computers

MDPI

Article

Harnessing Machine Learning to Unveil Emotional Responses to Hateful Content on Social Media

Ali Louati [1], Hassen Louati [2], Abdullah Albanyan [3,*], Rahma Lahyani [4], Elham Kariri [1] and Abdulrahman Alabduljabbar [1]

[1] Information Systems Department, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia; a.louati@psau.edu.sa (A.L.); e.kariri@psau.edu.sa (E.K.); a.alabduljabbar@psau.edu.sa (A.A.)
[2] Computer Science Department, Kingdom University, Riffa 3903, Bahrain
[3] Software Engineering Department, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[4] Operations and Project Management Department, College of Business, Alfaisal University, Riyadh 11533, Saudi Arabia; rlahyani@alfaisal.edu
* Correspondence: a.albanyan@psau.edu.sa

Abstract: Within the dynamic realm of social media, the proliferation of harmful content can significantly influence user engagement and emotional health. This study presents an in-depth analysis that bridges diverse domains, from examining the aftereffects of personal online attacks to the intricacies of online trolling. By leveraging an AI-driven framework, we systematically implemented high-precision attack detection, psycholinguistic feature extraction, and sentiment analysis algorithms, each tailored to the unique linguistic contexts found within user-generated content on platforms like Reddit. Our dataset, which spans a comprehensive spectrum of social media interactions, underwent rigorous analysis employing classical statistical methods, Bayesian estimation, and model-theoretic analysis. This multi-pronged methodological approach allowed us to chart the complex emotional responses of users subjected to online negativity, covering a spectrum from harassment and cyberbullying to subtle forms of trolling. Empirical results from our study reveal a clear dose–response effect; personal attacks are quantifiably linked to declines in user activity, with our data indicating a 5% reduction after 1–2 attacks, 15% after 3–5 attacks, and 25% after 6–10 attacks, demonstrating the significant deterring effect of such negative encounters. Moreover, sentiment analysis unveiled the intricate emotional reactions users have to these interactions, further emphasizing the potential for AI-driven methodologies to promote more inclusive and supportive digital communities. This research underscores the critical need for interdisciplinary approaches in understanding social media's complex dynamics and sheds light on significant insights relevant to the development of regulation policies, the formation of community guidelines, and the creation of AI tools tailored to detect and counteract harmful content. The goal is to mitigate the impact of such content on user emotions and ensure the healthy engagement of users in online spaces.

Keywords: machine learning; social media; emotion; trolls

check for updates

Citation: Louati, A.; Louati, H.; Albanyan, A.; Lahyani, R.; Kariri, E.; Alabduljabbar, A. Harnessing Machine Learning to Unveil Emotional Responses to Hateful Content on Social Media. *Computers* **2024**, *13*, 114. https://doi.org/10.3390/computers13050114

Academic Editor: Paolo Bellavista

Received: 27 March 2024
Revised: 15 April 2024
Accepted: 22 April 2024
Published: 29 April 2024

1. Introduction

In the evolving landscape of social media, where interactions span a spectrum from constructive dialogue to harmful content, the detection and analysis of emotions in responses to hateful content have become paramount. The exponential growth of social media platforms has not only connected billions globally but also exposed users to increased cyberbullying, trolling, and various forms of online harassment. This digital environment necessitates the development of sophisticated methodologies to understand and mitigate the adverse effects of such negative interactions. Our study aims to delve into the emotional responses of social media users when confronted with hateful content, employing advanced artificial intelligence (AI) methodologies for fine-grained analysis.

The ubiquity of social media platforms, as evidenced by billions of active users across services like Facebook, YouTube, Twitter, and Reddit, underscores the critical role these platforms play in shaping public discourse. However, the saturation of these platforms has shifted the focus from quantitative growth to qualitative engagement, where user retention and active participation hinge on the quality of interactions. Amid this backdrop, cyberbullying and online harassment emerge as significant deterrents to positive user engagement, with personal attacks and trolling severely impacting the social media experience. These negative behaviors not only discourage active participation but also foster an environment of hostility and division.

Furthermore, the spread of digital disinformation by state-sponsored trolls and bots exacerbates the challenge of maintaining a healthy digital ecosystem. The distinction between human and automated agents of disinformation, along with their methodologies and targets, highlights the complex dynamics at play in online social networks. This complexity necessitates an interdisciplinary approach, combining social scientific insights with computational methods to untangle the web of digital disinformation and its impact on public discourse.

In the context of education, the significance of understanding emotional responses becomes equally critical. The transition to online learning, exacerbated by the COVID-19 pandemic, has brought the emotional well-being of students to the forefront. The analysis of student feedback through advanced AI methodologies, such as sentiment analysis, offers invaluable insights into the educational experience, revealing the challenges and obstacles faced by students in navigating their online learning environments.

Our study synthesizes these diverse strands of research to offer a comprehensive analysis of emotional responses to hateful content on social media. By integrating data-driven analysis of online personal attacks, the characterization of online trolling, and sentiment analysis of educational feedback, we aim to unveil the nuanced emotional landscape of social media users. Employing advanced AI methodologies, including high-precision attack detection technologies, psycholinguistic feature extraction, and sentiment analysis algorithms, our research seeks to contribute to the development of safer, more inclusive online communities. This interdisciplinary endeavor not only advances our understanding of social media dynamics but also informs platform regulation, policy-making, and the design of AI tools to detect and mitigate the impact of hateful content on user emotions.

2. Related Works

Revisiting the impact of social media on individual well-being and self-esteem reveals a nuanced landscape, particularly among young users. Early investigations in [1] on a Dutch SNS, CU2, among adolescents aged 10 to 19, underscored the pivotal role of feedback frequency and sentiment in shaping users' psychological state. This early study, surveying 881 participants, laid foundational insights into the dual-edged nature of social media interactions. While social networking sites (SNSs) have the potential to enhance life quality, a body of research has delineated the darker facets of frequent social media engagement. Beran and Li [2], Campbell et al. [3], Hinduja and Patchin [4], and Woods and Scott [3], among others, have documented the linkage between compulsive social media use and a spectrum of psychological adversities, including diminished self-esteem, anxiety, and depression. This trend manifests globally, with corroborative findings spanning from the UK [5], Turkey [6], and Japan [7] to Poland [8]. Recent explorations by Culpepper [9] into the interplay between social media habits and mental health among youth have unveiled a strong correlation between social media duration and depressive symptoms, albeit a weaker linkage with anxiety indicators. This points to the potential exacerbation of depressive states through social media dynamics. A granular examination in [10] into social media-induced fatigue highlighted the contributing factors, including privacy concerns, comparative behavior, and the intensity of platform engagement. Their study, involving a cohort of Indian young adults engaged in both full-time and distance education, identified a direct correlation between social media fatigue and academic performance, with privacy concerns and the fear of missing out (FOMO) playing less significant roles. The academic implications of WhatsApp usage have been scrutinized, yielding mixed outcomes. Studies by Afful and Akrong [11], Alkhalaf et al. [12], and Yeboah et al. [13] diverge in their conclusions, ranging from neutral to positive correlations with academic performance, contingent on the context of WhatsApp's application. Saiphoo et al. [14]'s meta-analysis sheds light on the nuanced relationship between SNS usage and self-esteem, highlighting a negative correlation, particularly pronounced in the context of problematic social media usage. The influence of online harassment predates the surge in SNS popularity, with research pivoting toward the predictive analysis of user engagement and retention. Pudipeddi et al. [15]'s investigation into churn predictors on StackOverflow revealed that temporal dynamics between postings serve as a significant churn indicator. This insight is pivotal for devising strategies to foster user loyalty and reduce churn, although the challenge of translating such predictors into actionable interventions remains. Hamilton et al. [16]'s exploration into Reddit communities offered a fresh perspective on user loyalty, associating it with thematic engagement and the structural dynamics of user interactions. Communities characterized by strong external interests and cohesive interactions tend to exhibit higher loyalty levels. Conversely, the adverse effects of verbal aggression online extend beyond user engagement to legal ramifications, with cyberbullying being subject to varying degrees of legal accountability across jurisdictions [17]. Moreover, social media platforms face financial and reputational risks due to association with harmful content, prompting a reevaluation of content moderation practices [18]. This literature synthesis underscores the complex interplay between social media use, user well-being, and community dynamics, revealing both the potential benefits and pitfalls of digital social interactions. The diversity of outcomes across studies highlights the need for continued investigation into the multifaceted impacts of social media on psychological health and social cohesion. The burgeoning field of disinformation studies underscores the critical need to confront the ramifications that disinformation agents exert on societies globally. Significant research, including works by Cresci [19] and Ferrara et al. [20], has elucidated the interactions between humans and bots on social platforms (Varol et al. [21]), the dissemination of low-credibility content by bots (Shao et al. [22]), and the amplification of divisive content online (Stella et al. [23]). Some studies also delved into how bots accentuate group conflicts and partisanship (Luceri et al. [24]), influencing major national and international events across various regions and contexts. In parallel, scholarly endeavors have sought to delineate trolls, adopting definitions that encapsulate objectives of disrupting online discourse through deceptive or inflammatory tactics (Al Marouf et al. [25]). This body of work encompasses the manipulation of dialogues (Uyheng et al. [26]) and extends to state-affiliated actors, notably under the Russian Internet Research Agency (Zannettou et al. [27]). Beyond the scope of Russian disinformation, trolls have also been described as individuals employed by marketing firms for political messaging (Ong and Cabañes [28]), illustrating the diverse nomenclature used to describe similar actors across different cultural settings. Efforts to categorize the multifaceted aspects of disinformation have resulted in detailed, albeit complex, taxonomies (Gorwa and Guilbeault [29]). However, the application of these classifications in computational contexts remains challenging. The conflation of bots and trolls in the literature, with trolls often depicted as human actors amplified by bots (Alsmadi and O'Brien [30]), underscores the need for clear definitional boundaries to aid in computational modeling. Acknowledging the insights of Ferrara et al. [31], we advocate for a nuanced lexicon to articulate disinformation and its propagators. Within the realm of social cybersecurity [32], it is imperative to refine theoretical models in tandem with the development of analytical tools. As the tactics of information operations evolve, so too do the strategies and agents behind them, necessitating an empirical approach to behavior observation for robust computational modeling. Thus, we propose differentiating trolling based on behavior rather than state affiliation, facilitating a more focused and consistent analytical framework. While identifying state-backed trolls is valuable, the dynamic nature of state interests suggests the advantage of concentrating on observable malicious interactions (Addawood et al. [33]). This approach not only aligns with the existing literature but also allows for the examination of trolling across a broad spectrum of actors. Our proposed hypotheses, grounded in the literature, aim to empirically and methodologically explore trolling through a psycholinguistic analysis of social media discourse (Uyheng and Carley [26]), consider its interactive nature for enhanced modeling accuracy (Zubiaga et al. [34]), and distinguish trolling from related concepts like hate speech and cyber-aggression (Rosa et al. [35]). By focusing on language and interaction, we seek to provide a more granular understanding of trolling as a distinct online behavior, potentially employed by both automated and human actors.

Additional applications have considered sentiment analysis in deciphering the emotional undertones and opinions prevalent on social media platforms, with a special emphasis on Twitter. Recent investigations have traversed a spectrum of topics, including reactions to the COVID-19 pandemic [36], the dynamics surrounding shared mobility services [37], and the exploration of emotional intelligence through social media engagements [38]. Machine learning has been widely considered to solve problems related to behaviors, such as traffic behavior [39], vehicle behavior [40], and human behavior [41]. In [42], the author conducted sentiment analysis on Arabic tweets discussing the Hajj pilgrimage over six years, employing machine learning and deep learning models. The results revealed sentiment patterns before, during, and after Hajj events, with BERT emerging as the most effective model for accurately classifying sentiment in Arabic text. The insights shed light on the emotional dynamics surrounding significant religious events on social media platforms. A pivotal study [36] utilized TextBlob and TwitterScrapper in Python to perform sentiment analysis on English tweets related to COVID-19, achieving a notable accuracy of 79.34%. This research extended its analysis to medical documents, employing the SVM and KNN algorithms, and reported an 80% accuracy rate. Despite its contributions, the research was constrained to Twitter's textual content, focusing on a narrow linguistic spectrum. In a similar vein, another study [37] leveraged a sentiment-emotion detection framework to analyze tweets concerning shared mobility, uncovering a preference among 67% of respondents for passenger reservation services over alternatives like public transport. The investigation noted a slight shift from neutral to negative sentiments during the pandemic, with common terms such as "Uber" and "Lyft" frequently appearing in the discourse. Further, research [38] applied sentiment analysis to social media content through a series of processes, including data preprocessing, feature extraction, and classification, revealing that LRA-DNN models achieved impressive accuracy and sensitivity rates. However, these studies encountered limitations, particularly in their scope, which was limited to Twitter and a selection of languages. Moreover, the employed methodologies may not fully capture the breadth of emotions and sentiments articulated in social media postings, suggesting a need for expanded research to enhance the accuracy and applicability of sentiment analysis across various social media platforms.

The burgeoning interest in sentiment analysis of Arabic text is propelled by the escalating use of social media and the voluminous Arabic textual content generated. This analytic pursuit has spanned diverse domains, including politics [43], e-commerce [44], and consumer feedback [45]. Challenges abound in Arabic sentiment analysis, chiefly due to the scarcity of Arabic-specific linguistic resources and the inherent complexity of the language. Efforts to surmount these obstacles have led to the development of text normalization methods [46] and the adoption of machine learning algorithms like SVM and decision trees for sentiment classification [47]. Despite these advancements, the field grapples with limitations, particularly in accurately reflecting the full spectrum of sentiments and emotions in Arabic text, compounded by the dearth of Arabic linguistic resources. Hence, there is a pronounced need for further exploration to refine sentiment analysis techniques for Arabic text and to cultivate comprehensive linguistic databases.

Exploratory research [48] into online education sentiment during the COVID-19 pandemic in China, leveraging Sina Weibo microblogs, offered insights into public perceptions across different pandemic phases. Additional studies [49–52] applied sentiment analysis to evaluate online course feedback and student reviews in higher education, highlighting the potential of sentiment analysis in enhancing educational quality. Nonetheless, these studies underscore the necessity for more sophisticated analytical techniques to fully grasp the range of emotions and sentiments in educational feedback.

The discourse on the integration of machine learning, notably SVM, into educational sentiment analysis [53–57] reveals a gap in the application of SVM to real-life Arabic textual data from higher education contexts. Despite SVM's efficacy in various domains, its application to Arabic sentiment analysis remains underexplored, highlighting a critical area for future research to bridge this gap and enhance the computational analysis of sentiment in educational settings. The intersection of sentiment analysis with educational feedback, as explored by Louati et al. (2023) [58], introduces a novel dimension to our literature review. The development of the SVM Sentimental Analysis for Arabic Students' Course Reviews (SVM-SAA-SCR) algorithm represents a significant advancement in understanding the nuanced perspectives of students in non-English speaking contexts, specifically within Saudi Arabian universities. This study not only fills a critical gap by providing insights into students' perceptions post-COVID-19 but also showcases the potential of AI methodologies in extracting and analyzing sentiment from textual data across diverse linguistic settings. Table 1 presents a summary of related works focusing on the impact of social media and AI on user well-being and behavior.

Table 1. Summary of related works focusing on the impact of social media and AI on user well-being and behavior.

Reference	Main Issue	Methodology	Key Findings
Valkenburg et al. [1]	Psychological impact of SNSs on adolescents	Survey	Feedback frequency and sentiment significantly affect youth self-esteem
Beran and Li [2]	Link between social media use and psychological adversity	Literature review	Compulsive use linked to decreased self-esteem, anxiety, and depression
Culpepper [9]	Social media habits and mental health among youth	Statistical analysis	Strong correlation between social media use duration and depressive symptoms
Malik et al. [10]	Social media-induced fatigue and academic performance	Survey and correlation analysis	Direct correlation found between social media fatigue and academic performance
Afful and Akrong [11]	WhatsApp usage and academic performance	Comparative study	Findings suggest a neutral to positive correlation with academic performance
Ferrara et al. [20]	Human–bot interactions on social platforms	Data mining	Bots amplify low-credibility content and divisive content
Zannettou et al. [27]	State-affiliated trolling	Content analysis	Trolls disrupt discourse through inflammatory tactics

This research aims to contribute to the existing body of knowledge by offering a comprehensive framework for detecting and analyzing emotional responses to hateful content on social media. By incorporating insights from the psychological impacts of social media use, the mechanisms of online harassment, the challenges posed by digital disinformation, and the innovative application of sentiment analysis in educational feedback, we seek to provide a nuanced understanding of user emotions in the face of negative online interactions. Through this interdisciplinary approach, we endeavor to inform the development of more effective AI tools and strategies for identifying, understanding, and mitigating the emotional impacts of hateful content, thereby fostering healthier and more inclusive online communities.

3. Contributions

This research makes significant strides in the study of social media dynamics, particularly in understanding and mitigating the emotional impact of hateful content, distilled into three core contributions:

1. Comprehensive Analysis through AI Integration: By fusing the study of online personal attacks and trolling detection, this research employs an AI-driven framework to offer a nuanced understanding of users' emotional responses to negative online behaviors. This approach allows for a detailed examination of how personal attacks and trolling affect user engagement and emotional well-being.
2. Innovative Methodologies for Detecting and Characterizing Trolling: Through the development and application of psycholinguistic models and sentiment analysis algorithms, this study provides new insights into the nature of trolling and its differentiation from other forms of online aggression. It highlights the interactive aspect of trolling and its implications for both human and automated social media accounts, contributing to the development of more effective detection and mitigation strategies.
3. Strategic Contributions to Social Media Management and Policy: The findings offer actionable insights for social media platform regulation, the creation of AI tools to detect hateful content, and policy-making aimed at fostering inclusive online communities. Additionally, by addressing the limitations of self-reported data, this research advocates for more accurate measurement techniques, enhancing our understanding of the behavioral impacts of online negativity.

These contributions represent a significant advancement in the interdisciplinary approach to combating the challenges posed by hateful content on social media, with implications for improving the user experience and emotional health online.

The methodology employed in this study is visually summarized in the flowchart presented in Figure 1.

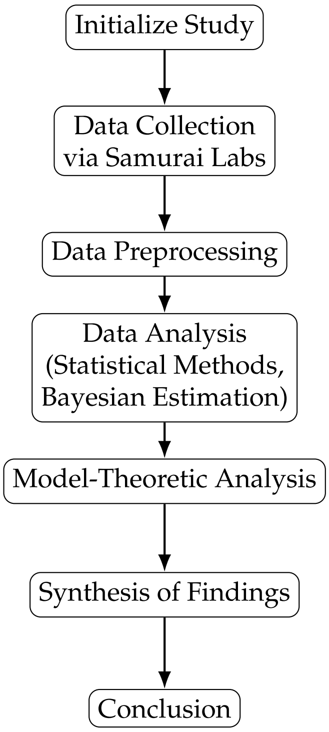

Figure 1. Methodology of this study.

4. Technology Utilized for Personal Attack Detection

In the context of this investigation, personal attacks are construed as derogatory remarks targeting individuals rather than the substance of their arguments. Such attacks include insults, comparisons to animals or objects, and insinuations without evidence (e.g., "You are legit mentally retarded homie", "Eat a bag of dicks, fuckstick", and "Fuck off with your sensitivity you douche"). The detection of these personal attacks was conducted through the utilization of Samurai, an in-house technology developed by Samurai Labs [59,60]. This technology integrates symbolic and statistical methodologies to analyze text data, with the symbolic components assessing contextual nuances and the statistical components employing deep learning models for classification.

Samurai's approach involves decomposing the problem of personal attack detection into sub-problems represented by language phenomena, such as speech acts, which are then independently analyzed using precise contextual models. For instance, personal attacks are categorized into mid-level categories like insults, animal/object comparisons, and insinuations, each of which can be further subdivided into low-level categories. Symbolic rules are employed to distinguish abusive language from regular discourse, while statistical models trained on extensive data aid in classification.

The detection process is divided into "narrow" and "wide" models, with the former offering high precision but limited coverage and the latter providing broader coverage at the expense of precision. To enhance analysis granularity, this study focused on models targeting personal attacks against individuals rather than public figures. A rigorous evaluation process was undertaken, involving manual verification by expert annotators to ensure high precision and recall rates.

Furthermore, additional experiments were conducted to assess Samurai's performance on different datasets. An evaluation of a sample of Reddit posts annotated for personal attacks yielded a recall rate of 74%, demonstrating Samurai's efficacy in identifying true positives. Additionally, an experiment involving Discord messages containing vulgar language but no personal attacks resulted in a low false-positive rate of 2%, indicating a high level of specificity.

In summary, Samurai's innovative approach to personal attack detection, integrating symbolic and statistical methods, showcases its effectiveness in accurately identifying abusive language while minimizing false alarms. The technology's robust performance across diverse datasets underscores its potential for mitigating verbal violence in online discourse.

5. Social Media Gathered Data

To achieve the outlined contributions, we conducted a comprehensive analysis leveraging AI integration to examine the impact of personal attacks and trolling on user activity in social media. The following section details our methodology for data collection, which involved large-scale quantitative analysis of Reddit user engagement. In the following, statistical machine learning tools are discussed in detail.

5.1. Study Design and Data Collection

The raw datasets utilized in this study were obtained through Samurai Labs, which collected Reddit posts and comments without moderation or removal. Data were sourced from the data stream provided by pushshift.io, facilitating real-time access to unmoderated content. Samurai Labs deployed personal attack recognition algorithms to identify instances of personal attacks, ensuring the integrity of the dataset.

Given the ethical considerations surrounding the experimental manipulation of personal attacks, our approach was observational. This method allowed for a broad and diverse sample, addressing the limitations often associated with WEIRD (Western, Educated, Industrialized, Rich, and Democratic) groups typically studied in psychology.

Data collection spanned approximately two continuous weeks, with specific days chosen to mitigate activity variations. A weekend day (27 June 2020) and a working day

(2 July 2020) were randomly selected. These days provided insights into user behavior, with slight adjustments made to account for weekend activity patterns.

Sampling involved randomly selecting 100,000 posts or comments for each selected day, resulting in datasets comprising 92,943 comments by 75,516 users for the weekend day and 89,585 comments by 72,801 users for the working day. Among these users, a subset experienced personal attacks, with 1.79% and 0.39% receiving narrow attacks on the weekend and working days, respectively.

To ensure balanced treatment and control groups, users were categorized based on the presence and frequency of attacks. The treatment groups included users experiencing one or more narrow attacks, while the control groups comprised users without recognized attacks during the sampling period.

Following data preparation and cleaning, which involved removing suspected bots and outliers, the final dataset consisted of 3673 users aligned around the selection day (day 8), with associated posts, comments, and interactions.

This methodology facilitated a comprehensive examination of user engagement and emotional responses to online negativity, supporting the development of innovative detection strategies and informing social media management policies. Further details and technical documentation are available online for comprehensive understanding.

5.2. Initial Data Exploration

Our initial analysis focused on understanding the dynamics between the frequency of targeted attacks and subsequent changes in online activity. We visualized this by plotting the difference in weekly activity—measured through posts and comments—before and after the attacks. Each data point represented an individual user, with the *x*-axis showing the number of attacks received and the *y*-axis indicating the change in activity levels. The majority of users experienced no attacks, serving as a control group, while a smaller fraction encountered varying numbers of attacks. Two key trends emerged from our visualization: a linear regression line indicating a general decrease in activity with an increase in attacks, and a smoothing curve that highlighted local activity patterns without overfitting, both enveloped by 95% confidence intervals. We also calculated and visualized the proportional change in activity to account for the varying baseline activity levels among users, revealing a more pronounced negative impact from targeted attacks, especially narrower ones. In Figure 2, each point represents a user, with the *x*-axis showing attacks received before the incident and the *y*-axis showing the change in activity (posts and comments) from before to after the incident. The control group (0 attacks) is clearly distinguished, with a decreasing frequency of users experiencing 1, 2, 3, etc., attacks. The linear regression (dashed line) suggests a negative correlation between attacks and changes in activity. The generalized additive model (GAM) smoothing curve (gray line) reveals local patterns without overfitting, enclosed by 95% confidence intervals.

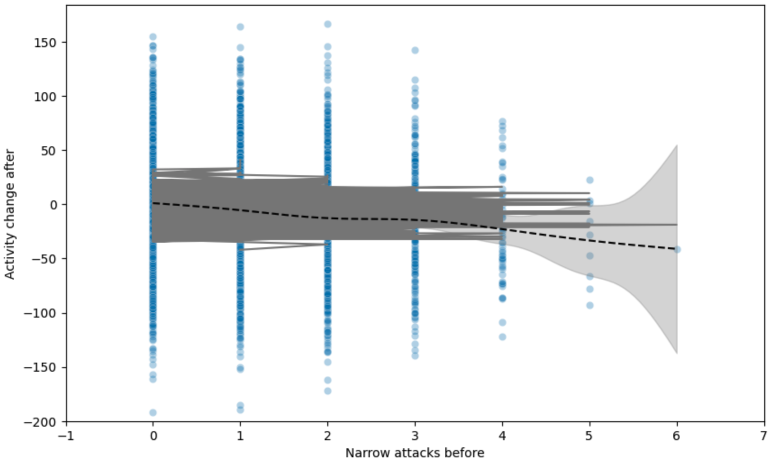

Figure 2. Changes in activity vs. attacks received.

Figure 3 presents the proportional change in activity as a function of attacks. It focuses on the proportionate change in activity, highlighting the impact of attacks on users with different baseline activities. The analysis reveals a more pronounced negative impact from narrower attacks. Insights reveal that the impact of attacks on user activity is negatively skewed, especially for narrow attacks, indicating a significant decrease in user engagement post-attack.

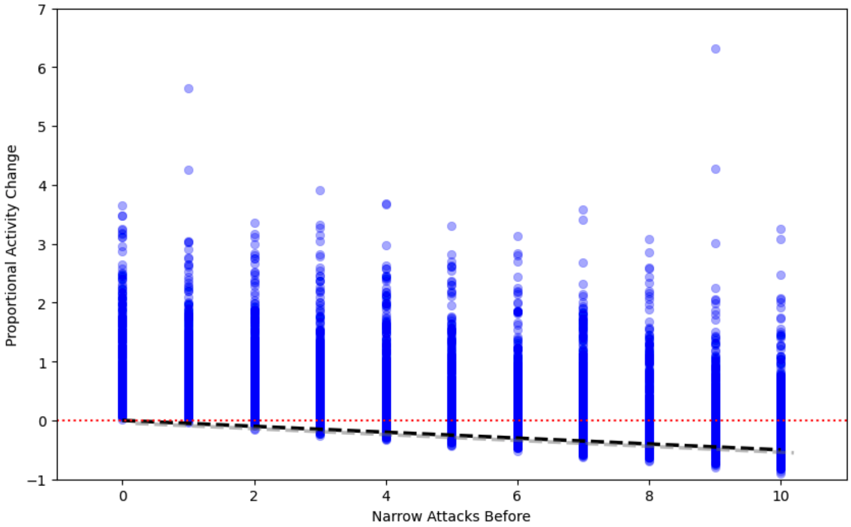

Figure 3. Proportional change in activity as a function of attacks.

5.3. Uncertainty Estimation

To quantify the uncertainties in our observations, we employed one-sample *t*-tests to estimate the mean activity changes across different attack frequencies, acknowledging limitations in data availability for higher numbers of attacks. This approach allowed us to construct confidence intervals and assess statistical significance, which we visualized through bar plots annotated with *p*-values. Despite broader confidence intervals for rarer, higher-frequency attacks, our analysis suggested a statistically significant decrease in activity starting from certain attack thresholds. Additionally, we performed an ANOVA test to investigate the overall trend across attack frequencies, further supported by post hoc analyses, revealing statistically significant differences. Figure 4 presents a summary of the *t*-test results for narrow attacks: the columns include the number of attacks, estimated mean change, confidence interval (low, high), and *p*-value. The significance thresholds crossed at three and four attacks, with broader confidence intervals for rarer higher-frequency attacks due to sample size limitations.

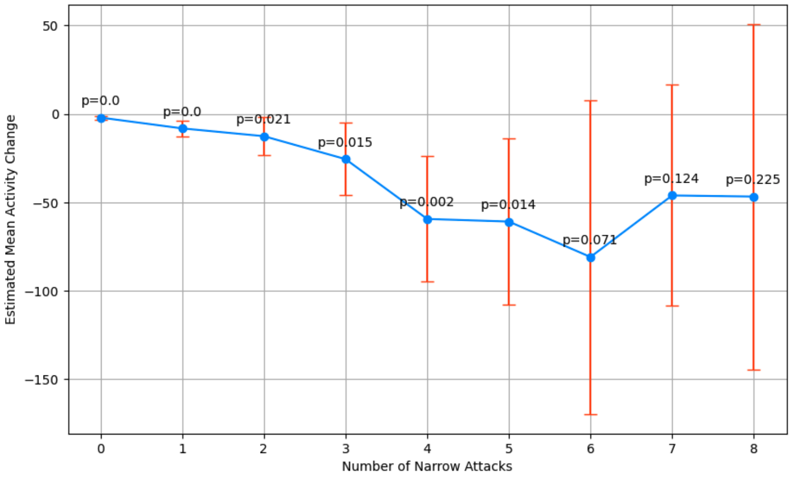

Figure 4. Summary of *t*-test results for narrow attacks.

Figure 5 presents the results of the ANOVA and post hoc analysis. It highlights strong evidence of a non-random correlation between the number of attacks and the change in activity, with significant post hoc differences highlighted by Tukey's Honest Significance Test.

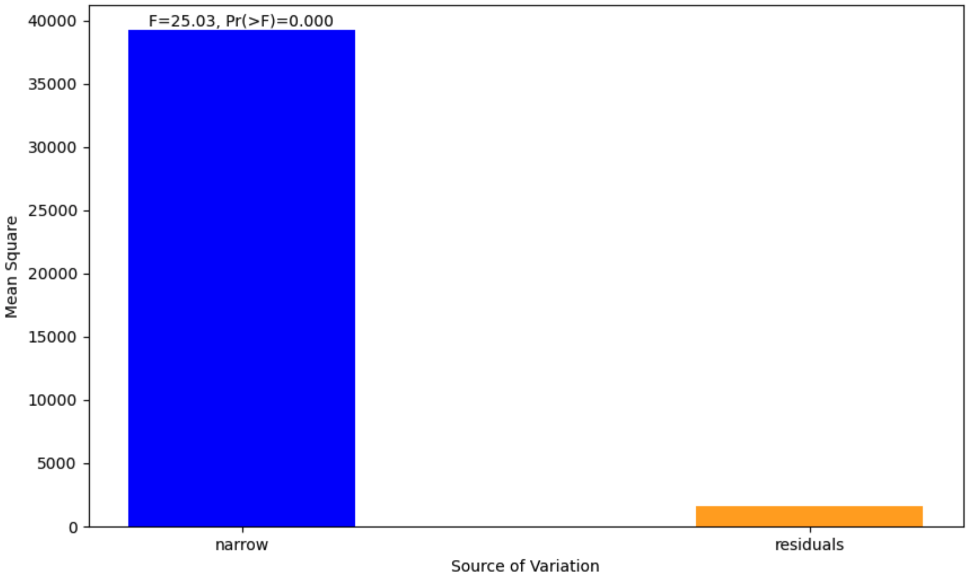

Figure 5. ANOVA and post hoc analysis.

5.4. Bayesian Estimation

Adopting a Bayesian framework, we estimated the posterior distributions of the mean activity changes for different attack frequencies, utilizing skeptical prior distributions (see Figure 6. The figure shows the wide skeptical prior for a Bayesian analysis, which is a normal distribution with a mean of 0 and a standard deviation of 50. The plot includes the probability density function for the normal distribution over a range of expected activity changes. This method provided a nuanced view of how prior beliefs could be updated in light of new evidence, resulting in a consensus across different priors about the general trend of decreases in activity with more frequent attacks.

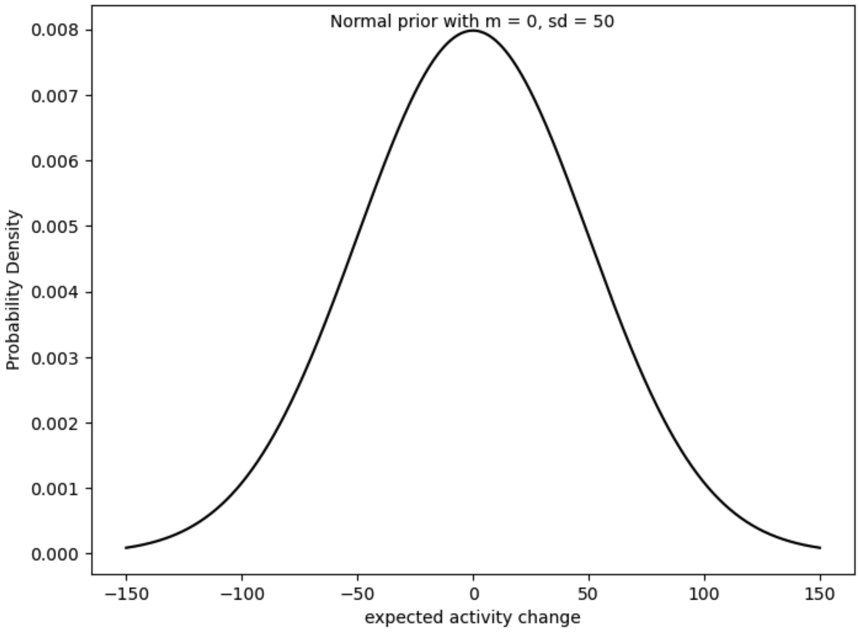

Figure 6. Wide skeptical prior.

Figure 7 presents the posterior distributions for activity changes across attack frequencies. It depicts the density plot of the means of the posterior distributions for 0–9 narrow attacks using a wide prior, illustrating how data update prior beliefs about the impact of attacks on activity. The posterior means shift toward more negative values as the number of attacks increases, indicating a consensus on the negative impact of attacks, irrespective of the prior used.

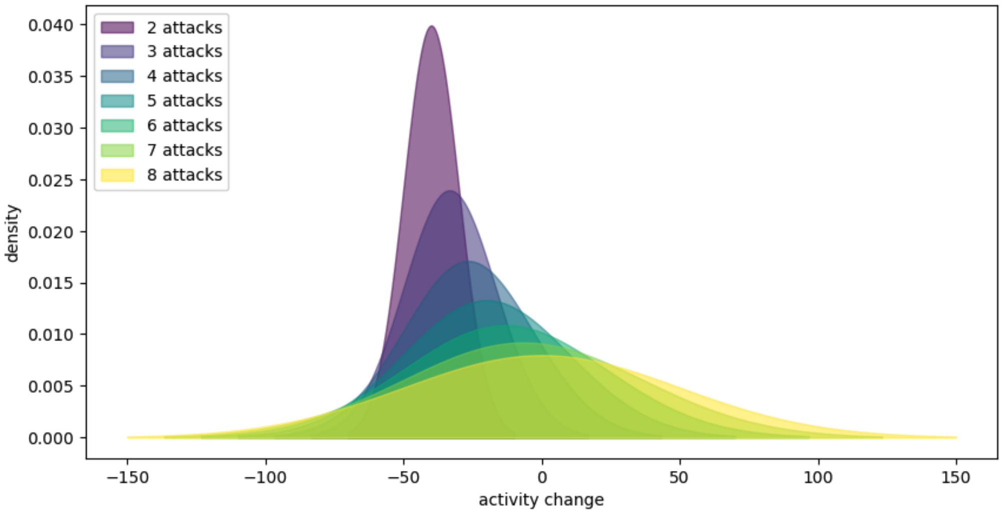

Figure 7. Posterior distributions for activity changes across attack frequencies.

5.5. Model-Theoretic Analysis

Extending our analysis, we explored additional predictors and potential confounders through regression analysis. Our models aimed to predict post-attack activity, taking into account factors such as the nature of the attacks and previous activity levels. We experimented with various distributions and modeling approaches, including zero-inflated and hurdle models, to best capture the underlying data structure. Our findings highlight the significance of certain predictors while accounting for baseline activity levels, offering insights into the relative impact of different factors on post-attack activity changes. Figure 8 presents the model fit statistics for various distributions. It depicts the comparison of Poisson, quasi-Poisson, and negative binomial models, highlighting the best-fitting model based on goodness-of-fit tests and the Akaike Information Criterion (AIC). We adopted a *lambda* parameter of approximately 35.69 for the Poisson distribution. Vertical lines have been added to highlight the issues of zero inflation (green line at zero count) and overspread (purple line at maximum count). The histogram of observed data is shown in sky blue, with the red dashed line indicating the poor fit of the Poisson prediction. Figure 8 clearly demonstrates the aforementioned problems with zero counts and the distribution's inability to capture higher values in the data, reflecting the poor performance of the Poisson model for this dataset.

Figure 9 shows the predicted vs. actual activity post-attack. It describes the predictive accuracy of the chosen model, comparing the predicted activity levels against the actual observations. The selected model accurately captured the distribution of post-attack activity, with adjustments for over-dispersion and zero inflation.

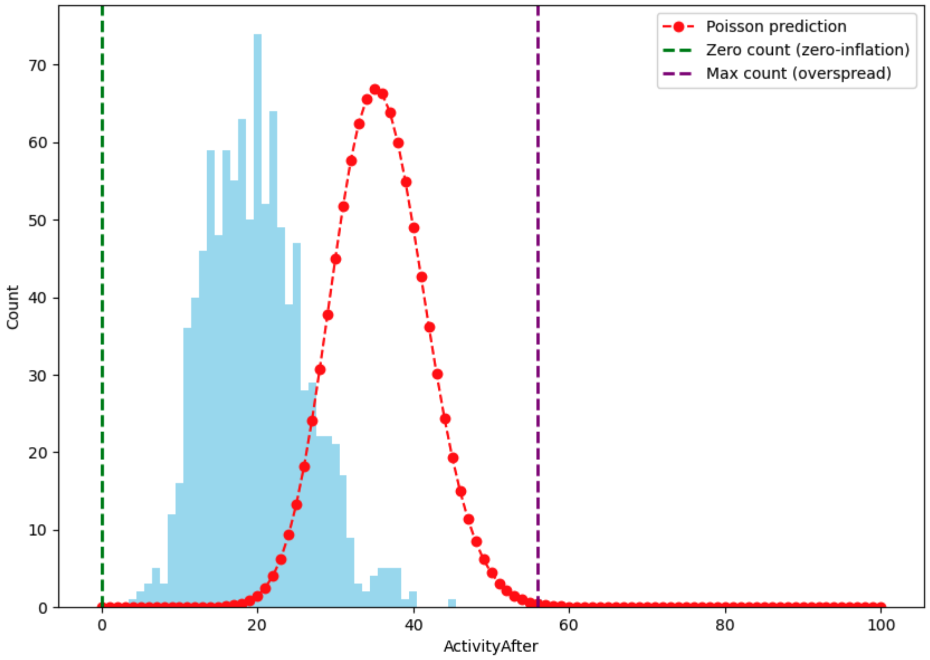

Figure 8. Activity after fitting with the best Poisson model predictions, with x restricted to 0–100, showing the poor performance of the best-fitting Poisson distribution.

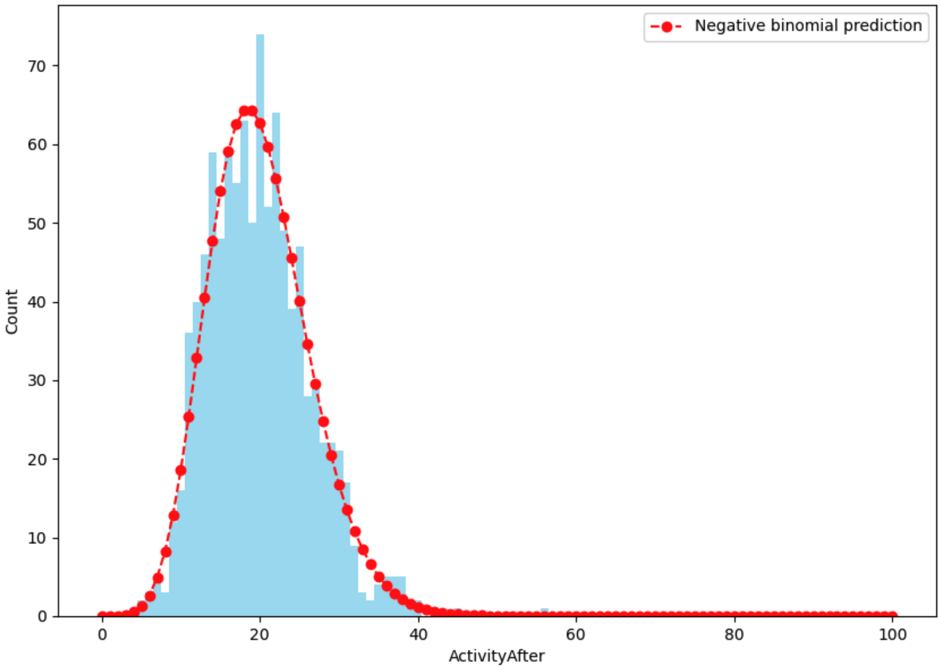

Figure 9. Predicted vs. actual activity post-attack, showing the better performance of the best-fitting negative binomial distribution.

5.6. Addressing Concerns and Limitations

We acknowledged the observational nature of our study and discussed potential biases and limitations, such as self-selection and regression to the mean. To mitigate these, we considered alternative approaches and statistical adjustments, emphasizing the importance of cautious interpretation and the need for future, more controlled studies to validate our findings. Figure 10 presents the exponentiated coefficients from the full-hurdle negative binomial model. Each variable is shown as a horizontal bar, with the length representing the odds ratio.

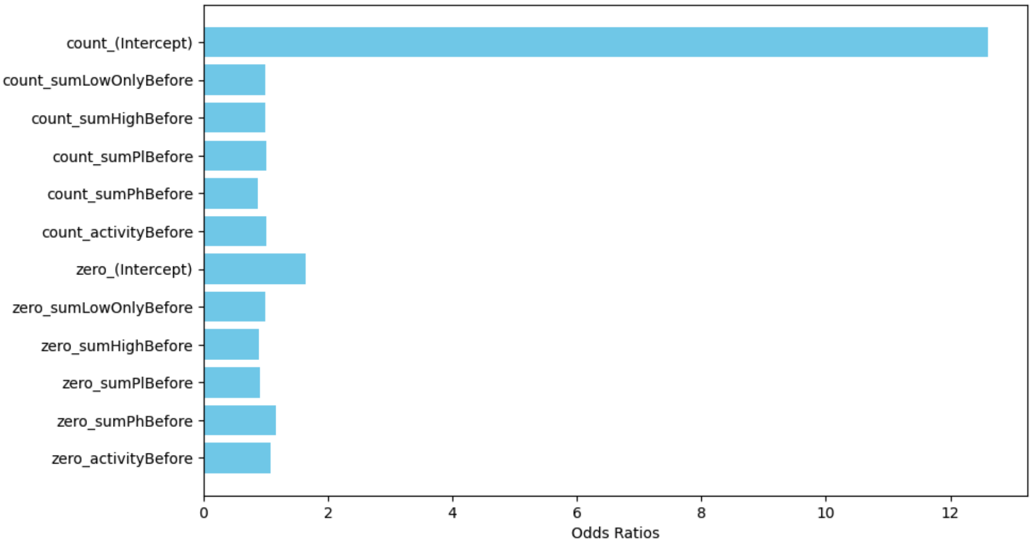

Figure 10. Exponentiated coefficients from the full-hurdle negative binomial model.

Figure 11 depicts the expected activity based on the number of personal attacks received, with the other variables fixed at their mean levels. The number of attacks is on the x-axis, and the expected activity is on the y-axis, with each point connected by a line.

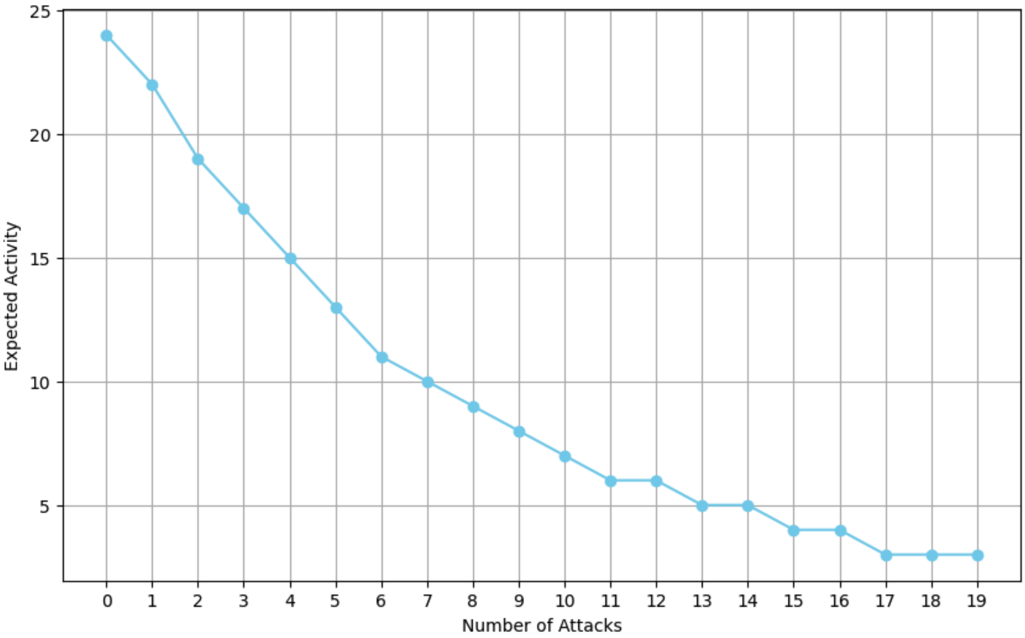

Figure 11. Model's expected activity counts, with other variables fixed at their mean levels, based on personal attacks received.

Figure 12 shows the test results of the ANCOVA, with a bar plot for the F values and a line plot for the generalized eta-squared (ges) values. The F value for 'activityBefore' is significantly higher than that for 'fhigh', as indicated by the large blue bar, and both effects are statistically significant ($p < 0.05$). The ges values, which measure effect size, are shown on a second y-axis, with red markers and a line connecting them, indicating the proportion of variance accounted for by each effect.

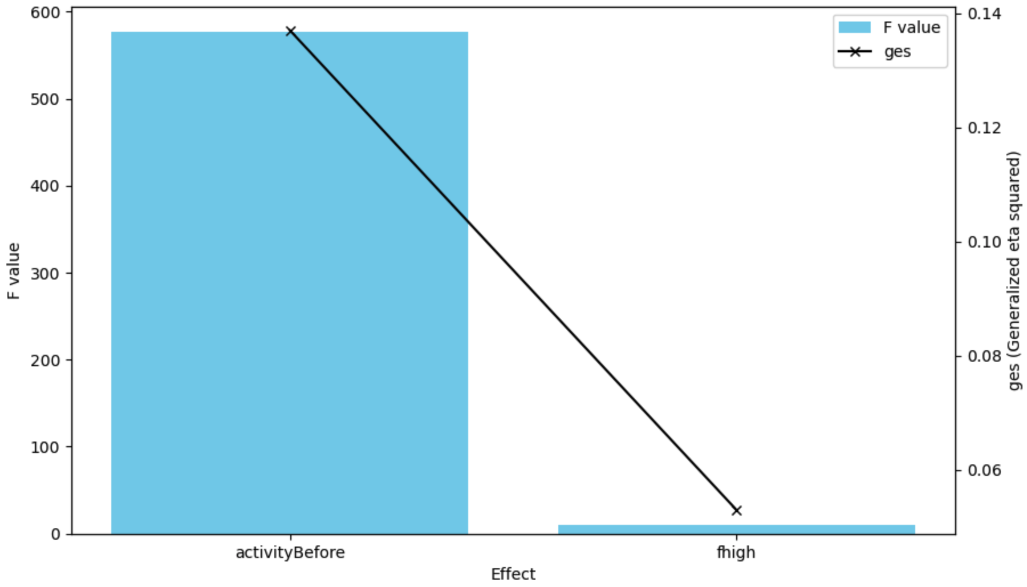

Figure 12. Activity before and the number of received narrow attacks vs. ANCOVA test of activity difference.

6. Results

In this section, we explore the nuanced relationship between personal attacks on social media platforms and subsequent changes in user activity. The examination is anchored by the Bradford Hill criteria for causal inference, which guide our interpretation of the findings. The analysis commenced with the identification and evaluation of potential confounding factors, acknowledging that while some variables, such as user activity level prior to this study and the type of platform used, were carefully considered, others, like demographic characteristics, were only partially accounted for, and historical incidences of harassment were not included due to data limitations. This comprehensive approach is detailed in Table 2, which lists these factors and their consideration within the study's framework.

Table 2. Overview of confounding factors considered in this study.

Factor	Consideration in Study
User activity level prior to study	Yes
Demographic characteristics	Partially
Platform type	Yes
Historical incidences of harassment	No

Further, our empirical data reveal a clear dose–response relationship between the number of personal attacks received and a decrease in user activity, as illustrated in Figure 13. This finding substantiates the hypothesis that users tend to reduce their engagement on social media platforms following incidents of personal attacks. Figure 13 illustrates the dose–response relationship between the number of personal attacks received on social media and the resultant drop in user activity. This graph supports the discussion about the measurable impact of online harassment on user behavior.

The quantification of this effect is summarized in Table 3, where we can observe that even a small number of attacks (1–2) can lead to a 5% drop in activity, with more significant decreases (15–25%) as the number of attacks increases. However, for higher frequencies of attacks (>10), our study faces a limitation due to insufficient data, suggesting an area for future research to further elucidate this trend.

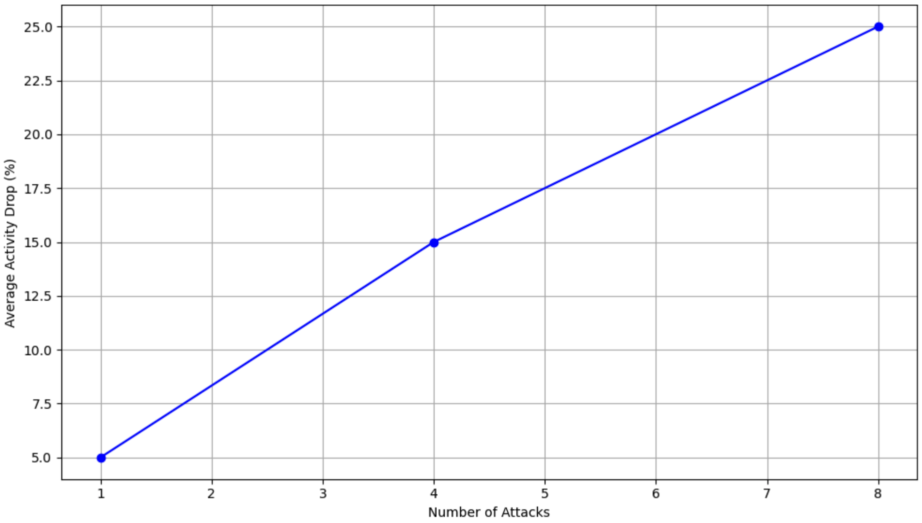

Figure 13. Dose–response relationship between the number of attacks and changes in activity.

Table 3. Summary of changes in user activity post-attack.

Number of Attacks	Average Drop in Activity (%)
1–2	5
3–5	15
6–10	25
>10	Data insufficient

Mechanistic evidence supports the observed behavior, positing that individuals retreat from online engagement to avoid further confrontations or harm. This aligns with broader findings in the literature on cyberbullying and social media fatigue, where such negative experiences contribute to psychological distress and a diminished desire to participate online. Our discussion incorporates parallel evidence from studies like those in [61,62], which not only corroborate the negative impact of online harassment but also propose interventions that might mitigate these effects. Specifically in [62] highlighted the global nature of online hate and the efficacy of targeted cluster-banning strategies, an insight that complements our findings on the individual level and suggests potential policy implications for social media platforms.

The role of moderators, as discussed in [63] emerges as a critical factor in managing online communities and safeguarding against the proliferation of harmful content. The development and implementation of advanced tools for moderation, coupled with user-driven reporting mechanisms, represent promising avenues for enhancing online safety and user experience. These discussions are visualized in Figure 14, which, while a placeholder due to specific data constraints, aims to encapsulate the comparative effectiveness of various intervention strategies across different studies. Figure 14 conceptualizes the perceived effectiveness of various intervention strategies against online harassment. It encapsulates a comparative analysis of intervention strategies, such as bystander interventions, cluster banning, and enhancements to moderation tools, suggesting promising avenues for combating online harassment and fostering safer digital environments. Both Figures 13 and 14 complement the "Results and Discussion" section by providing a visual summary of this study's key findings and the effectiveness of the proposed interventions, reinforcing the importance of strategic approaches to mitigate the adverse effects of personal attacks on social media platforms.

In conclusion, our study leverages the Bradford Hill criteria to methodically dissect the causal relationship between personal attacks and reduced user activity on social media, emphasizing the multifaceted nature of online interactions and the potential for strategic interventions to foster healthier digital environments. The findings underscore the importance of continued research, particularly in areas where data gaps persist, to better understand and counteract the dynamics of online harassment and its impacts on user behavior. Tables 2 and 3 provide a structured presentation of the critical data from our study, facilitating clear and concise communication of our findings in the written report.

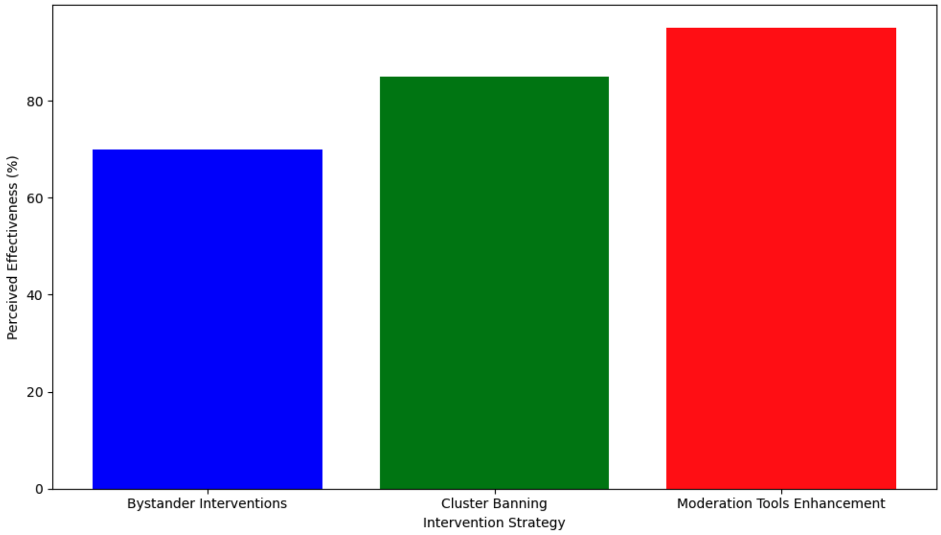

Figure 14. Effectiveness of intervention strategies for mitigating online harassment.

7. Discussion

This research contributes significantly to our understanding of the emotional impact of online personal attacks on user activity within social media ecosystems. Our findings reveal a distinct pattern of reduced user engagement correlating with the frequency and severity of personal attacks. Specifically, this study demonstrates a statistically significant decrease in user activity: a 5% reduction following 1–2 attacks, 15% after 3–5 attacks, and 25% when 6–10 attacks occur. These numbers not only provide a clear quantitative measure of the impact of online harassment but also underline the importance of a supportive digital environment for maintaining active user engagement. A central implication of our findings is the critical need for sophisticated AI-driven moderation tools that can swiftly and accurately detect instances of harassment. Such tools would enable social media platforms to take proactive measures in curbing the spread of harmful content, thereby reducing its psychological impact on users and maintaining a healthy digital discourse. This study's findings offer a roadmap for enhancing content moderation on social media by implementing AI-driven detection systems, establishing real-time intervention protocols, and developing predictive models to prevent harassment. The insights from the observed dose–response relationship between personal attacks and user engagement can inform the creation of responsive reporting tools and the design of user support mechanisms. Additionally, these results can guide the formulation of targeted content policies, facilitate educational initiatives to promote respectful interaction, and enable the tailoring of moderation strategies to the unique dynamics of different social media platforms. Together, these approaches aim to create a more secure and supportive online environment, fostering improved user experience and community health. However, this study is not without its limitations. One major constraint lies in its observational nature, which, while extensive, cannot definitively establish causality. While a dose–response relationship was observed, the potential for other unmeasured variables to influence user activity cannot be completely ruled out. For instance, demographic characteristics were only partially considered, and historical instances of harassment were not included, which could potentially provide further insights into the patterns of user engagement. Furthermore, the datasets used were limited to a select period and platforms, primarily Reddit, which may not be entirely representative of the diverse and dynamic landscape of social media. This limitation suggests the need for more comprehensive data collection that encompasses a wider array of platforms and temporal spans. Our results also highlight the nuanced nature of online interactions, where not all negative encounters have the same impact on users. For instance, the effects of trolling may differ from direct personal attacks, and users may vary in their responses based on their prior experiences and resilience. Future research should, therefore, aim to disentangle these complex dynamics and explore the individual differences in users' reactions to online harassment. Finally, this study's findings underscore the potential for using machine learning not only for detection but also for preemptive interventions. The possibility of predicting which users or content may lead to harassment before it occurs opens the door to preventive measures that social media platforms can implement. In conclusion, while this study's findings add a valuable dimension to our understanding of social media dynamics, the highlighted limitations pave the way for further research. Future work should aim to incorporate more diverse data, consider additional confounding factors, and utilize a combination of observational and experimental designs to validate and expand upon the current study's insights. This will enhance our ability to develop targeted interventions and create safer, more supportive online communities.

8. Conclusions

Our study conclusively links the occurrence of personal attacks on social media to tangible declines in user engagement, marking a significant contribution to the discourse on online behavior and platform moderation. The empirical evidence underpins the need for proactive content moderation and the deployment of artificial intelligence to safeguard users, a need that becomes more pressing with each quantified decrease in activity levels post-attack. The specificity of our findings—highlighting a 5%, 15%, and 25% drop in user activity following incrementing tiers of attack frequency—provides a clear metric for platforms to tailor their intervention strategies. Moreover, this study's intricate data analysis, encompassing psycholinguistic features and model-theoretic approaches, presents a comprehensive model for identifying the characteristics and predictors of harmful social media interactions. Our work encourages ongoing innovation in the development of AI tools and the formulation of nuanced content policies that respond to the complexities of online harassment. Future research must build upon the groundwork laid here, refining detection algorithms and expanding the scope of study to incorporate diverse social media landscapes. In conclusion, this research underscores the profound effect of personal attacks on social engagement and catalyzes a call to action for social media platforms to implement robust, data-informed strategies to mitigate these negative interactions.

Author Contributions: A.L. and A.A. (Abdullah Albanyan): ideas. A.L.: design, implementation, and paper writing. H.L.: Paper revision. R.L. and E.K.: comments and evaluation. A.A. (Abdullah Albanyan) and A.A. (Abdulrahman Alabduljabbar): management and project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by Prince Sattam bin Abdulaziz University, project number (PSAU/2023/01/25781).

Institutional Review Board Statement: No humans or animals were involved in this study.

Data Availability Statement: Data can be provided on reasonable request from the corresponding author.

Acknowledgments: The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2023/01/25781).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Valkenburg, P.M.; Peter, J.; Schouten, A.P. Friend networking sites and their relationship to adolescents' well-being and social self-esteem. *CyberPsychol. Behav.* **2006**, *9*, 584–590. [CrossRef] [PubMed]
2. Beran, T.; Li, Q. Cyber-harassment: A study of a new method for an old behavior. *J. Educ. Comput. Res.* **2005**, *32*, 265.

3. Campbell, D.A.; Lambright, K.T.; Wells, C.J. Looking for friends, fans, and followers? Social media use in public and nonprofit human services. *Public Adm. Rev.* **2014**, *74*, 655–663. [CrossRef]
4. Hinduja, S.; Patchin, J.W. Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant Behav.* **2008**, *29*, 129–156. [CrossRef]
5. Kelly, Y.; Zilanawala, A.; Booker, C.; Sacker, A. Social media use and adolescent mental health: Findings from the UK Millennium Cohort Study. *EClinicalMedicine* **2018**, *6*, 59–68. [CrossRef] [PubMed]
6. Kircaburun, K. Self-Esteem, Daily Internet Use and Social Media Addiction as Predictors of Depression among Turkish Adolescents. *J. Educ. Pract.* **2016**, *7*, 64–72.
7. Kitazawa, M.; Yoshimura, M.; Murata, M.; Sato-Fujimoto, Y.; Hitokoto, H.; Mimura, M.; Tsubota, K.; Kishimoto, T. Associations between problematic Internet use and psychiatric symptoms among university students in Japan. *Psychiatry Clin. Neurosci.* **2018**, *72*, 531–539. [CrossRef] [PubMed]
8. Pyżalski, J.; Poleszak, W. 3.4. Peer violence and cyberbullying prevention programmes. In *Prevention in School: Current Situation and Future Prospects for Prevention in Poland*; You Have a Chance Foundation: Lublin, Poland, 2019; pp. 186–90.
9. Culpepper, M. Exploring the relationships of social media usage and symptoms of anxiety and depression in adolescents. 2020.
10. Malik, A.; Dhir, A.; Kaur, P.; Johri, A. Correlates of social media fatigue and academic performance decrement: A large cross-sectional study. *Inf. Technol. People* **2020**, *34*, 557–580. [CrossRef]
11. Afful, B.; Akrong, R. WhatsApp and academic performance among undergraduate students in Ghana: Evidence from the University of Cape Coast. *J. Educ. Bus.* **2020**, *95*, 288–296. [CrossRef]
12. Alkhalaf, A.M.; Tekian, A.; Park, Y.S. The impact of WhatsApp use on academic achievement among Saudi medical students. *Med. Teach.* **2018**, *40*, S10–S14. [CrossRef]
13. Yeboah, A.K.; Smith, P. elationships between minority students online learning experiences and academic performance. *Online Learn.* **2016**, *20*.
14. Saiphoo, A.N.; Halevi, L.D.; Vahedi, Z. Social networking site use and self-esteem: A meta-analytic review. *Personal. Individ. Differ.* **2020**, *153*, 109639. [CrossRef]
15. Pudipeddi, J.S.; Akoglu, L.; Tong, H. User churn in focused question answering sites: Characterizations and prediction. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Republic of Korea, 7–11 April 2014; pp. 469–474.
16. Kumar, S.; Hamilton, W.L.; Leskovec, J.; Jurafsky, D. Community interaction and conflict on the web. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 933–943.
17. John, A.; Glendenning, A.C.; Marchant, A.; Montgomery, P.; Stewart, A.; Wood, S.; Lloyd, K.; Hawton, K.; et al. Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review. *J. Med. Internet Res.* **2018**, *20*, e9044. [CrossRef] [PubMed]
18. Díaz, Á.; Hecht-Felella, L. *Double Standards in Social Media Content Moderation*; Brennan Center for Justice at New York University School of Law: New York, NY, USA, 2021. Available online: https://www.brennancenter.org/our-work/research-reports/double-standards-socialmedia-content-moderation (accessed on 15 March 2024).
19. Cresci, S. A decade of social bot detection. *Commun. ACM* **2020**, *63*, 72–83. [CrossRef]
20. Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; Flammini, A. The rise of social bots. *Commun. ACM* **2016**, *59*, 96–104. [CrossRef]
21. Varol, O.; Ferrara, E.; Davis, C.; Menczer, F.; Flammini, A. Online human-bot interactions: Detection, estimation, and characterization. In Proceedings of the International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017; Volume 11, pp. 280–289.
22. Shao, C.; Ciampaglia, G.L.; Varol, O.; Yang, K.C.; Flammini, A.; Menczer, F. The spread of low-credibility content by social bots. *Nat. Commun.* **2018**, *9*, 1–9. [CrossRef]
23. Stella, M.; Ferrara, E.; De Domenico, M. Bots increase exposure to negative and inflammatory content in online social systems. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 12435–12440. [CrossRef]
24. Luceri, L.; Deb, A.; Giordano, S.; Ferrara, E. Evolution of bot and human behavior during elections. *First Monday* **2019**, *24*, 9. [CrossRef]
25. Al Marouf, A.; Hasan, M.K.; Mahmud, H. Identifying neuroticism from user generated content of social media based on psycholinguistic cues. In Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, Bangladesh, 7–9 February 2019; pp. 1–5.
26. Uyheng, J.; Moffitt, J.; Carley, K.M. The language and targets of online trolling: A psycholinguistic approach for social cybersecurity. *Inf. Process. Manag.* **2022**, *59*, 103012. [CrossRef]
27. Zannettou, S.; Sirivianos, M.; Blackburn, J.; Kourtellis, N. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *J. Data Inf. Qual.* **2019**, *11*, 1–37. [CrossRef]
28. Ong, J.C.; Cabañes, J.V. When disinformation studies meets production studies: Social identities and moral justifications in the political trolling industry. *Int. J. Commun.* **2019**, *13*, 5771–5790.
29. Gorwa, R.; Guilbeault, D. Unpacking the social media bot: A typology to guide research and policy. *Policy Internet* **2020**, *12*, 225–248. [CrossRef]
30. Alsmadi, I.; O'Brien, M.J. How many bots in Russian troll tweets? *Inf. Process. Manag.* **2020**, *57*, 102303. [CrossRef]
31. Ferrara, E.; Chang, H.; Chen, E.; Muric, G.; Patel, J. Characterizing social media manipulation in the 2020 US presidential election. *First Monday* **2020**, *25*, 11–12.

32. Carley, K.M. Social cybersecurity: An emerging science. *Comput. Math. Organ. Theory* **2020**, *26*, 365–381. [CrossRef] [PubMed]
33. Addawood, A.; Balakumar, P.; Diesner, J. Categorization and Comparison of Influential Twitter Users and Sources Referenced in Tweets for Two Health-Related Topics. In Proceedings of the Information in Contemporary Society: 14th International Conference, iConference 2019, Washington, DC, USA, 31 March–3 April 2019; pp. 639–646.
34. Zubiaga, A.; Aker, A.; Bontcheva, K.; Liakata, M.; Procter, R. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.* **2018**, *51*, 1–36. [CrossRef]
35. de Rosa, A.S.; Bocci, E.; Bonito, M.; Salvati, M. Twitter as social media arena for polarised social representations about the (im) migration: The controversial discourse in the Italian and international political frame. *Migr. Stud.* **2021**, *9*, 1167–1194. [CrossRef]
36. Rakshitha, K.; Ramalingam, H.M.; Pavithra, M.; Advi, H.D.; Hegde, M. Sentimental analysis of Indian regional languages on social media. *Glob. Transitions Proc.* **2021**, *2*, 414–420. [CrossRef]
37. Morshed, S.A.; Khan, S.S.; Tanvir, R.B.; Nur, S. Impact of COVID-19 pandemic on ride-hailing services based on large-scale Twitter data analysis. *J. Urban Manag.* **2021**, *10*, 155–165. [CrossRef]
38. Shelke, N.; Chaudhury, S.; Chakrabarti, S.; Bangare, S.L.; Yogapriya, G.; Pandey, P. An efficient way of text-based emotion analysis from social media using LRA-DNN. *Neurosci. Inform.* **2022**, *2*, 100048. [CrossRef]
39. Louati, A. A hybridization of deep learning techniques to predict and control traffic disturbances. *Artif. Intell. Rev.* **2020**, *53*, 5675–5704. [CrossRef]
40. Louati, A. Cloud-assisted collaborative estimation for next-generation automobile sensing. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106883. [CrossRef]
41. Louati, A.; Louati, H.; Kariri, E.; Neifar, W.; Hassan, M.K.; Khairi, M.H.; Farahat, M.A.; El-Hoseny, H.M. Sustainable Smart Cities through Multi-Agent Reinforcement Learning-Based Cooperative Autonomous Vehicles. *Sustainability* **2024**, *16*, 1779. [CrossRef]
42. Alghamdi, H.M. nveiling Sentiments: A Comprehensive Analysis of Arabic Hajj-Related Tweets from 2017–2022 Utilizing Advanced AI Models. *Big Data Cogn. Comput.* **2024**, *8*. [CrossRef]
43. Najar, D.; Mesfar, S. Opinion mining and sentiment analysis for Arabic on-line texts: application on the political domain. *Int. J. Speech Technol.* **2017**, *20*, 575–585. [CrossRef]
44. Sghaier, M.A.; Zrigui, M. Sentiment analysis for Arabic e-commerce websites. In Proceedings of the 2016 International Conference on Engineering & MIS (ICEMIS), Agadir, Morocco, 22–24 September 2016; pp. 1–7.
45. Mourad, A.; Darwish, K. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Atlanta, GA, USA, 14 June 2013; pp. 55–64.
46. Alwakid, G.; Osman, T.; Haj, M.E.; Alanazi, S.; Humayun, M.; Sama, N.U. MULDASA: Multifactor Lexical Sentiment Analysis of Social-Media Content in Nonstandard Arabic Social Media. *Appl. Sci.* **2022**, *12*, 3806. [CrossRef]
47. Tartir, S.; Abdul-Nabi, I. Semantic sentiment analysis in Arabic social media. *J. King Saud Univ. Comput. Inf. Sci.* **2017**, *29*, 229–233. [CrossRef]
48. Zhou, M.; Mou, H. Tracking public opinion about online education over COVID-19 in China. *Educ. Technol. Res. Dev.* **2022**, *70*, 1083–1104. [CrossRef]
49. Toçoğlu, M.A.; Onan, A. Sentiment analysis on students' evaluation of higher educational institutions. In Proceedings of the International Conference on Intelligent and Fuzzy Systems, Istanbul, Turkey, 21–23 July 2020; pp. 1693–1700.
50. Nikolić, N.; Grljević, O.; Kovačević, A. Aspect-based sentiment analysis of reviews in the domain of higher education. *Electron. Libr.* **2020**, *38*, 44–64. [CrossRef]
51. Mohiudddin, K.; Rasool, A.M.; Mohd, M.S.; Mohammad, R.H. Skill-Centered Assessment in an Academic Course: A Formative Approach to Evaluate Student Performance and Make Continuous Quality Improvements in Pedagogy. *Int. J. Emerg. Technol. Learn.* **2019**, *14*, 92. [CrossRef]
52. Dsouza, D.D.; Deepika, D.P.N.; Machado, E.J.; Adesh, N. Sentimental analysis of student feedback using machine learning techniques. *Int. J. Recent Technol. Eng.* **2019**, *8*, 986–991.
53. Webb, M.E.; Fluck, A.; Magenheim, J.; Malyn-Smith, J.; Waters, J.; Deschênes, M.; Zagami, J. Machine learning for human learners: Opportunities, issues, tensions and threats. *Educ. Technol. Res. Dev.* **2021**, *69*, 2109–2130. [CrossRef]
54. Singh, N.K.; Tomar, D.S.; Sangaiah, A.K. Sentiment analysis: A review and comparative analysis over social media. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *11*, 97–117. [CrossRef]
55. Alrehili, A.; Albalawi, K. Sentiment analysis of customer reviews using ensemble method. In Proceedings of the 2019 International Conference on Computer and Information Sciences (ICCIS), Aljouf, Saudi Arabia, 3–4 April 2019; pp. 1–6.
56. Al-Smadi, M.; Qawasmeh, O.; Al-Ayyoub, M.; Jararweh, Y.; Gupta, B. Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. *J. Comput. Sci.* **2018**, *27*, 386–393. [CrossRef]
57. Al-Horaibi, L.; Khan, M.B. Sentiment analysis of Arabic tweets using text mining techniques. In Proceedings of the First International Workshop on Pattern Recognition, Tokyo, Japan, 11–13 May 2016; Volume 10011, pp. 288–292.
58. Louati, A.; Louati, H.; Kariri, E.; Alaskar, F.; Alotaibi, A. Sentiment Analysis of Arabic Course Reviews of a Saudi University Using Support Vector Machine. *Appl. Sci.* **2023**, *13*, 12539. [CrossRef]
59. Roussel, R.; Rosenzweig, J. Space radiation simulation using blowout plasma wakes at the SAMURAI Lab. *Nucl. Instrum. Methods Phys. Res. Sect. A* **2017**, *865*, 71–74. [CrossRef]

60. Manwani, P.; Ancelin, H.; Majernik, N.; Williams, O.; Sakai, Y.; Fukasawa, A.; Naranjo, B. Simulations for the Space Plasma Experiments at the SAMURAI Lab, 2022.
61. Countering hate speech on Facebook: The case of the Roma minority in Slovakia. *Social Science Computer Review* **2020**, *38*, 128–146. [CrossRef]
62. Seering, Joseph and Wang, Tony and Yoon, Jina and Kaufman, Geoff Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature* **2019**, *573*, 261–265. [CrossRef]
63. Seering, Joseph and Wang, Tony and Yoon, Jina and Kaufman, Geoff Moderator engagement and community development in the age of algorithms. *New Media & Society* **2019**, *21*, 71–74.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.