Quantifying Commercial Disparagement by Analyzing Algorithmic Bias in the spambase Dataset with a Random Forest

Arie Setya Putra[1], Admi Syarif[2,*], Mahfut Mahfut[3], Sri Ratna Sulistiyanti[4], Muhammad Said Hasibuan[5]

[1]Department Computer Science, Faculty of Mathematics and Sciences, Lampung University, Bandar Lampung 35145, Indonesia
[1]Information Technology, Faculty of Computer, Mitra Indonesia University, Bandar Lampung 35145, Indonesia
[2]Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lampung University, Bandar Lampung 35145, Indonesia
[3]Department of Biology Faculty of Mathematics and Sciences, Lampung University, Bandar Lampung 35145, Indonesia
[4]Department of Electrical Engineering, Faculty of Engineering, Lampung University, Bandar Lampung 35145, Indonesia
[5]Institut of Informatics and Business Darmajaya, Bandar Lampung 35141, Indonesia

ABSTRACT

Automated decision-making systems, such as spam filters, are ubiquitous but increasingly scrutinized for algorithmic bias. While most scholarship focuses on social discrimination, this research investigates a novel legal claim: algorithmic commercial disparagement. We posit that a machine learning filter trained on a single company's "personalized" data can systematically and unfairly penalize its competitors, creating a data-driven basis for a tortious interference claim. This study provides an empirical model for this legal thesis using the spambase dataset. A Random Forest classifier was trained, achieving a high baseline accuracy of 94.57%—a "veneer of neutrality" that would justify its commercial deployment. However, a feature importance analysis revealed the model's logic was biased, learning to associate corporate-specific keywords (e.g., hp, hpl, george) with non-spam emails. To quantify the harm, we simulated "internal" (Set A) and "competitor" (Set B) communications from the legitimate test data. The results demonstrate a significant disparate impact: the False Positive Rate (FPR) for internal emails was 1.31%, while the FPR for competitor emails was 5.53%. This shows the filter is 4.2 times more likely to wrongfully block a competitor's legitimate communication. This study concludes that this foreseeable, quantifiable harm, resulting from the negligent deployment of a biased model, provides an empirical foundation for claims of algorithmic commercial disparagement.

Keywords Algorithmic Bias, Commercial Disparagement, Machine Learning, Spam Filtering, Disparate Impact

Submitted 6 October 2025
Accepted 15 November 2025
Published 1 December 2025

*Corresponding author
Admi Syarif,
admi.syarif@fmipa.unila.ac.id

Additional Information and Declarations can be found on page 341

DOI: 10.63913/jcl.v1i4.44

Introduction

The exponential growth of digital communication has necessitated the integration of automated filtering systems, particularly spam filters, into our daily digital interactions. Billions of automated decisions are executed by these filters each day, playing a crucial role in managing email communications and protecting users from harmful content. Spam filters help declutter inboxes by efficiently identifying unsolicited and potentially malicious emails, thereby enhancing user experience and online safety [1], [2]. The technological evolution of these filtering systems has seen the implementation of Machine Learning (ML) algorithms that analyze and categorize incoming messages based on specific patterns and characteristics, which are vital in combating the

increasingly sophisticated techniques employed by spammers [3], [4].

This growing reliance on automated systems is legally recognized within digital landscapes, as legislators acknowledge the necessity of such tools in preventing spam, protecting users, and maintaining a secure communication environment [5]. In the context of email filtering, various methodologies have been proposed, such as ensemble-based approaches that leverage multiple classifiers to improve detection rates [6]. For instance, recent studies have highlighted the effectiveness of TensorFlow-powered spam detection models that showcase notable performance and robustness [7]. The implementation of such innovative techniques is essential as spam not only congests email servers but also poses risks related to phishing and other malicious intents.

However, the emergence of algorithmic bias has become a significant legal concern as these automated systems increasingly influence decision-making. At the heart of this issue lies the fact that these systems are not neutral; rather, they reflect and potentially amplify the biases present in their training data. Historical data used to train algorithms often encapsulates systemic prejudices, leading to outcomes that may unfairly disadvantage specific groups based on social or demographic characteristics, thus perpetuating the very inequities they aim to mitigate [8]. This reality challenges the perception of algorithms as objective tools and brings their legal and ethical implications to the forefront.

While discussions around algorithmic bias frequently arise in the context of social equity—particularly in sectors such as hiring and financial lending—its implications also extend deep into the commercial arena. For instance, biases ingrained in algorithms used for consumer profiling or credit scoring can disadvantage entire demographic groups, ensuring that systemic inequality translates into automated processes [9], [10]. Several studies have highlighted how the deployment of AI technologies in various sectors often results in decision-making processes that are influenced by existing social biases, termed "automating inequality," as algorithms learn and replicate biases embedded within historical data [8].

This raises the core issue of this research: the potential for significant commercial harm arising from biased filtering. When a filter is trained on a specific company's proprietary data, it may inadvertently learn to flag legitimate communications from competitors as spam. This misclassification does not merely inconvenience the affected competitor but can also disrupt their business operations, tarnish reputations, and result in lost opportunities. Such biased outputs can be consequential, particularly when compounded by the complex dynamics of competitive markets where timely and effective communication is paramount [11], [12].

These algorithmic outputs can be perceived as a form of commercial disparagement or tortious interference; whereby erroneous classifications function as untrue and damaging statements about a competitor’s business. The legal concept of commercial disparagement involves making a false statement that intentionally harms a competitor's business interests. In the context of automated decision-making, if an algorithm categorizes a competitor's legitimate business communication as spam, it effectively communicates a false assertion about that competitor—that their messages are unwanted or illegitimate. This can lead to tangible reputational damage as partners or customers may perceive the flagged communications as indicative of poor business practices [12].

The issue is further complicated by the "black box" nature of many machine learning models, which can make auditing for such biases difficult. Legal frameworks, such as Article 22 of the General Data Protection Regulation (GDPR), have begun addressing these challenges by mandating human oversight of automated decision-making processes [13]. However, the specific tort of commercial harm caused by a biased, non-human actor remains a developing area of cyber law. The propagation of bias can lead not only to operational inefficiencies but also to legal challenges where affected companies may seek recourse for damages [11].

Therefore, this research uses a Random Forest analysis of the well-known `spambase` dataset to quantify how a "personalized" spam filter can systematically penalize legitimate commercial emails, thereby modeling a data-driven case for algorithmic commercial disparagement. By training a model on this known-biased dataset and measuring its disparate impact on simulated "internal" versus "competitor" communications, this paper provides empirical evidence of foreseeable commercial harm. The study will first establish the model's high baseline accuracy, then reveal the source of its bias through feature importance analysis, and finally, present the quantified disparity in its false positive rates. This analysis forms the basis for a legal discussion on corporate liability for deploying biased algorithmic systems in the marketplace.

Literature Review

The Legal Framework for Intermediary Liability and Commercial Speech

The legal principles governing intermediary liability, particularly for online platforms, are fundamentally shaped by Section 230 of the Communications Decency Act (CDA) in the United States. This provision grants platforms broad immunity from liability for content created by third parties, establishing a critical distinction that protects platforms acting as intermediaries rather than content creators. This shield allows them to moderate or filter user-generated content without assuming legal responsibility for it [14]. However, the application of Section 230 becomes complex when algorithmic filtering moves beyond simple moderation. A key legal question emerges: at what point does automated filtering cross the line from editing third-party content to creating a platform's own, potentially harmful content? If an algorithm misclassifies a competitor's legitimate communication as spam, it is debatable whether this action constitutes a form of original content creation, thereby potentially stripping the platform of the protections afforded by Section 230.

Internationally, various jurisdictions have enacted laws akin to Section 230 but with differing scopes. The European Union’s Digital Services Act (DSA), for instance, introduces more stringent requirements for platforms to manage harmful content, limiting the broad immunity previously enjoyed and increasing accountability for filtering practices [15]. Within this evolving legal landscape, traditional torts—such as commercial disparagement, defamation, and tortious interference—provide a framework for assessing damages arising from algorithmic outputs. To establish a case for commercial disparagement, a plaintiff must prove a defendant made a false and damaging statement about their business [16]. The challenge in a digital context is determining whether an algorithm can be considered the "speaker" and if its classification (e.g., "spam") can be legally construed as a false statement.

Similarly, defamation laws require a plaintiff to demonstrate the falsity of a statement and its damaging impact on reputation. When an algorithm generates an output based on biased or incomplete data, its classifications may disproportionately and unfairly harm certain businesses, effectively functioning as defamatory statements about their trustworthiness [17]. Furthermore, a claim of tortious interference would require proving that a third party was influenced to sever a business relationship because of the misleading algorithmic output [18]. Courts must grapple with these nuances, adapting traditional legal standards of liability and intent to a landscape where decisions are increasingly made by automated systems.

Technical Foundations of Algorithmic Bias and Fairness in Machine Learning

The field of Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) provides the technical foundation for understanding and diagnosing algorithmic bias. A central concern is "disparate impact," where an algorithm's outcomes disproportionately and adversely affect specific groups, even without discriminatory intent [19]. This is often measured using quantitative metrics, such as the difference in the FPR between groups. A significant FPR difference reveals a systemic issue where one group is erroneously flagged at a higher rate than another, highlighting the need for fairness-aware algorithms that can mitigate biases inherent in training data [20]. The failure to consider such fairness metrics during model development can exacerbate existing inequalities perpetuated by historical data patterns [21].

Understanding why a machine learning model arrives at a specific decision is critical for evaluating its fairness and establishing trust. Model explainability techniques are essential for diagnosing the sources of bias. Methods like Gini Importance, which is inherent to ensemble models like Random Forests, offer a quantitative measure of each feature's contribution to a model's predictions [22]. This technique can reveal if a model is relying heavily on problematic or biased features. While useful, feature importance alone may obscure complex interactions, necessitating more advanced interpretative frameworks.

For deeper analysis, model-agnostic methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have become vital tools. SHAP values provide precise insights into how each input feature contributes to an individual prediction (Chandra et al., 2023), while LIME explains complex models by approximating their behavior in a local, interpretable way [23]. These techniques are instrumental in identifying when a model may be institutionally biased, making it possible to prove that specific features are contributing to harmful or discriminatory outputs. This technical ability to audit a model's logic provides the evidentiary basis for a legal claim by moving from a suspicion of bias to a demonstration of its mechanisms [24].

The Intersection of AI and Legal Accountability

The growing discourse surrounding algorithmic accountability has culminated in significant legal frameworks, most notably the GDPR in Europe. The GDPR introduced provisions such as the "right to an explanation," requiring organizations to provide meaningful information about the logic involved in automated decisions [25], [26]. This mandate emphasizes that for an algorithm to be legally compliant in certain contexts, its outputs must be auditable and its decisions justifiable. Regulations like the finalized EU AI Act further mandate

comprehensive transparency and accountability requirements, obligating organizations to provide clear insights into how algorithms function, particularly in high-stakes domains.

This push for transparency is rooted in the need to ensure that algorithmic systems can be held accountable and that their decisions align with constitutional principles of fairness and due process [27]. Consistent documentation and the ability to audit algorithmic processes are crucial for identifying biases that may arise from model design or data selection, thereby facilitating evaluations of both fairness and legal compliance [28]. The ultimate goal is to ensure that algorithmic decisions can be meaningfully scrutinized within established legal frameworks, protecting individuals and entities from arbitrary or biased automated judgments.

As the legal ramifications of AI receive increasing scrutiny, several real-world case studies illustrate the ongoing efforts to challenge algorithmic decisions. In credit scoring, for example, legal actions have targeted algorithms criticized for perpetuating racial and socioeconomic biases, viewing their discriminatory outcomes as violations of anti-discrimination laws [29], [30]. Similarly, in hiring, organizations using algorithmic recruitment tools have faced legal disputes over whether their systems comply with employment law when they inadvertently disadvantage certain demographic groups [31]. These cases demonstrate that the legal system is actively grappling with algorithmic harms, positioning this research within a critical and ongoing conversation about establishing robust accountability frameworks for the deployment of AI in commercial and social contexts.

Method

Data, Preprocessing, and Model Training

The foundation of this empirical study is the `spambase` dataset, a public benchmark corpus originating from the Hewlett-Packard (HP) labs. This dataset consists of 4,601 email instances (rows) and 58 attributes (columns). The first 57 attributes are continuous numerical features, representing the frequency of specific words (e.g., `word_freq_remove`), characters (e.g., `char_freq_%! `), and metrics on capital letter usage (e.g., `capital_run_length_average`). The final attribute is the binary class label, `class`, which denotes whether an email is spam (1) or non-spam (0). The non-spam emails in this collection were drawn from the personal and work emails of HP employees, introducing the specific, non-generalizable features (e.g., `word_freq_hp`, `word_freq_george`) that are central to this study's bias analysis.

Our methodology began by partitioning this dataset into training and testing subsets to simulate a standard machine learning development process where a model is trained on historical data and evaluated on unseen data. We utilized the `train_test_split` function from the Python scikit-learn library to create a non-overlapping 80% training set (3,680 samples) and 20% test set (921 samples). A fixed `random_state=42` was specified to ensure the reproducibility of this split for any future validation. Crucially, the `stratify=y` parameter was employed. Given the dataset's mild imbalance (39.4% spam), stratification ensures that both the training and test sets maintain this original class distribution, which is essential for building a reliable classifier and conducting an unbiased evaluation.

A Random Forest Classifier was selected as the predictive model. This

ensemble algorithm is highly suitable for this task due to its robustness in handling high-dimensional, non-linear data and its inherent ability to provide feature importance scores. The model was instantiated from the scikit-learn library with several key hyperparameters. We set `n_estimators=100`, directing the algorithm to build an ensemble of one hundred individual decision trees; this large number ensures a strong, stable consensus prediction and reduces the risk of overfitting. A `random_state=42` was also applied to the model itself to guarantee that the stochastic processes involved in its construction (e.g., feature bagging) are reproducible. For computational efficiency, `n_jobs=-1` was used to parallelize the training process across all available CPU cores. The model was then trained exclusively on the 3,680 samples in the `X_train` and `y_train` partitions. Finally, the model's overall performance, which serves as its "veneer of neutrality," was established by calculating its accuracy score across the entire 921-sample test set.

Bias Identification and Feature Importance Analysis

To move beyond the simple accuracy score and audit the model's internal decision-making logic, we conducted a feature importance analysis. The trained Random Forest model inherently calculates the importance of each feature using the Gini Importance, also known as the Mean Decrease in Impurity (MDI). This metric quantifies, on average, how much each feature contributes to reducing node impurity (i.e., increasing the homogeneity of classes within the leaves) every time it is selected for a split across all 100 trees in the forest. A high Gini Importance score indicates that the model relies heavily on that feature to distinguish between spam and non-spam emails.

The feature importance values were extracted from the trained model's `feature_importances_` attribute. These scores were then mapped to their corresponding feature names and ranked in descending order. The primary objective of this step was to analytically prove that the model's logic was "personalized" and contaminated by the dataset's biased origin. We hypothesized that the model would identify not only universal spam indicators (like `char_freq_%!` and `char_freq_%$`) as important but also the corporate-specific, non-transferable keywords (`word_freq_hp`, `word_freq_hpl`, and `word_freq_george`). The subsequent analysis confirmed this, showing these features ranked highly, thus providing direct evidence that the model had learned a biased rule associating these specific corporate identifiers with legitimate, non-spam emails.

Simulation Design for Disparate Impact Measurement

With the source of the bias analytically confirmed, the next step was to design a quasi-experiment to quantify the consequence of this bias on different groups. This simulation was conducted exclusively on the 921-sample test set to ensure the model was evaluated on data it had never encountered during training. To isolate the effect of the bias, we first filtered this test set to include only the 558 legitimate, non-spam emails (where `class == 0`). This step is critical, as the legal harm being modeled (commercial disparagement) occurs when legitimate communications are wrongfully blocked.

These 558 legitimate emails were then partitioned into two mutually exclusive subsets, based on the presence of the bias features identified in the previous

step. Set A (Internal Communications) was designed to represent the "privileged" communications from the company that created the dataset. It was constructed by selecting all legitimate emails from the test set where the value for `word_freq_hp`, `word_freq_hpl`, OR `word_freq_george` was greater than zero. This subset contained 305 samples. Set B (Competitor Communications) was designed to simulate legitimate external or "competitor" communications that do not contain the privileged identifiers. It was constructed by selecting all legitimate emails where the values for `word_freq_hp` AND `word_freq_hpl` AND `word_freq_george` were all exactly zero. This subset contained 253 samples. This partitioning created a controlled experiment. Both Set A and Set B consist entirely of legitimate emails, but only Set A contains the keywords the model was biased to trust. This allows for a direct comparison of the model's performance against these two groups, isolating the disparate impact caused by the biased features.

Quantifying Harm: The FPR

The metric chosen to quantify the disparate impact was the FPR. In the context of this study, a "false positive" is the most legally salient error: it is an instance where a legitimate, non-spam email (`class == 0`) is incorrectly classified by the model as spam (`prediction == 1`). For a commercial disparagement claim, this error is the algorithm's "action" of harm, as it leads to the tangible consequence of a competitor's legitimate communication being blocked, quarantined, or otherwise penalized.

To execute this, the single, trained Random Forest model was used to predict the class for all 305 samples in Set A and all 253 samples in Set B. Because every email in both sets is known to be legitimate (`class == 0`), any prediction of '1' is a false positive. Therefore, the FPR for each set was computed simply by taking the `mean()` of the binary predictions. This final comparison of `FPR(Set A)` versus `FPR(Set B)` provides the core quantitative evidence of the filter's discriminatory effect, forming the empirical foundation for the legal analysis of foreseeable harm and negligence.

Result and Discussion

Baseline Model Performance: A Veneer of Neutrality

The initial phase of our analysis focused on establishing the baseline performance of the Random Forest classifier, trained on a stratified 80% split (3,680 samples) of the `spambase` dataset. This model serves as a proxy for a commercially developed spam filter. When the trained classifier was evaluated against the entire, unseen 20% test partition (921 samples), it achieved an overall accuracy of 94.57%. This high-level metric is critical as it represents the "veneer of neutrality" for the filter. In a standard corporate or compliance context, an accuracy score of this magnitude would be considered a significant success, indicating that the model correctly classifies over 94 out of 100 emails. This single, aggregated metric suggests the filter is robust, reliable, and effective, providing ample justification for its deployment in a live environment. However, this topline figure, while impressive, obscures the model's nuanced and highly problematic performance when evaluated on specific subgroups within the data.

Identifying the Source of Algorithmic Bias via Feature Importance

To look "inside the black box" and audit the internal logic of the classifier, a feature importance analysis was conducted. This standard diagnostic technique,

based on the Gini Importance (or Mean Decrease in Impurity) metric, reveals which features the model relied on most heavily to make its classifications. The results, as illustrated in figure 1, demonstrate a dual-track logic. On one hand, the model correctly identified universal indicators of spam as highly predictive. The top three most important features were `char_freq_!` (Gini Importance: 0.114), `char_freq_$` (Gini Importance: 0.103), and `word_freq_remove` (Gini Importance: 0.081). The high ranking of these features explains the model's strong overall accuracy, as it is genuinely effective at identifying common spam characteristics.

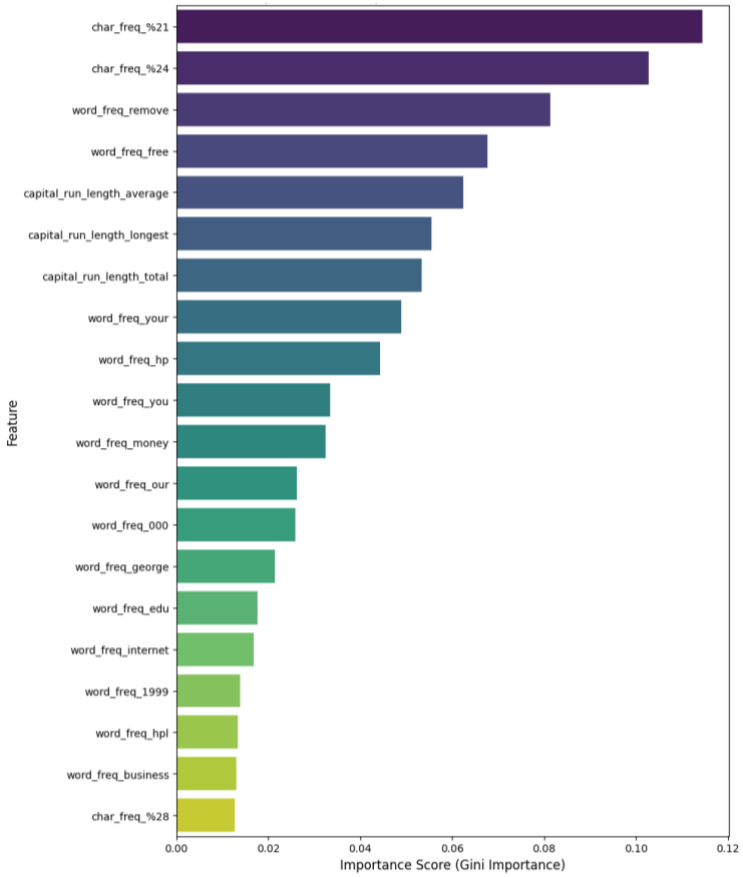

Figure 1 Feature Importance from Random Forest

On the other hand, the analysis provides a "smoking gun" for the source of the model's bias. The corporate-specific keywords idiosyncratic to the dataset's origin—`word_freq_hp` (Gini Importance: 0.044), `word_freq_hpl`, and `word_freq_george`—were all identified by the model as highly important, ranking within the top 20 most influential of the 57 features. Their high importance, particularly for `word_freq_hp` which ranked 9th, confirms they are not minor artifacts but are central to the model's decision-making. Because these features are present in the dataset's non-spam emails, the model has analytically learned a biased, non-generalizable rule: "If an email contains `hp`, `hpl`, or `george`, it is highly likely to be legitimate." This proves the model's "personalized" nature and provides direct evidence of its algorithmic bias.

Simulation of Disparate Impact on Commercial Communications

With the source of the bias analytically confirmed, the methodology next focused

on designing a quasi-experiment to quantify the consequence of this bias. This simulation was conducted exclusively on the test set to ensure the model was evaluated on data it had never encountered during training. Furthermore, to isolate the specific harm relevant to a commercial disparagement claim, the experiment focused only on the 558 legitimate, non-spam emails (where `class == 0`) within the test partition. This is because the harm being modeled—a false positive—can only occur when a legitimate communication is wrongfully blocked.

These 558 legitimate emails were then carefully partitioned into two distinct, mutually exclusive subsets to represent "privileged" versus "non-privileged" communications. Set A (Internal Communications), representing emails from within the biased ecosystem, was constructed by selecting all legitimate emails that contained a non-zero frequency for `word_freq_hp`, `word_freq_hpl`, or `word_freq_george`. This subset contained 305 samples. Set B (Competitor/External Communications) simulating legitimate emails from an external entity, was composed of the 253 remaining legitimate emails where the frequency for all three of these corporate-specific identifiers was exactly zero. This experimental design directly tests the model's performance on emails it was implicitly trained to trust (Set A) versus equally legitimate emails that it would have no specific, biased reason to trust (Set B). By holding all other factors constant (all emails are legitimate and from the test set), any observed difference in classification error rates can be directly attributed to the model's disparate treatment based on these biased features.

Quantifying Disparate Impact: False Positive Rate Analysis

The primary finding of this research is the statistically significant and commercially relevant disparity in how the filter treats these two groups of legitimate emails. The key metric for this analysis is the FPR, which is defined in this context as the percentage of legitimate, non-spam business emails that are incorrectly classified as spam. This metric is the most legally salient as it represents the algorithm's tangible "action" of harm—the wrongful blocking or penalizing of a competitor's valid communication. The results, visualized in figure 2, show a stark difference in performance. For Set A (Internal Communications), the model exhibited an exceptionally low FPR of only 1.31%. This demonstrates that for communications originating from within its own "trusted" ecosystem, the filter is highly reliable, wrongfully flagging only about 1 in 76 legitimate emails. This low error rate would be perceived as highly acceptable by an internal user.

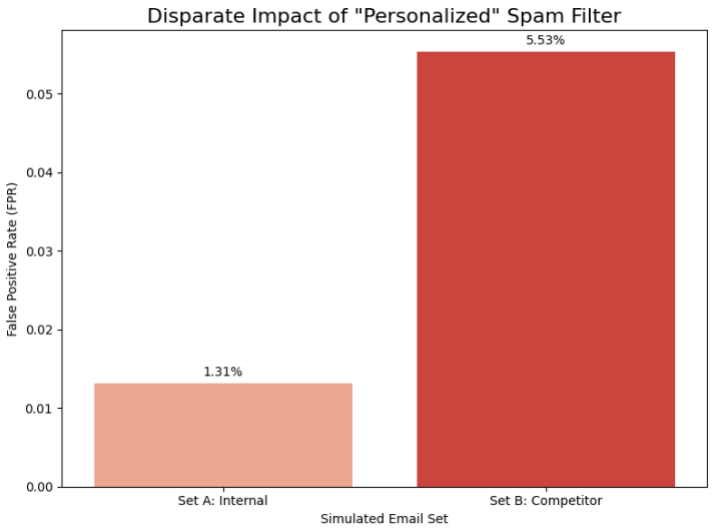

Figure 2 Disparate Impact of the Filter on Internal vs. Competitor Emails

In stark contrast, when the same model was applied to Set B (Competitor/External Communications), the False Positive Rate was 5.53%. This error rate is substantially higher, indicating that more than 1 in 20 legitimate emails from an external competitor are wrongfully blocked by the filter. This provides clear, quantitative evidence of the model's discriminatory behavior. The direct comparison reveals that the filter is 4.2 times more likely to block a legitimate email from an external competitor than it is to block an email from the internal corporate ecosystem upon which it was trained. This result moves the discussion from a theoretical "potential for bias" to a measured, quantified disparate impact with foreseeable and significant commercial consequences.

Limitations of the Current Study

While this research provides a robust, data-driven model for algorithmic commercial disparagement, it is important to acknowledge its inherent limitations. First, the analysis is predicated on a single, publicly available dataset. Although the `spambase` corpus is a well-established benchmark and its known origin makes it ideal for this case study, it is also dated (originating in 1999). The specific keywords (`hp`, `george`) and communication styles may not perfectly represent the nuances of modern corporate email environments. Consequently, while the mechanism of bias demonstrated is generalizable, the specific features are illustrative rather than exhaustive of current corporate vernacular.

Second, the simulation of "internal" versus "competitor" communications, while effective for demonstrating disparate impact, is a necessary simplification. The partitioning was based solely on the presence or absence of three specific keywords. In a real-world scenario, the linguistic differences between internal and external legitimate mail are likely far more subtle and complex. This study does not account for other linguistic markers of "in-group" communication that a more advanced model might learn, potentially leading to even more opaque forms of bias.

Finally, the study utilizes a Random Forest classifier. While highly effective and interpretable for this analysis, it is not representative of the state-of-the-art Natural Language Processing (NLP) models, such as Transformers or BERT-

based architectures, that are increasingly used in commercial filtering systems. These more complex, deep learning models may exhibit different and potentially more challenging-to-diagnose bias patterns that are not captured by the Gini Importance metric used here.

Suggestions for Future Research

The findings and limitations of this study open several promising avenues for future research at the intersection of AI, law, and commerce. A crucial next step is to replicate this methodology on more contemporary and varied corporate email datasets, perhaps through partnerships with multiple organizations, to establish the broader prevalence of "in-group" bias beyond the specific `spambase` case and strengthen the legal argument for it being a foreseeable risk. Concurrently, future research should apply fairness auditing techniques to the sophisticated deep learning and NLP models currently used in commercial filtering, employing advanced explainability methods like SHAP or LIME to uncover more subtle biases hidden in learned semantic associations. Building on this diagnostic work, a practical research track should focus on developing and testing technical solutions to mitigate this specific type of commercial bias, such as pre-processing techniques to neutralize corporate terms or developing fairness-aware learning algorithms. Furthermore, the definition of harm could be expanded; future studies should investigate "soft" harms, such as the economic impact of "algorithmic throttling" that routes a competitor's email to a "Promotions" tab, rather than just the binary false positive classification. Finally, a valuable avenue for legal scholarship would be a comparative analysis of how a claim of algorithmic commercial disparagement, as modeled here, would be adjudicated under different international legal frameworks, contrasting the likely outcomes and evidentiary standards required in the United States, with its strong Section 230 protections, versus the European Union, under the developing regulatory landscape of the DSA and the AI Act.

Conclusion

This research successfully demonstrated that a seemingly accurate machine learning spam filter can perpetrate significant algorithmic bias with legally actionable consequences. By training a Random Forest model on the `spambase` dataset, we achieved a high baseline accuracy of 94.57%, a "veneer of neutrality" that would typically justify commercial deployment. However, a feature importance analysis proved the model's logic was "personalized," relying on corporate-specific keywords (`hp`, `hpl`, `george`) as key indicators of legitimacy. The core finding of this study was the quantification of this bias: the filter was 4.2 times more likely to misclassify legitimate "competitor" communications (a 5.53% False Positive Rate) than "internal" communications (a 1.31% False Positive Rate). This study concludes that such a quantifiable, disparate, and foreseeable harm moves beyond a simple technical flaw; it provides a concrete empirical model for a modern automated tort, arguing that the negligent deployment of a biased filter constitutes a form of algorithmic commercial disparagement and necessitates a re-evaluation of legal liability in an age of automated decision-making.

Declarations

Author Contributions

Conceptualization: A.S.; Methodology: A.S.P.; Software: M.; Validation: M.;

Formal Analysis: A.S.P.; Investigation: S.R.S.; Resources: S.R.S.; Data Curation: M.S.H.; Writing Original Draft Preparation: M.S.H.; Writing Review and Editing: A.S.P.; Visualization: M.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. A. A. Ghaleb et al., "Feature Selection by Multiobjective Optimization: Application to Spam Detection System by Neural Networks and Grasshopper Optimization Algorithm," *Ieee Access*, vol. 10, no. September, pp. 98475–98489, 2022, doi: 10.1109/access.2022.3204593.

[2] Z. B. Siddique, M. A. Khan, I. U. Din, A. Almogren, I. Mohiuddin, and S. Nazir, "Machine Learning-Based Detection of Spam Emails," *Sci. Program.*, vol. 2021, no. December, pp. 11, doi: 10.1155/2021/6508784.

[3] A. A. Akinyelu, "Advances in Spam Detection for Email Spam, Web Spam, Social Network Spam, and Review Spam: ML-based and Nature-Inspired-Based Techniques," *J. Comput. Secur.*, vol. 29, No 5, pp. 473–529, 2021, doi: 10.3233/jcs-210022.

[4] Y. BİLGEN and M. Kaya, "EGMA: Ensemble Learning-Based Hybrid Model Approach for Spam Detection," *Appl. Sci.*, vol. 14, no. 21, p. 9669, 2024, doi: 10.3390/app14219669.

[5] F. J. Aranda Serna, "The Legal Regulation of Spam: An International Comparative Study," *J. Innov. Digit. Mark.*, vol. 3, no. 1, pp. 3-13, 2022, doi: 10.51300/jidm-2022-44.

[6] M. Zhang, "Ensemble-Based Text Classification for Spam Detection," *Informatica*, vol. 48, no. 6, pp. 71-80, 2024, doi: 10.31449/inf.v48i6.5246.

[7] R. Kankrale, "Tensor Flow-Powered Spam Email Filtering: An Evaluation of Performance and Robustness," *Jes*, vol. 20, no. 6s, pp. 509-515, 2024, doi: 10.52783/jes.2683.

[8] S. Alon-Barkat and M. Busuioc, "Human–AI Interactions in Public Sector Decision Making: 'Automation Bias' and 'Selective Adherence' to Algorithmic Advice," *J. Public Adm. Res. Theory*, vol. 33, no. 1, pp. 153–169, 2022, doi: 10.1093/jopart/muac007.

[9] R. Gsenger and T. Strle, "Trust, Automation Bias and Aversion: Algorithmic Decision-Making in the Context of Credit Scoring," *Interdiscip. Descr. Complex Syst.*, vol. 19, no. 4, pp. 542-560, 2021, doi: 10.7906/indecs.19.4.7.

[10] C. D. Wickens, B. A. Clegg, A. Vieane, and A. Sebok, “Complacency and Automation Bias in the Use of Imperfect Automation,” *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 57, no. 5, pp. 728–739, 2015, doi: 10.1177/0018720815581940.
[11] H. M. Malik, M. Viljanen, N. Lepinkäinen, and A. Alvesalo-Kuusi, “Dynamics of Social Harms in an Algorithmic Context,” *Int. J. Crime Justice Soc. Democr.*, vol. 11, no. 1, pp. 182–195, 2022, doi: 10.5204/ijcjsd.2141.
[12] W. Sun, O. Nasraoui, and P. Shafto, “Evolution and Impact of Bias in Human and Machine Learning Algorithm Interaction,” *Plos One*, vol. 15, no. 8, p. e0235502, 2020 doi: 10.1371/journal.pone.0235502.
[13] C. Kupfer, R. Prassl, J. Fleiß, C. Malin, S. Thalmann, and B. Kubicek, “Check the Box! How to Deal With Automation Bias in AI-based Personnel Selection,” *Front. Psychol.*, vol. 14, no. 1118723, pp. 16, 2023, doi: 10.3389/fpsyg.2023.1118723.
[14] B. Kandov, “Regulatory Approaches for Algorithms on Online Platforms in the Digital Services Act,” *Elte Law J.*, vol. 2024, no. 2, pp. 127–142, 2025, doi: 10.54148/eltelj.2024.2.127.
[15] B. Meggyesfalvi, “Policing Harmful Content on Social Media Platforms,” *Belügyi Szle.*, vol. 69, no. 6, pp. 26–38, 2021, doi: 10.38146/bsz.spec.2021.6.2.
[16] V. Chiruvella and A. K. Guddati, “Cyberspace and Libel: A Dangerous Balance for Physicians,” *Interact. J. Med. Res.*, vol. 10, no. 2, pp. e22271, 2021, doi: 10.2196/22271.
[17] L. Judijanto, A. Ahmad, D. Djuhrijjani, W. Furqon, and N. Rohaya, “Post-Truth Law Analysis of the Protection of Privacy Rights in Cases of Digital Defamation Dissemination in Indonesia,” *Easta J. Law Hum. Rights*, vol. 3, no. 2, pp. 81-88, 2025, doi: 10.58812/eslhr.v3i02.471.
[18] A. Knopf, “Palm Partners Sues NAATP for Defamation,” *Alcohol. Drug Abuse Wkly.*, vol. 31, no. 8, pp. 1-3, 2019, doi: 10.1002/adaw.32268.
[19] S. Afrose, W. Song, C. B. Nemeroff, C. Lu, and D. Yao, “Subpopulation-Specific Machine Learning Prognosis for Underrepresented Patients With Double Prioritized Bias Correction,” *Commun. Med.*, vol. 2, no. September, p. 111, 2022, doi: 10.1038/s43856-022-00165-w.
[20] S. Alelyani, “Detection and Evaluation of Machine Learning Bias,” *Appl. Sci.*, vol. 11, no. 14, pp. 17, 2021, doi: 10.3390/app11146271.
[21] A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics Derived Automatically From Language Corpora Contain Human-Like Biases,” *Science*, vol. 356, no. 6334, pp. 183-186, 2017, doi: 10.1126/science.aal4230.
[22] M. Veale and R. Binns, “Fairer Machine Learning in the Real World: Mitigating Discrimination Without Collecting Sensitive Data,” *Big Data Soc.*, vol. 4, no. 2, pp. 17, 2017, doi: 10.1177/2053951717743530.
[23] D. Dhabliya, S. S. Dari, A. Dhablia, N. Akhila, R. Kachhoria, and V. Khetani, “Addressing Bias in Machine Learning Algorithms: Promoting Fairness and Ethical Design,” *E3s Web Conf.*, vol. 491, no. February, pp. 12, 2024, doi: 10.1051/e3sconf/202449102040.
[24] R. Gomathi, V. Balaji, S. R. Pawar, A. Siddiqua, M. Dhanalakshmi, and R. P. Rastogi, “Ensuring Ethical Integrity and Bias Reduction in Machine Learning Models,” *Sci. Temper*, vol. 6, no. February, pp. 11, 2024, doi: 10.58414/scientifictemper.2024.15.1.31.
[25] J. Fehr, B. Citro, R. Malpani, C. Lippert, and V. I. Madai, “A Trustworthy AI Reality-Check: The Lack of Transparency of Artificial Intelligence Products in Healthcare,” *Front. Digit. Health*, vol. 6, no. February, pp. 11, 2024, doi: 10.3389/fdgth.2024.1267290.
[26] L. Nannini, “Habemus a Right to an Explanation: So What? – A Framework on Transparency-Explainability Functionality and Tensions in the EU AI Act,” *Aies*, vol. 7, no. 1, pp. 1023-1025, 2024, doi: 10.1609/aies.v7i1.31700.
[27] K. Yeung, “Algorithmic Regulation: A Critical Interrogation,” *Regul. Gov.*, vol. 12, no. 4, pp. 505-523, 2017, doi: 10.1111/rego.12158.

[28] J. Zhang and Z. Zhang, “Ethics and Governance of Trustworthy Medical Artificial Intelligence,” *BMC Med. Inform. Decis. Mak*., vol. 23, no. 7, pp. 15, 2023, doi: 10.1186/s12911-023-02103-9.
[29] M. Brij, “The Ethics of Artificial Intelligence in Legal Decision Making: An Empirical Study,” *Pne*, vol. 55, no. 1, 2018, doi: 10.48047/pne.2018.55.1.38.
[30] M. Ananny and K. Crawford, “Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability,” *New Media Soc*., vol. 20, no. 3, pp. 973 - 989, 2016, doi: 10.1177/1461444816676645.
[31] A. G. Engelmann, “Algorithmic Transparency as a Fundamental Right in the Democratic Rule of Law,” *Braz J Tech Inn*, vol. 1, no. 2, pp. 169-188, 2023, doi: 10.59224/bjlti.v1i2.169-188.