NSRI STUDENT RESEARCH INSTITUTION

ARTICLE TYPE:

Review Article

An Overlooked Risk: Investigating Name-Based Bias in TF-IDF-Based AI Spam Filters - A Literature Review

M. Neema

Ardrey Kell High School, Charlotte NC

Mukund Neema

ABSTRACT

As artificial intelligence becomes increasingly integrated into digital communication, ensuring that these systems operate fairly has become increasingly important. This literature review examines whether existing research suggests that TF-IDF-based email spam filters may exhibit bias against culturally diverse names by synthesizing studies on spam filtering, machine learning fairness, natural language processing, and name-based bias. The literature shows that machine learning systems can inherit bias from training data and that personal names may function as demographic signals, yet little direct research has examined whether these findings extend to TF-IDF-based spam filters. Rather than concluding that such bias exists, this review identifies an important gap in the literature and argues that current evidence provides a strong theoretical basis for future empirical research evaluating fairness in traditional text classification systems.

Keywords: TF-IDF; email spam filtering; machine learning fairness; artificial intelligence; natural language processing; name-based bias

MAIN MANUSCRIPT

1. Introduction

Email remains one of the most widely used forms of digital communication, with billions of messages exchanged every day for personal, academic, and professional purposes. Because manually reviewing every incoming message is impossible, email providers increasingly rely on machine learning algorithms to automatically distinguish legitimate emails from spam. These systems have become an essential component of modern communication by reducing unwanted advertisements, phishing attempts, malware, and other malicious content while improving the efficiency and security of email services (1; 2).

Over the past two decades, machine learning has fundamentally changed how spam filtering systems operate. Rather than relying on manually written rules or predefined keyword lists, modern spam filters learn statistical patterns from large collections of labeled emails and use those patterns to classify future messages (1). One of the most widely adopted techniques for representing textual information in these systems is Term Frequency-Inverse Document Frequency

(TF-IDF), which assigns numerical importance to words based on how frequently they appear within a document relative to an entire collection of documents (3). Its computational efficiency and effectiveness have made TF-IDF a foundational feature extraction technique in information retrieval and text classification, and it continues to play an important role in contemporary spam detection systems (2; 4).

Although these methods have achieved high levels of predictive performance, recent advances in artificial intelligence have also raised important questions regarding fairness. Machine learning systems do not understand language or social context in the way humans do. Instead, they identify statistical relationships within training data and use those patterns to make predictions. As a result, biases present in datasets may be learned and reproduced by AI systems, even when discrimination is not intentionally programmed into the model (5; 6; 7). Researchers have shown that bias can emerge from multiple stages of the machine learning pipeline, including data collection, feature representation, model development, and evaluation, making fairness an increasingly important consideration alongside traditional measures of model performance (7; 8).

A growing body of research has demonstrated that personal names can serve as demographic signals within artificial intelligence systems. Studies have found that language models may produce different outputs depending on the names presented to them and that low-frequency or culturally distinctive names often receive less reliable representations due to their limited presence in training data (9; 10). More broadly, decades of research have shown that names alone can influence important decision-making processes, demonstrating that they frequently carry cultural, ethnic, or linguistic information that may affect human and algorithmic judgments alike (11). These findings have motivated researchers to investigate fairness across demographic groups rather than relying solely on overall model accuracy when evaluating artificial intelligence systems (8).

Despite significant progress in both spam filtering and AI fairness research, comparatively little attention has been given to the intersection of these two fields. Existing studies on email spam filtering primarily evaluate systems based on classification accuracy, precision, recall, and computational efficiency, with relatively little discussion of demographic fairness or unintended disparities (1; 2; 4). At the same time, emerging evidence suggests that spam filtering systems themselves can exhibit measurable biases under certain conditions, highlighting the importance of evaluating fairness alongside predictive performance (12). However, to the best of the author's knowledge, little published research has directly examined whether TF-IDF-based spam filters may systematically treat emails differently based on culturally diverse names.

This gap in the literature motivates the present study. Rather than attempting to determine whether such bias definitively exists, this paper synthesizes existing research on statistical text representation, machine learning fairness, dataset bias, and name-based bias to evaluate whether the current literature suggests a plausible mechanism through which disparities could emerge. By examining findings across these related fields, this review aims to identify what is currently known, where important gaps remain, and why fairness should become a greater consideration in the future development and evaluation of AI-powered email spam filtering systems.

Accordingly, this paper addresses the following research question:

To what extent does existing literature suggest that TF-IDF-based email spam filters may exhibit bias against culturally diverse names?

2. Methods

2.1 Research Design

This study employs a qualitative literature review methodology to examine the relationship between TF-IDF-based email spam filtering, machine learning fairness, and name-based bias in artificial intelligence systems. Rather than conducting original experiments or collecting new datasets, this research synthesizes findings from existing peer-reviewed articles, conference papers, technical books, and survey papers to evaluate the extent to which current literature suggests that TF-IDF-based spam filters may exhibit bias against culturally diverse names. A literature review was selected because little research has directly investigated this specific question, making it appropriate to first establish the current state of knowledge and identify gaps that warrant future empirical investigation.

2.2 Literature Search Strategy

Relevant literature was identified through academic databases and reputable research repositories, including Google Scholar, the ACM Digital Library, IEEE Xplore, SpringerLink, Nature, arXiv, the Stanford Information Retrieval Book, and PubMed Central. Searches were conducted using combinations of keywords related to the research topic, including:

- "TF-IDF spam filtering"
- "email spam detection"
- "machine learning bias"
- "AI fairness"
- "natural language processing fairness"
- "dataset bias"
- "name bias in AI"
- "fairness in text classification"
- "algorithmic bias"
- "representation bias"

Priority was given to peer-reviewed journal articles, conference proceedings, and widely cited foundational works. Recent publications were also included to ensure the review reflected current developments in spam filtering and AI fairness research.

2.3 Source Selection

Sources were selected based on their relevance to at least one of four major themes identified during the literature search:

1. TF-IDF and email spam filtering.
2. Sources of bias in machine learning.
3. Name-based bias in artificial intelligence.
4. Fairness evaluation in AI and natural language processing.

Studies that focused exclusively on unrelated machine learning applications without contributing to these themes were excluded. Preference was given to papers that introduced foundational concepts, comprehensive surveys, or influential empirical findings that have shaped current understanding within the field.

2.4 Literature Analysis

Each selected source was analyzed to identify its primary research objective, methodology, key findings, and relevance to the research question. Rather than summarizing individual papers independently, findings were compared across studies to identify recurring themes, areas of agreement, differences in perspective, and gaps within the existing literature. Particular attention was given to how researchers defined fairness, the sources of bias they identified, and whether they examined the influence of demographic information, such as personal names, on machine learning systems.

The literature was then organized into four thematic categories: the foundations of TF-IDF-based spam filtering, sources of bias in machine learning, name-based bias in artificial intelligence, and fairness in AI classification systems. This thematic approach allowed findings from multiple research domains to be synthesized into a unified discussion that directly addressed the research question.

2.5 Limitations of the Methodology

Because this study is based exclusively on existing literature, its conclusions are limited by the availability of prior research. No original experiments, datasets, or statistical analyses were conducted to directly evaluate whether TF-IDF-based spam filters exhibit bias against culturally diverse names. Furthermore, relatively few studies examine this specific topic, requiring this review to draw connections between related fields, including spam filtering, natural language processing, and AI fairness. Consequently, the conclusions presented in this paper should be interpreted as a synthesis of current knowledge and a framework for future empirical research rather than definitive evidence of bias within existing spam filtering systems.

3. Results

Because few studies directly address this research question, this review draws from multiple related fields. The following sections examine existing literature on spam filtering, machine learning bias, name-based bias, and AI fairness to identify common findings and research gaps.

3.1 TF-IDF and Email Spam Filtering

Early email spam filters relied primarily on manually defined rules, such as blocking specific keywords, phrases, or sender addresses. Although effective against simple spam campaigns, these systems required continual manual updates and struggled to adapt as spam tactics evolved. Machine learning approaches addressed these limitations by learning statistical patterns directly from labeled data, enabling more accurate and adaptable spam detection (1). Among these approaches, TF-IDF emerged as one of the foundational techniques for representing email text in statistical spam classifiers and remains widely used because of its simplicity, interpretability, and computational efficiency.

Across the spam-filtering literature, evaluation focuses almost entirely on accuracy, precision, recall, and efficiency, with little attention to fairness or demographic disparities (1; 2; 4). Studies in this area commonly evaluate algorithms using benchmark datasets such as Enron, SpamAssassin, Ling-Spam, and the SMS Spam Collection. If these datasets do not adequately represent diverse naming conventions, this narrow evaluation lens may overlook an important dimension of model behavior, a possibility explored in the sections that follow.

3.2 Sources of Bias in Machine Learning

Because machine learning systems learn statistical relationships rather than social context, bias present in training data or development pipelines can be learned and reproduced unintentionally (5; 7). Mehrabi et al.[7] catalog multiple sources, including historical, representation, measurement, aggregation, and evaluation bias, while Hutchinson et al.[6] show bias enters NLP systems at every stage from data collection to evaluation. Gebru et al.[5] argue that poorly documented datasets obscure who is represented and what limitations may shape model behavior, and Selvam et al.[13] similarly emphasize improving data representativeness as one of the most effective ways to reduce bias before a model is even trained. Separately, Le Bras et al.[14] show models often exploit unintended statistical artifacts rather than the concepts they are meant to learn, meaning high accuracy does not guarantee unbiased decision-making.

None of this literature identifies TF-IDF itself as inherently biased; rather, it shows that representations, algorithms, and datasets interact to shape outcomes. If culturally diverse names are unevenly represented in spam-filter training data, this body of work suggests disparities could theoretically emerge through ordinary statistical learning, even though this has not been tested directly.

3.3 Name-Based Bias in Artificial Intelligence

Names are commonly treated as neutral identifiers, but they often encode cultural, ethnic, or linguistic information that models can learn to associate with patterns in the data (9). Smith and Williams[9] used names to measure and mitigate demographic bias in dialogue models, while Wolfe and Caliskan[10] found that language models represent low-frequency names less reliably, since such names are underrepresented in training data. Directly relevant to the present question, Hafner et al.[15] found that name-ethnicity classification algorithms themselves can systematically misclassify or perform unevenly across names associated with different ethnic groups, demonstrating that name-based disparities can be measured empirically. Outside AI, Bertrand and Mullainathan[11] found in a landmark field experiment that resumes with traditionally White-sounding names received significantly more callbacks than identical resumes with traditionally African American-sounding names, showing how names work as powerful demographic signals in both human and algorithmic decisions.

The literature on name-based bias and machine learning fairness focuses almost exclusively on large language models, generation tasks, and standalone name-classification systems; virtually no research examines whether comparable name-driven effects occur in traditional text classification systems such as TF-IDF-based spam filters. Given that names measurably shape behavior elsewhere in NLP, and that methods already exist for detecting name-based disparities in other classifiers (15), it is a reasonable, currently untested question whether they could apply here as well.

3.4 Fairness Evaluation in AI Classification

Accuracy, precision, recall, and F1-score do not by themselves indicate whether a model treats demographic groups equitably (8), a concern documented in NLP more broadly (16) as well as in text classification specifically, where diverse test sets are needed to surface disparities that standard benchmarks miss (17). Within spam filtering itself, Iqbal et al.[12] found measurable partisan differences in how political emails were classified during the 2020 U.S. election, direct evidence that spam filters can produce unequal outcomes, even though that study examined political content rather than names. Parallel findings in algorithmic hiring (18; 19) reinforce that historical patterns in training data can produce unequal outcomes even in highly accurate systems.

3.5 Synthesis and Research Gap

Taken together, the literature establishes three consistent findings. First, machine learning systems can inherit and reproduce biases present within training data and model development pipelines. Second, personal names often function as demographic signals that influence AI behavior across a variety of natural language processing tasks, including language models and dedicated name-classification systems. Third, fairness cannot be evaluated using traditional performance metrics such as accuracy alone, as models with strong predictive performance may still produce unequal outcomes across different groups.

Despite these findings, no study directly investigates whether TF-IDF-based email spam filters exhibit bias against culturally diverse names. Spam filtering research has largely focused on improving detection performance, while AI fairness research has concentrated on language models, recommendation systems, hiring algorithms, and other machine learning applications. As a result, these two areas have developed largely independently, leaving an important gap in the existing literature. Rather than suggesting that bias exists, this review identifies a question that has received little direct attention despite the widespread use of TF-IDF-based spam filtering systems. Figure 1 illustrates how these separate research areas converge to motivate the research question examined in this review.

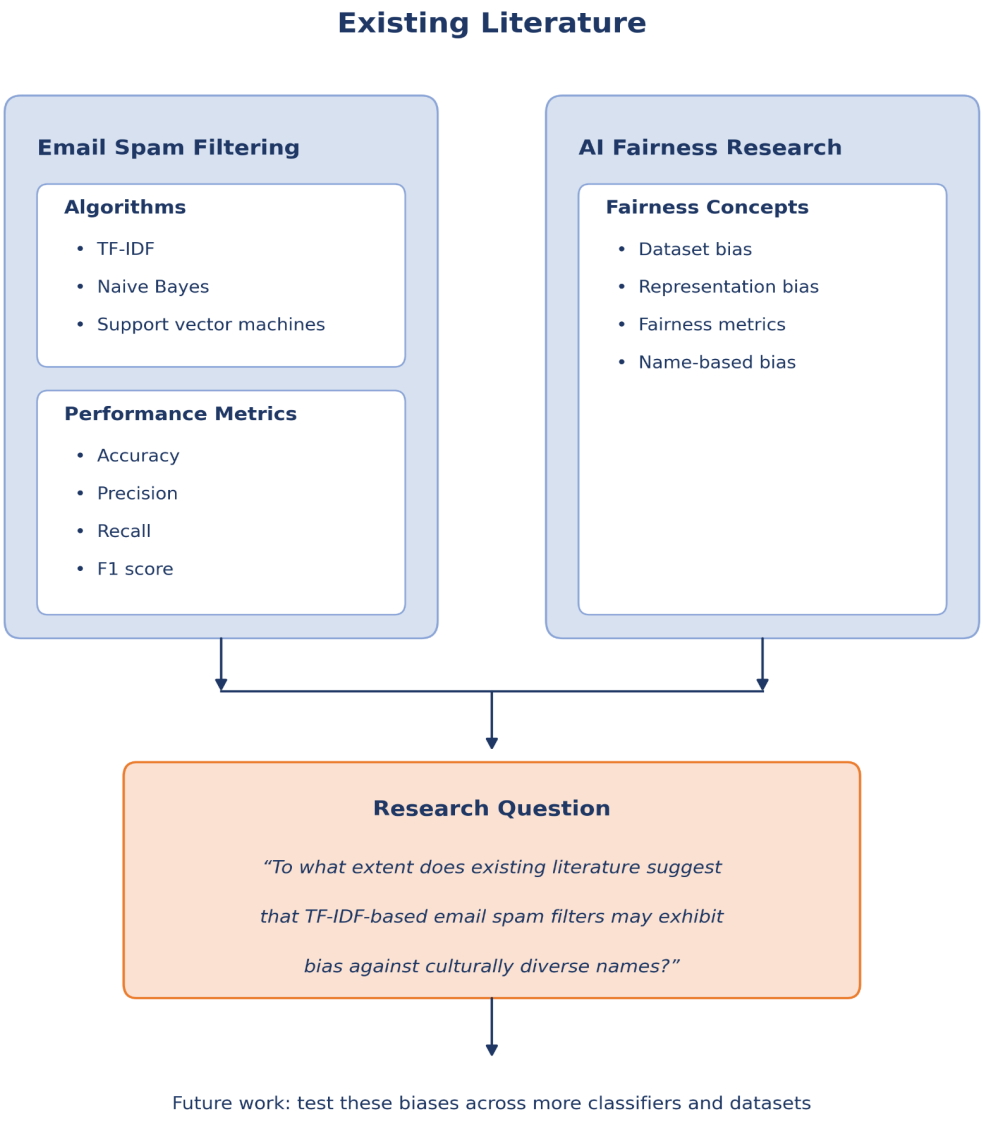

Figure 1.

Relationship Between Existing Research Areas and the Identified Research Question

Figure 1 summarizes how the major themes identified throughout the review converge on the research question this paper addresses.

4. Discussion

The literature reviewed throughout this paper suggests that the question of fairness in TF-IDF-based email spam filters deserves greater attention than it has received. While no existing study directly concludes that these systems are biased against culturally diverse names, the broader body of research provides several reasons why this possibility should not be dismissed. Rather than pointing to a single flaw in TF-IDF itself, the literature consistently shows that machine learning systems are shaped by the statistical patterns they learn from training data. If those patterns are incomplete or unrepresentative, unintended disparities may emerge even when the underlying algorithm functions exactly as designed.

One of the most interesting observations from this review is the disconnect between two mature areas of research. On one hand, spam filtering research has made significant progress in improving accuracy, precision, recall, and computational efficiency. On the other hand, AI fairness research has extensively explored how bias can arise from datasets, feature representations, and evaluation methods. Despite these parallel developments, very little work connects the two fields. As a result, fairness has not become a routine part of evaluating traditional spam filtering systems, even as it has become a central concern in many other areas of artificial intelligence.

The literature also highlights the importance of names as demographic signals. Existing studies demonstrate that names can influence AI systems in applications ranging from language models to text generation, particularly when certain names appear less frequently in training data. Although these findings cannot be directly applied to TF-IDF-based spam filters, they raise an important question: if names can influence model behavior in other natural language processing tasks, should researchers also examine whether similar patterns exist in email classification? Answering this question would improve our understanding of whether fairness concerns identified in modern AI systems also extend to traditional text classification methods that continue to be widely used.

Another important takeaway is that accuracy alone should not be considered a complete measure of model quality. A spam filter may correctly classify the overwhelming majority of emails while still producing small but consistent disparities for certain groups of users. These differences may not noticeably affect overall accuracy, yet they could still influence how individuals experience AI-powered communication systems. As artificial intelligence becomes increasingly integrated into everyday life, evaluating fairness alongside predictive performance will become an increasingly important part of assessing machine learning systems.

An important question raised by this review is why this issue has received so little attention. In recent years, AI fairness research has focused heavily on large language models, hiring algorithms, facial recognition, and recommendation systems because these technologies have obvious and highly visible societal impacts. Traditional machine learning systems, such as email spam filters, have largely remained in the background. Once they achieve high levels of accuracy, researchers often shift their attention toward developing newer models instead of asking whether existing systems are equally fair. However, millions of people still rely on spam filters every day, meaning even small disparities could affect a large number of legitimate emails over time. This suggests that fairness should be considered whenever AI systems make automated

decisions that influence everyday communication, regardless of how simple or established the underlying technology may appear.

Ultimately, this paper does not argue that TF-IDF-based spam filters are biased against culturally diverse names. Instead, it argues that the current literature leaves this question largely unanswered. Existing research provides strong theoretical reasons to investigate the possibility, but direct empirical evidence remains limited. By bringing together findings from spam filtering, machine learning fairness, and name-based bias, this review identifies an overlooked area of research and emphasizes the need for future studies that evaluate fairness within traditional text classification systems. As artificial intelligence continues to shape digital communication, understanding not only how accurately these systems perform but also how fairly they operate will become increasingly important.

5. Limitations

While this literature review provides a comprehensive synthesis of research related to TF-IDF-based spam filtering, machine learning fairness, and name-based bias in artificial intelligence, several limitations should be acknowledged. Recognizing these limitations is important for accurately interpreting the conclusions of this paper and identifying opportunities for future research.

First, this study is based entirely on existing literature and does not include original experiments or empirical testing. As a result, the conclusions presented are derived from previously published research rather than direct observations of real-world spam filtering systems. Although the reviewed literature provides a strong theoretical foundation, this paper cannot determine whether current TF-IDF-based email spam filters actually exhibit bias against culturally diverse names. Instead, it evaluates whether existing research suggests that such bias is theoretically plausible and worthy of further investigation.

A second limitation is the limited availability of research that directly addresses the specific research question. While substantial literature exists on email spam filtering, AI fairness, and name-based bias independently, very few studies examine the intersection of these three topics. Consequently, this review relies on synthesizing findings across multiple research domains to construct a broader understanding of the problem. Although this interdisciplinary approach helps identify important research gaps, it also requires drawing careful connections between studies conducted in different contexts and for different purposes.

Additionally, the findings discussed throughout this review are influenced by the datasets, evaluation methods, and application domains examined in the original studies. Many of the fairness papers focus on large language models, hiring algorithms, recommendation systems, or general natural language processing tasks rather than traditional email spam filtering. While these studies provide valuable theoretical insight into how bias can emerge in machine learning systems, their findings cannot be directly generalized to TF-IDF-based spam filters without additional empirical evidence.

Finally, the rapidly evolving nature of artificial intelligence represents an inherent limitation of this review. Advances in machine learning models, feature representations, and spam detection techniques continue to reshape the field, meaning that future developments may alter how fairness is evaluated or reduce the reliance on traditional approaches such as TF-IDF. Consequently, the conclusions presented in this paper should be interpreted within the context of the current body of literature rather than as permanent assessments of future spam filtering technologies.

Despite these limitations, this review provides a valuable synthesis of existing research and identifies an important gap within the current literature. By bringing together findings from multiple disciplines, this paper establishes a foundation

for future empirical studies that can directly investigate whether TF-IDF-based email spam filters exhibit disparities associated with culturally diverse names.

6. Future Research

The findings synthesized throughout this literature review highlight several opportunities for future research. While existing studies provide valuable insight into email spam filtering, machine learning fairness, and name-based bias independently, there remains little empirical evidence examining how these areas intersect. Addressing this gap will require research that directly evaluates whether traditional spam filtering systems produce different outcomes for emails associated with culturally diverse names.

One promising direction would be to conduct controlled experiments using balanced email datasets in which the sender's name is the primary variable being manipulated. By keeping the content, formatting, subject line, and other email characteristics consistent while varying only the cultural or linguistic background suggested by the sender's name, researchers could determine whether classification outcomes differ across demographic groups. Such experiments would provide the first direct empirical evidence addressing the research question examined in this review and help determine whether any observed disparities originate from the statistical representation of names rather than differences in email content. Establishing this evidence would also provide a foundation for evaluating fairness in traditional text classification systems more broadly.

Future research should also compare multiple text representation techniques rather than focusing exclusively on TF-IDF. Although TF-IDF remains widely used because of its simplicity and efficiency, modern spam detection systems increasingly incorporate word embeddings and transformer-based language models. Comparing TF-IDF with approaches such as Word2Vec, BERT, or sentence embeddings could help determine whether advances in language representation reduce, amplify, or otherwise change potential fairness concerns associated with traditional feature extraction methods.

Another important area for future research is the development of fairness evaluation benchmarks specifically designed for email spam filtering. Current studies primarily report metrics such as accuracy, precision, recall, and F1-score, while fairness metrics are rarely considered. Incorporating demographic fairness measures alongside traditional performance metrics would provide a more complete evaluation of spam filtering systems and help identify disparities that overall accuracy alone may overlook.

Finally, future studies should investigate the role of dataset composition in shaping spam filter behavior. Existing fairness literature suggests that the diversity and representativeness of training data significantly influence model performance. Examining how the frequency and distribution of culturally diverse names within training datasets affect classification outcomes could improve understanding of whether observed disparities originate from feature extraction methods, training data, or other components of the machine learning pipeline.

As artificial intelligence continues to influence digital communication, ensuring that automated systems operate both accurately and fairly will become increasingly important. By addressing the research questions identified throughout this review, future studies can help establish best practices for evaluating fairness in traditional text classification systems and contribute to the development of more transparent, equitable, and trustworthy AI-powered spam filtering technologies.

7. Conclusion

As artificial intelligence becomes increasingly integrated into everyday communication, ensuring that these systems operate fairly is becoming just as important as improving their accuracy. While email spam filters have become highly effective at detecting unwanted messages, existing research has focused primarily on optimizing classification performance through improvements in feature extraction, machine learning algorithms, and evaluation metrics. Comparatively little attention has been given to whether these systems produce equitable outcomes for all users.

This literature review examined the relationship between TF-IDF-based email spam filtering, machine learning fairness, and name-based bias in artificial intelligence. Collectively, the literature demonstrates that machine learning systems can learn and reproduce biases present within training data, that personal names may function as demographic signals influencing AI behavior, and that fairness should be evaluated alongside traditional performance measures. However, the review found little direct evidence examining whether TF-IDF-based email spam filters exhibit bias against culturally diverse names, highlighting an important gap in the current literature.

One of the most significant findings of this review is that spam filtering research and AI fairness research have developed largely independently. While spam filtering studies have concentrated on improving predictive performance, fairness research has focused on understanding and mitigating bias across a wide range of AI applications. Bringing these fields together highlights an overlooked question that has received relatively little attention despite the widespread use of email communication. Addressing this gap would improve our understanding of spam filtering systems and help researchers evaluate fairness in traditional machine learning models more thoroughly.

Ultimately, this paper does not argue that TF-IDF-based spam filters are inherently biased against culturally diverse names. Instead, it argues that the existing body of research provides sufficient theoretical motivation to investigate this possibility more directly. By synthesizing findings from spam filtering, machine learning fairness, natural language processing, and name-based bias research, this review provides a foundation for future empirical studies that can directly evaluate fairness in traditional text classification systems. As artificial intelligence continues to shape digital communication, ensuring that these systems are both accurate and fair will become increasingly important.

DECLARATIONS

Abbreviations

- TF-IDF – Term Frequency–Inverse Document Frequency
- AI – Artificial Intelligence
- NLP – Natural Language Processing
- BERT – Bidirectional Encoder Representations from Transformers

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The authors declare no competing financial interests or personal relationships that could have influenced this work.

Ethics Statement

This study was a literature review based exclusively on publicly available published research. No human participants, private identifiable data, or animal subjects were involved. Therefore, ethics approval was not required.

Data Availability Statement

No original datasets were generated or analyzed during this study. All information used in this review was obtained from publicly available published literature cited in the References section.

AI Use Statement

The author used ChatGPT (OpenAI) to assist with brainstorming, organization, and editorial feedback during manuscript preparation. All written content was independently reviewed, substantially revised, and verified by the author, who takes full responsibility for the accuracy, interpretation, and integrity of the manuscript.

Author Contributions

Mukund Neema: Conceptualization, methodology, literature search, investigation, analysis, visualization, writing - original draft, writing - review and editing.

Acknowledgements

The author gratefully acknowledges Dr. Nicolò Brandizzi for his valuable feedback and constructive suggestions during the preparation of this manuscript. His guidance significantly improved the organization, clarity, and overall quality of this review.

REFERENCES

1. Blanzieri E, Bryl A. A survey of learning-based techniques of email spam filtering. Artif Intell Rev. 2008;29(1):63-92.
2. Jáñez-Martino F, Alaiz-Rodríguez R, González-Castro V, Fidalgo E, Alegre E. Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach. arXiv. 2024. doi:10.48550/arXiv.2402.05296
3. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press; 2008.
4. Yang X, Zhou W, Duan X, Ma N, Wang Y. A spam detection model based on the discriminative TF-IDF belief rule base. Sci Rep. 2026;16:11962.
5. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86-92.
6. Hutchinson B, Smart A, Hanna A, Denton E, Greer C, Kjartansson O, et al. Towards accountability for machine learning datasets: practices from software engineering and infrastructure. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery; 2021. p. 560-575.
7. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54(6):115.
8. Dev S, Sheng E, Zhao J, Amstutz A, Sun J, Hou Y, et al. On measures of biases and harms in NLP. In: Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022. Association for Computational Linguistics; 2022. p. 246-267.
9. Smith EM, Williams A. Hi, my name is Martha: using names to measure and mitigate bias in generative dialogue models. arXiv. 2021. doi:10.48550/arXiv.2109.03300
10. Wolfe R, Caliskan A. Low frequency names exhibit bias and overfitting in contextualizing language models. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2021. p. 518-532.
11. Bertrand M, Mullainathan S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. Am Econ Rev. 2004;94(4):991-1013.
12. Iqbal H, Khan UM, Khan HA, Shahzad M. Left or right: a peek into the political biases in email spam filtering algorithms during US Election 2020. In: Proceedings of the ACM Web Conference 2022. Association for Computing Machinery; 2022. p. 2491-2500.
13. Selvam N, Dev S, Khashabi D, Khot T, Chang K-W. The tail wagging the dog: dataset construction biases of social bias benchmarks. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics; 2023. p. 1373-1386.
14. Le Bras R, Swayamdipta S, Bhagavatula C, Zellers R, Peters ME, Sabharwal A, et al. Adversarial filters of dataset biases. In: Proceedings of the 37th International Conference on Machine Learning. PMLR; 2020. p. 1078-1088.
15. Hafner L, Peifer TP, Hafner FS. Equal accuracy for Andrew and Abubakar—detecting and mitigating bias in name-ethnicity classification algorithms. AI Soc. 2023. Advance online publication.
16. Zadid TH, Jahin SM, Dhruba AR, Tanim SA, Hamja AM, Hasan MR, et al. A review of fairness challenges in natural language processing. Array. 2026;30:100773.
17. Djennane L, Kardkovács ZT, Benatallah B, Gaci Y, Gaaloul W, Farah Z. Enhancing bias detection in text classification models through high-quality diverse test cases. Inf Softw Technol. 2026;199:108253.
18. Fabris A, Baranowska N, Dennis MJ, Graus D, Hacker P, Saldivar J, et al. Fairness and bias in algorithmic hiring: a multidisciplinary survey. ACM Trans Intell Syst Technol. 2025;16(1):1-54.

19. Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR. Algorithms as discrimination detectors. Proc Natl Acad Sci U S A. 2020;117(48):30096-30100.