. Dixon, "Media stereotypes: Content, effects, and theory," in *Media effects*, Taylor & Francis, 2019.

[53] S. T. Murphy, "The impact of factual versus fictional media portrayals on cultural stereotypes," *Ann Am Acad Pol Soc Sci*, vol. 560, no. 1, pp. 165–178, 1998.

[54] L. Aaldering and D. J. Van Der Pas, "Political leadership in the media: Gender bias in leader stereotypes during campaign and routine times," *Br J Polit Sci*, vol. 50, no. 3, pp. 911–931, 2020.

[55] J. Grimmer, M. E. Roberts, and B. M. Stewart, "Machine learning for social science: An agnostic approach," *Annual Review of Political Science*, vol. 24, no. 1, pp. 395–419, 2021.

[56] I. Lundberg, J. E. Brand, and N. Jeon, "Researcher reasoning meets computational capacity: Machine learning for social science," *Soc Sci Res*, vol. 108, p. 102807, 2022.

[57] B. Burscher, D. Odijk, R. Vliegenthart, M. De Rijke, and C. H. De Vreese, "Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis," *Commun Methods Meas*, vol. 8, no. 3, pp. 190–206, 2014.

[58] O. Eisele, T. Heidenreich, O. Litvyak, and H. G. Boomgaarden, "Capturing a news frame–comparing machine-learning approaches to frame analysis with different degrees of supervision," *Commun Methods Meas*, vol. 17, no. 3, pp. 205–226, 2023.

[59] R. A. Al-Maroof, I. Arpaci, M. Al-Emran, S. A. Salloum, and K. Shaalan, "Examining the acceptance of WhatsApp stickers through machine learning algorithms," *Recent advances in intelligent systems and smart applications*, pp. 209–221, 2021.

[60] J. Rochotte, A. Sanap, V. Silenzio, and V. K. Singh, "Predicting anxiety using Google and Youtube digital traces," *Emerging Trends in Drugs, Addictions, and Health*, vol. 4, p. 100145, 2024.

[61] M. McCombs, "A look at agenda-setting: Past, present and future," *Journal Stud*, vol. 6, no. 4, pp. 543–557, 2005.

[62] D. A. Scheufele and D. Tewksbury, "Framing, agenda setting, and priming: The evolution of three media effects models," *Journal of communication*, vol. 57, no. 1, pp. 9–20, 2007.

[63] T. E. Ruggiero, “Uses and gratifications theory in the 21st century,” *Mass Commun Soc*, vol. 3, no. 1, pp. 3–37, 2000.

[64] M. Kaur and V. Kumar, “Optimization of Text Classification using Supervised and Unsupervised Learning Approach,” *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 6, no. 4, pp. 3385–3387, 2015.

[65] M. W. Berry, A. Mohamed, and B. W. Yap, *Supervised and unsupervised learning for data science*. Springer, 2019.

[66] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, “The impact of features extraction on the sentiment analysis,” *Procedia Comput Sci*, vol. 152, pp. 341–348, 2019.

[67] V. Nasteski, “An overview of the supervised machine learning methods,” *Horizons. b*, vol. 4, no. 51–62, p. 56, 2017.

[68] F. De Grove, K. Boghe, and L. De Marez, “(What) can journalism studies learn from supervised machine learning?,” *Journal Stud*, vol. 21, no. 7, pp. 912–927, 2020.

[69] B. Agarwal, N. Mittal, B. Agarwal, and N. Mittal, “Machine learning approach for sentiment analysis,” *Prominent feature extraction for sentiment analysis*, pp. 21–45, 2016.

[70] S. González-Carvajal and E. C. Garrido-Merchán, “Comparing BERT against traditional machine learning text classification,” *arXiv preprint arXiv:2005.13012*, 2020.

[71] L. A. Chang, “Detecting Asian values in Asian news via machine learning text classification,” in *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, 2021, pp. 123–128.

[72] R. G. Winther, “The structure of scientific theories,” 2015, Accessed: Sep. 14, 2024. [Online]. Available: https://plato.stanford.edu/entrieS/structure-scientific-theories/

[73] P. Cunningham, M. Cord, and S. J. Delany, “Supervised learning,” in *Machine learning techniques for multimedia: case studies on organization and retrieval*, Springer, 2008, pp. 21–49.

[74] B. Burscher and others, “Machine learning-based content analysis: Automating the analysis of frames and agendas in political communication research,” Universiteit van Amsterdam [Host], 2016.

[75] J. Salminen, V. Yoganathan, J. Corporan, B. J. Jansen, and S.-G. Jung, “Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type,” *J Bus Res*, vol. 101, pp. 203–217, 2019.

[76] G. James, D. Witten, T. Hastie, R. Tibshirani, and others, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[77] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[78] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[79] F. Y. Osisanwo *et al.*, “Supervised machine learning algorithms: classification and comparison,” *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, 2017.

[80] R. D. Wimmer, J. R. Dominick, and others, *Mass media research: An introduction*. Cengage India, 2015.

[81] H. A. Schwartz and L. H. Ungar, “Data-driven content analysis of social media: A systematic overview of automated methods,” *Ann Am Acad Pol Soc Sci*, vol. 659, no. 1, pp. 78–94, 2015.

[82] J. Salminen, V. Yoganathan, J. Corporan, B. J. Jansen, and S.-G. Jung, “Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type,” *J Bus Res*, vol. 101, pp. 203–217, 2019.

[83] M. W. Kearney, “rtweet: Collecting and analyzing Twitter data,” *J Open Source Softw*, vol. 4, no. 42, p. 1829, 2019.

[84] C. Stokel-Walker, “Why is Twitter becoming X?,” 2023, *Elsevier*.

[85] D. K. Loomis and S. Paterson, “A comparison of data collection methods: Mail versus online surveys,” *J Leis Res*, vol. 49, no. 2, pp. 133–149, 2018.

[86] A. Palanivinayagam and R. Damaševičius, “Effective handling of missing values in datasets for classification using machine learning methods,” *Information*, vol. 14, no. 2, p. 92, 2023.

[87] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, “Effect of data scaling methods on machine learning algorithms and model performance,” *Technologies (Basel)*, vol. 9, no. 3, p. 52, 2021.

[88] N. Akkarapatty, A. Muralidharan, N. S. Raj, and P. Vinod, “Dimensionality reduction techniques for text mining,” in *Collaborative filtering using data mining and analysis*, IGI Global, 2017, pp. 49–72.

[89] J. Cai, J. Luo, S. Wang, and S. Yang, “Feature selection in machine learning: A new perspective,” *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[90] A. Khan, B. Baharudin, L. H. Lee, and K. Khan, “A review of machine learning algorithms for text-documents classification,” *Journal of advances in information technology*, vol. 1, no. 1, pp. 4–20, 2010.

[91] Z. Fatemi *et al.*, “Understanding stay-at-home attitudes through framing analysis of tweets,” in *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, 2022, pp. 1–10.

[92] A. Gelman, *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2007.

[93] C. Chen, W. K. Härdle, and A. Unwin, *Handbook of data visualization*. Springer Science & Business Media, 2008.

[94] A. Ng, “Machine learning yearning: Technical strategy for AI engineers, in the era of deep learning,” 2018.

[95] W. H. Organization and others, “Tools for making good data visualizations: the art of charting,” 2021.

[96] S. Buyrukoğlu and A. Akbaş, “Machine learning based early prediction of type 2 diabetes: A new hybrid feature selection approach using correlation matrix with heatmap and SFS,” *Balkan Journal of Electrical and Computer Engineering*, vol. 10, no. 2, pp. 110–117, 2022.

[97] N. Shahid, M. U. Ilyas, J. S. Alowibdi, and N. R. Aljohani, “Word cloud segmentation for simplified

exploration of trending topics on Twitter," *IET Software*, vol. 11, no. 5, pp. 214–220, 2017.

[98] D. Goldenberg, "Social network analysis: From graph theory to applications with python," *arXiv preprint arXiv:2102.10014*, 2021.

[99] S. Raschka, J. Patterson, and C. Nolet, "Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence," *Information*, vol. 11, no. 4, p. 193, 2020.

[100] G. Ciaburro and B. Venkateswaran, *Neural Networks with R: Smart models using CNN, RNN, deep learning, and artificial intelligence principles*. Packt Publishing Ltd, 2017.

[101] F. Chollet, *Deep learning with Python*. Simon and Schuster, 2021.

[102] H. Wickham and G. Grolemund, *R for Data Science*. O'Reilly Media, Inc., 2016.

[103] W. McKinney and P. D. Team, "Pandas-Powerful python data analysis toolkit," *Pandas—Powerful Python Data Analysis Toolkit*, vol. 1625, 2015.

[104] M. Ferreira, V. Rolim, R. F. Mello, R. D. Lins, G. Chen, and D. Gašević, "Towards automatic content analysis of social presence in transcripts of online discussions," in *Proceedings of the tenth international conference on learning analytics & knowledge*, 2020, pp. 141–150.

[105] R. S. Wadawadagi and V. B. Pagi, "Sentiment analysis on social media: recent trends in machine learning," *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, pp. 780–799, 2022.

[106] Z. Khanam, B. N. Alwasel, H. Sirafi, and M. Rashid, "Fake news detection using machine learning approaches," in *IOP conference series: materials science and engineering*, 2021, p. 12040.

[107] N. Gupta *et al.*, "Data quality for machine learning tasks," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 4040–4041.

[108] L. Budach *et al.*, "The effects of data quality on machine learning performance," *arXiv preprint arXiv:2207.14529*, 2022.

[109] D. Sarkar, *Text analytics with python*, vol. 2. Springer, 2016.

[110] P. Mariani, A. Marletta, and M. Locci, "Missing values and data enrichment: an application to social media liking," *Comput Stat*, vol. 39, no. 1, pp. 217–237, 2024.

[111] C. Fiesler and N. Proferes, "'Participant' perceptions of Twitter research ethics," *Soc Media Soc*, vol. 4, no. 1, p. 2056305118763366, 2018.

[112] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices," *Sci Eng Ethics*, vol. 26, no. 4, pp. 2141–2168, 2020.

[113] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[114] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[115] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[116] E. Rahm, H. H. Do, and others, "Data cleaning: Problems and current approaches," *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.

[117] N. Japkowicz and M. Shah, "Performance evaluation in machine learning," *Machine Learning in Radiation Oncology: Theory and Applications*, pp. 41–56, 2015.

[118] H. Semmelrock, S. Kopeinik, D. Theiler, T. Ross-Hellauer, and D. Kowald, "Reproducibility in machine learning-driven research," *arXiv preprint arXiv:2307.10320*, 2023.

[119] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," *Big Data and Cognitive Computing*, vol. 4, no. 1, p. 1, 2020.

[120] R. V Hogg, E. A. Tanis, and D. L. Zimmerman, *Probability and statistical inference*. Pearson Education, 2015.