Predicting Individual Media Preferences and News Exposure: An Algorithmic Approach Using Demographic Data and Media Use Habits

Nour Assili[1] and Neeley Pate[1] and Adiba Proma[1] and Dihan Shi[2]

***Abstract*— Media plays a crucial role in shaping public opinion and political discourse. In this study, we develop an algorithm to predict an individual's preferences for media outlets based on demographics and media use habits. We examine the performance of our algorithm across eight classical and neural machine learning classification models on 22,179 samples from the 2022 Cooperative Election Study (CES) with complete responses on the consumption of ABC, CBS, NBC, CNN, Fox News, MSNBC, and PBS. Our findings have implications for understanding individuals' media preferences and content exposure. Political scientists can leverage our algorithm for downstream analyses of public opinion formation, issue salience, media bias, and voting behavior.**

I. INTRODUCTION

The media plays a crucial role in shaping public opinion and political discourse [21]. In an increasingly digital age, individuals have access to a wide range of media outlets, each with its own unique perspective and bias [8]. This abundance of choice has led to concerns about selective exposure, where individuals primarily consume media that aligns with their pre-existing beliefs [23], potentially leading to increased polarization and the formation of echo chambers [24].

Therefore, understanding the factors that influence an individual's media preferences and the content they are exposed to is essential for analyzing public opinion formation, issue salience, media bias, and voting behavior. Previous research has explored the relationship between demographic characteristics and media preferences [17], [25], as well as the role of media use habits in shaping exposure to political information [10], [26].

Building upon this foundation, our study aims to develop an algorithm for predicting individual media preferences based on demographic information and media use habits, using data from the 2022 Cooperative Election Study (CES). This approach draws from the growing field of computational social science, which leverages computational methods to analyze large-scale social and political data [19], [27].

Our findings have important applications in political science research on public opinion formation, issue salience, media bias, and polarization. To begin with, the "Receive-Accept-Sample" (RAS) model, developed by John Zaller [28], provides a framework for understanding how individuals form their opinions by receiving, accepting, and sampling information from various sources, including the media. Zaller argues that more informed individuals are more likely to resist persuasion from media messages that contradict their pre-existing beliefs. Our study can be used to examine this model by predicting the media outlets an individual is likely to receive information from based on their demographic characteristics and media use habits, thereby providing insights into the potential for media influence on opinion formation.

Additionally, media bias and polarization have been widely studied in the context of political communication. Gentzkow and Shapiro [9] found that newspaper slant is driven by the preferences of consumers, suggesting that media outlets cater to the ideological leanings of their audience. Prior [30] argues that the proliferation of media choices has led to increased polarization, as individuals can more easily select outlets that align with their political views. By identifying the political leaning of the news information individuals receive, further research using our media preference prediction algorithm can contribute to the ongoing discussion on the role of media in fostering polarization and populist tendencies.

Moreover, the perceived salience of political issues is often influenced by the media's agenda-setting function. McCombs and Shaw's [21] agenda-setting theory posits that the media's emphasis on certain issues influences the public's perception of their importance. Guo and McCombs [11] extended this theory to the network agenda-setting model, which accounts for the interconnection of issues and attributes in shaping public opinion. Paired with survey data, our study can provide a framework for analyzing the relationship between an individual's media exposure and the salience they assign to specific political issues, taking into account the topic and sentiment of the news content they receive.

Finally, our findings can be used to study media influence on voting intentions, as exposure to specific types of news content may shape political attitudes and behavior. The impact of media on voting behavior has been examined in the early work of Lazarsfeld et al. [18] on the "two-step flow" of communication and more recent studies investigating in the context of increasing media fragmentation and polarization [7], [23]. Iyengar and Kinder [12] argue that the way media frames and primes political issues can influence the criteria by which voters evaluate candidates. By associating different news sources with their coverage of candidates and political leaning, our research could inform campaign strategies, contribute to our understanding of how media shapes public opinion and electoral outcomes, and shed light on the growing role of digital media in election campaigns [5].

*This work was not supported by any organization
[1]University of Rochester Computer Science Department
[2]Washington University in St. Louis Political Science Department

II. BACKGROUND AND PRIOR WORK

Traditionally in political communication, political preferences have been considered key predictors of media preferences, with individuals often seeking out media content that aligns with their pre-existing political beliefs [23]. However, a growing body of psychological and biological research suggests that demographics and media use habits alone can provide valuable insights into individual media preferences, even without taking political preferences into account.

Specifically, a number of studies have demonstrated the influence of demographic factors on individuals' preferences for specific media outlets. Age, for instance, has been shown to play a significant role in shaping media outlet choices, with older adults preferring traditional media outlets and younger individuals gravitating towards digital-native outlets and social media platforms [3], [13], [31]. Gender has also been linked to differences in media outlet preferences, with women more likely to prefer lifestyle and human-interest content and men favoring sports and business news [15], [32]. Education level and socioeconomic status can also influence media outlet preferences. Individuals with higher levels of education tend to prefer news outlets that provide in-depth analysis and complex coverage of political and social issues, and they may be more likely to seek out outlets that align with their political ideologies [33], [22]. In contrast, individuals with lower levels of education may gravitate towards outlets that provide more accessible and entertaining content [34]. Socioeconomic status, encompassing factors such as income and occupation, can affect access to a wider range of media outlets and the leisure time available to consume news and engage with complex issues [6], [29].

Additionally, media use habits, particularly the medium and intensity of media consumption, have been shown to significantly influence individuals' preferences for specific media outlets. The choice of medium, such as television, radio, print, or digital platforms, can shape the types of media outlets individuals are exposed to and, consequently, their outlet preferences. For example, individuals who primarily consume news through television may gravitate towards cable news networks or local broadcasts, while those who rely on digital platforms like social media or news websites may prefer online news sources or digital-native outlets [1]. The intensity of media use, which refers to the frequency and duration of media consumption, has also been linked to media outlet preferences. Heavy media users tend to have more diverse media diets and are more likely to consume content from a variety of outlets across different platforms [17]. In contrast, light media users may have more limited outlet preferences and stick to a narrower range of sources.

The relationship between media use habits and outlet preferences can be understood through the uses and gratifications theory [38]. This theory suggests that individuals actively seek out media outlets that satisfy their specific needs and desires, such as information-seeking, entertainment, or social connection. As individuals develop habitual patterns of media consumption based on their needs, these habits can shape their outlet preferences over time. For example, individuals who frequently use social media platforms for news consumption may develop preferences for outlets that produce shorter, more visual content that is easily shareable [20].

Algorithmically, extant studies in computer science have employed machine learning techniques to predict user preferences and personalize content recommendations. Collaborative filtering algorithms, such as matrix factorization [16] and deep learning approaches [35], have been widely used to predict user preferences based on their past behaviors and the preferences of similar users. These techniques have been applied to various domains, including movie recommendations [2], music recommendations [36], and news article recommendations [37].

In the context of news media, Karimi et al. [14] develop a news recommender system that takes into account user preferences, article content, and collaborative filtering to provide personalized news recommendations. Bharadhwaj et al. [4] propose a hierarchical attention model for news recommendation, which considers both user preferences and the temporal dynamics of news articles. These studies demonstrate the potential for applying machine learning techniques to predict individual media preferences and personalize news exposure.

Our research differs from prior work by incorporating demographic information and media use habits into the prediction of media preferences, with downstream social-science applications in mind. Leveraging non-political predictors only, we aim to develop a more parsimonious and generalizable approach to understanding individual media consumption.

III. METHODOLOGY

We develop our algorithm using data from the 2022 Cooperative Election Study (CES), a nationally representative survey that includes information on respondents' demographic characteristics, media use habits, and news outlet preferences. These characteristics are then fed into several different machine learning models for prediction evaluation.

First, we preprocess the data. Of the 60,000 respondents, only 22,179 participants answered our outcome questions of interest (news outlet preferences); observations with incomplete outcome variables are removed. Then, demographic and media use habit variables are extracted from the data. Demographic information includes the state of the participant, the birth year of the participant, the gender of the participant, the race of the participant, whether or not the participant is Hispanic, the participant's occupational category, the participant's current employment status, and the participant's annual family income. For media use habits, we include the medium of media used in the past 24 hours, how often the individual reads paper news, how often the individual watches broadcast news, and social media actions performed in the past 24 hours. This data was then processed to be one-hot encoded for each categorical predictor; we treat all predictors as categorical except for birth year. Missing values are treated as a separate category for all predictors (birth year does not contain missing values), because we deem non-response tendencies to survey questions to be substantively important for media outlet preferences.

For this study, the following classical machine learning algorithms are tested: K-Nearest Neighbor, Gradient Boosting, Logistic Regression, XGBoost, Random Forests. Moreover, we also examine basic Neural Networks, Cosine Similarity, and Transformers. We predict the probability of the individual to consume each of seven unique news sources in the United States: ABC, CBS, NBC, CNN, Fox News, MSNBC, and PBS. We compare model performance using F1, precision and recall metrics.

IV. Results

Our Random Forest model has the best accuracy performance of all models tested. As a decision tree algorithm, random forests are less influenced by outliers than other algorithms.

Tables 1 and 2 present the accuracy for each news source, for both positive prediction (the individual consumes the corresponding news source) and negative prediction (the individual does not consume the corresponding news source). Tables 3 and 4 outline the macro average accuracy for each news source.

TABLE I
ACCURACY PERFORMANCE FOR EACH NEWS OUTLET, PREDICTING POSITIVE RESPONSE

ABC	CBS	NBC	CNN	Fox News	MSNBC	PBS
0.83	0.85	0.84	0.86	0.84	0.9	0.94

TABLE II
ACCURACY PERFORMANCE FOR EACH NEWS OUTLET, PREDICTING NEGATIVE RESPONSE

ABC	CBS	NBC	CNN	Fox News	MSNBC	PBS
0.79	0.82	0.8	0.82	0.8	0.86	0.92

TABLE III
MACRO AVERAGE PERFORMANCE FOR EACH NEWS OUTLET, PREDICTING POSITIVE RESPONSE

ABC	CBS	NBC	CNN	Fox News	MSNBC	PBS
0.57	0.59	0.52	0.61	0.69	0.45	0.47

TABLE IV
MACRO AVERAGE PERFORMANCE FOR EACH NEWS OUTLET, PREDICTING NEGATIVE RESPONSE

ABC	CBS	NBC	CNN	Fox News	MSNBC	PBS
0.71	0.76	0.73	0.77	0.73	0.83	0.9

Further analyses of the dataset show that the dataset had a slight imbalance, where most participants selected no for any given source, which may explain why macro average for predicting positive response is lower. Table 5 shows the percentage of "yes" answers for each source. This impact is shown through the consistently lower precision scores for predicting "yes" than "no", because it tends to underestimate the number of "yes" responses due to the imbalance.

TABLE V
PERCENTAGES OF "YES" RESPONSES FOR EACH NEWS OUTLET

ABC	CBS	NBC	CNN	Fox News	MSNBC	PBS
0.379	0.356	0.365	0.343	0.389	0.27	0.15

Notwithstanding the gap in accuracy across metrics, our model is robust in the sense that it is less likely to recommend an individual a source they will not like. Thus, if our model predicts that the consumption of a particular news outlet is likely, we have greater confidence that it is actually consumed by the individual.

V. Discussion

Our findings have important implications for research in political science and computer science. By understanding the factors that influence individual media preferences and the content they are exposed to, one can examine the processes of public opinion formation outlined in Zaller's [28] RAS model. Our study also sheds light on the perceived salience of political issues, as the media's agenda-setting function may vary based on an individual's predicted media exposure.

Furthermore, our research contributes to the ongoing scholarly debates on media bias and polarization. By identifying the political leaning of the news information individuals receive, we can assess the potential for the media to foster polarization and populist tendencies. One potential use case of our design would be to begin to recommend the next similar news source to individuals, encouraging them to break filter bubble behaviors and could help mitigate polarization and media bias.

Our findings also have implications for understanding media influence on voting intentions, as exposure to specific types of news content may shape political attitudes and behavior. By associating different news sources with their coverage of candidates and political leaning, we could provide insights into the general public's current position on national or local elections.

For computer scientists, our findings can be used to build more advanced models such as large language model pipelines. We can fine-tune an LLM that can consider people's personality and media habits, and then use that information to recommend news sources to the individual. This allows the model to take into account established consumption patterns of the individual with specific news sources, and therefore, such technology may be more effective in informing citizens or debunking misinformation.

We selected the classical ML models because we were constrained by the size of the dataset, and classical ML models have been shown to perform better in such cases. Although we did not examine complex deep-learning based models, we tested basic neural networks and transformer models. Neural network models and transformer models did not perform well for this dataset, and it is likely that with a larger dataset and more epochs, it may exhibit better performance.

VI. Conclusion

In this study, we developed an algorithm that predicts individual media preferences based on demographic information and media use habits, using data from the 2022 Cooperative Election Study. It was determined that Random Forests provides the best results, and tends to underpredict rather than overpredict positive responses, which is the preferred behavior when compared to the latter. Our findings highlight the importance of considering individual-level factors when studying media and politics.

Moving forward, our research will focus on refining and validating the predictive algorithm by incorporating additional datasets from diverse sources to enhance its generalizability. We will integrate behavioral data, such as social media interactions and temporal consumption patterns, to provide a more comprehensive understanding of media use habits. Advanced content analysis techniques, including multimodal data and contextual sentiment analysis, will be employed to improve the accuracy of our insights. Field experiments will be conducted to assess the real-world impact of our predictions on public opinion and political behavior, potentially influencing media regulation and policy. We will foster interdisciplinary collaborations to incorporate diverse perspectives and methodologies, and engage with stakeholders to disseminate findings and gather practical feedback. Ethical considerations, such as privacy, data security, and bias mitigation, will remain a priority throughout the research process. These steps aim to enhance the robustness and applicability of our algorithm, contributing to a deeper understanding of the interplay between media exposure and public opinion.

ACKNOWLEDGMENT

Thank you to SICSS-Rochester for this opportunity.

References

[1] Althaus, S. L., & Tewksbury, D. (2000). Patterns of Internet and traditional news media use in a networked community. Political Communication, 17(1), 21-45.

[2] Bennett, J., & Lanning, S. (2007, August). The Netflix prize. In Proceedings of KDD cup and workshop (Vol. 2007, p. 35).

[3] Bergström, A., & Jervelycke Belfrage, M. (2018). News in social media: Incidental consumption and the role of opinion leaders. Digital Journalism, 6(5), 583-598.

[4] Bharadhwaj, H., Park, S., & Luo, B. (2018). Personalized news recommendation using hierarchical attention. In Proceedings of the 2018 IEEE International Conference on Big Data (pp. 2321-2330). IEEE.

[5] Bimber, B. (2014). Digital media in the Obama campaigns of 2008 and 2012: Adaptation to the personalized political communication environment. Journal of Information Technology & Politics, 11(2), 130-150.

[6] Chan, T. W., & Goldthorpe, J. H. (2007). Social status and newspaper readership. American Journal of Sociology, 112(4), 1095-1134.

[7] Dilliplane, S. (2011). All the news you want to hear: The impact of partisan news exposure on political participation. Public Opinion Quarterly, 75(2), 287-316.

[8] Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. Public Opinion Quarterly, 80(S1), 298-320.

[9] Gentzkow, M., & Shapiro, J. M. (2010). What drives media slant? Evidence from U.S. daily newspapers. Econometrica, 78(1), 35-71.

[10] Gil de Zúñiga, H., Jung, N., & Valenzuela, S. (2012). Social media use for news and individuals' social capital, civic engagement and political participation. Journal of Computer-Mediated Communication, 17(3), 319-336.

[11] Guo, L., & McCombs, M. (2011, May). Network agenda setting: A third level of media effects. In International Communication Association Conference, Boston, MA.

[12] Iyengar, S., & Kinder, D. R. (2010). News that matters: Television and American opinion. University of Chicago Press.

[13] Jackob, N. (2010). No alternatives? The relationship between perceived media dependency, use of alternative information sources, and general trust in mass media. International Journal of Communication, 4, 589-606.

[14] Karimi, M., Jannach, D., & Jugovac, M. (2018). News recommender systems–Survey and roads ahead. Information Processing & Management, 54(6), 1203-1227.

[15] Knobloch-Westerwick, S., & Alter, S. (2007). The gender news use divide: Americans' sex-typed selective exposure to online news topics. Journal of Communication, 57(4), 739-758.

[16] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30-37.

[17] Ksiazek, T. B., Malthouse, E. C., & Webster, J. G. (2010). News-seekers and avoiders: Exploring patterns of total news consumption across media and the relationship to civic participation. Journal of Broadcasting & Electronic Media, 54(4), 551-568.

[18] Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). The people's choice: How the voter makes up his mind in a presidential campaign. Columbia University Press.

[19] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., ... & Van Alstyne, M. (2009). Computational social science. Science, 323(5915), 721-723.

[20] Lee, C. S., & Ma, L. (2012). News sharing in social media: The effect of gratifications and prior experience. Computers in Human Behavior, 28(2), 331-339.

[21] McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. Public Opinion Quarterly, 36(2), 176-187.

[22] Stroud, N. J. (2008). Media use and political predispositions: Revisiting the concept of selective exposure. Political Behavior, 30(3), 341-366.

[23] Stroud, N. J. (2010). Polarization and partisan selective exposure. Journal of Communication, 60(3), 556-576.

[24] Sunstein, C. R. (2017). #Republic: Divided democracy in the age of social media. Princeton University Press.

[25] Thurman, N., & Schifferes, S. (2012). The future of personalization at news websites: Lessons from a longitudinal study. Journalism Studies, 13(5-6), 775-790.

[26] Weeks, B. E., Lane, D. S., Kim, D. H., Lee, S. S., & Kwak, N. (2017). Incidental exposure, selective exposure, and political information sharing: Integrating online exposure patterns and expression on social media. Journal of Computer-Mediated Communication, 22(6), 363-379.

[27] Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. The ANNALS of the American Academy of Political and Social Science, 659(1), 6-13.

[28] Zaller, J. R. (1992). The nature and origins of mass opinion. Cambridge University Press.

[29] Prior, M. (2007). Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections. Cambridge University Press.

[30] Prior, M. (2013). Media and political polarization. Annual Review of Political Science, 16, 101-127.

[31] Opgenhaffen, M., & d'Haenens, L. (2015). Heterogeneity within homogeneity: Impact of online skills on the use of online news media and interactive news features. Communications, 40(1), 85-103.

[32] Rosengren, K. E., & Windahl, S. (1989). Media matter: TV use in childhood and adolescence. Ablex Publishing.

[33] Poindexter, P. M., & McCombs, M. E. (2001). Revisiting the civic duty to keep informed in the new media environment. Journalism & Mass Communication Quarterly, 78(1), 113-126.

[34] Van Zoonen, L. (2004). Imagining the fan democracy. European Journal of Communication, 19(1), 39-52.

[35] Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys (CSUR), 52(1), 1-38.

[36] van den Oord, A., Dieleman, S., & Schrauwen, B. (2013, October). Deep content-based music recommendation. In Advances in neural information processing systems (pp. 2643-2651).

[37] Zheng, G., Zhang, F., Zheng, Z., Xiang, Y., Yuan, N. J., Xie, X., & Li, Z. (2018, April). DRN: A deep reinforcement learning framework for news recommendation. In Proceedings of the 2018 World Wide Web Conference (pp. 167-176).

[38] Rubin, A. M. (2009). Uses-and-gratifications perspective on media effects. In J. Bryant & M. B. Oliver (Eds.), Media effects: Advances in theory and research (pp. 165-184). Routledge.