Information Research - Vol. 30 No. iConf (2025)

Analyzing the language of rejection: a study of user flagging responses to hate speech on Reddit

Sharon Lisseth Perez, Xiaoying Song, and Lingzi Hong

DOI: https://doi.org/10.47989/ir30iConf47191

Abstract

Introduction. Online hate speech poses significant threats to individuals and society, exacerbating psychological harm, discrimination, and potential real-world violence. While automated detection models are available, their inability to recognize subtle variations of hate speech, particularly implicit forms, emphasizes the need for supplementary methods.

Method. This study investigates the potential of user-written flagging messages in enhancing hate speech detection, focusing on the characteristics and identification of flagging messages. We created a dataset of flagging messages and the comments they respond to, employing transformer-based models (BERT, RoBERTa, ALBERT, DistilBERT, and XLNet) for classification.

Analysis. Linguistic analysis using SEANCE and Named Entity Recognition was conducted to reveal unique characteristics of flagging messages.

Results. Our findings show that BERT and DistilBERT models achieved the highest accuracy in classifying flagging messages, with distinct linguistic patterns emerging in flagging content.

Conclusion. This research contributes to the development of more nuanced hate speech detection methods by leveraging user-generated flagging content. These findings have implications for improving automated content moderation systems and supporting more inclusive online environments. Future work will focus on the effectiveness of flagging messages in identifying implicit hate speech across diverse cultural contexts.

Introduction

Online hate speech poses significant threats to individuals and communities, exacerbating psychological harm, discrimination, and potential real-world violence. Extensive research has been dedicated to identifying and addressing this issue, with various studies utilizing automated methods, including natural language processing techniques, to detect and categorize hate speech on social media platforms (Salminen et al., 2020, Yu et al., 2022).

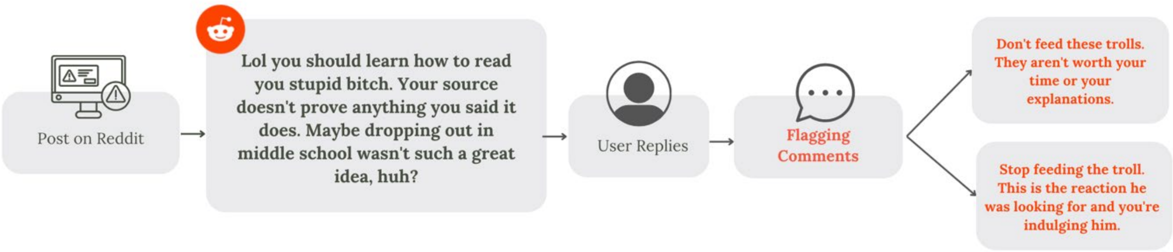
However, current detection methods primarily focus on explicit forms of hate, often overlooking more subtle manifestations and evolving abusive language. This limitation highlights the need for more nuanced approaches that can capture the complex dynamics of online hate. One potential source of insight that has been under-explored is user-written flagging messages.

Flagging is widely employed by users across social media platforms to report offensive content, as shown in Figure 1, and it provides moderators with a rhetorical defence for content removal decisions (Crawford and Gillespie, 2016). In an environment where control and transparency are limited, flags are crucial in giving users a voice (Zhang et al., 2023). Chandrasekharan et al. (2018) found that online posts flagged by regular users and later reviewed by moderators were important in determining the best ways to intervene.

This study aims to identify flagging messages, laying a foundation for the empirical investigation of flagging messages in enhancing hate speech detection, particularly for subtle forms of abuse or microaggressions (De la Peña Sarracén and Rosso, 2023; MacAvaney et al., 2019). Specifically, we address the following research questions:

1. What are the linguistic features that distinguish flagging from non-flagging messages?
2. What NLP models are most effective in accurately detecting flagged content?
3. To what extent can flagging messages aid in identifying implicit hate, and what insights can they provide into this form of hate expression?

We aim to provide insights that can inform the development of more comprehensive and nuanced hate speech detection models, ultimately contributing to safer and more inclusive online environments.



Lol you should learn how to read you stupid bitch. Your source doesn't prove anything you said it does. Maybe dropping out in middle school wasn't such a great idea, huh?

Don't feed these trolls. They aren't worth your time or your explanations.

Stop feeding the troll. This is the reaction he was looking for and you're indulging him.

Figure 1. Online users post flagging messages to warn others. These comments indicate the comment they respond to is harmful speech

Related work

Implicit hate speech detection

Detecting implicit forms of hate speech poses significant challenges due to linguistic nuances and contextual dependencies. ElSherief et al. (2021) presented a comprehensive benchmark dataset aimed at understanding and detecting implicit hate speech, providing a foundation for evaluating detection algorithms. Ghosh et al. (2023) proposed CoSyn, a context-synergized hyperbolic network that considers both personal and dialogue context in conversation trees. Their approach highlights the importance of contextual information in this task. Jafari et al. (2023) investigated the influence of fine-grained emotions on implicit hate speech classification. Their research employed single task learning and multi-task learning models and revealed that integrating emotional features improved hate speech detection. Recently, Ahn et al. (2024) introduced a SharedCon model that leverages shared semantic information to enhance the detection of implicit hate speech.

These works collectively advance the understanding of implicit hate speech. However, as hate language evolves, the curated implicit hate speech datasets could slowly become obsolete, affecting detection accuracy. We propose a detection method that is anchored in flagging messages, which are less likely to change due to the lack of motivation for alteration among users.

Method

Data collection and processing

We started by compiling a small selection of comments from Reddit discussion threads. We collected a total of 1,050 pairs of hate speech comments and their corresponding replies from 39 subreddits that have been identified as hateful (Vidgen et al., 2021). Examples of these subreddits include r/TwoXChromosomes, r/TrueReddit, and r/ChangeMyView. We used Reddit's API to collect the comments while ensuring that we followed the platform's terms of service.

Our research team manually examined and labeled each reply message. We utilized a binary (0,1) coding system to categorize the messages based on their content, purpose, tone, effectiveness, and rejection in addressing hate speech. Inter-rater agreement was calculated to ensure the reliability of the labeling process. The obtained results are as follows: Cohen's kappa: 0.81 and Krippendorff's alpha: 0.82. A Kappa coefficient between 0.61 and 0.80 indicates substantial agreement (Viera et al., 2005), demonstrating the reliability of our labels.

Data analysis

We utilize the SEANCE tool (Sentiment Analysis and Cognition Engine) to perform in-depth sentiment, social cognition, and social order analysis on textual data. Its comprehensive set of indices and component scores allows for a nuanced and insightful analysis of language use, emotions, and social dynamics (Crossley et al., 2017). In addition to the SEANCE analysis, we conducted a Named Entity Recognition (NER) using Spacy (Naseer et al., 2021).

We use the Wilcoxon Rank Sum Test to conduct statistical tests on SEANCE and NER features between the flagging and non-flagging messages. Then Bonferroni correction is conducted to identify the most significant features.

Classification experiments

We utilize several state-of-the-art transformer-based models, including BERT (Kenton and Toutanova, 2019), RoBERTa (Liu, 2019), ALBERT (Lan, 2019), DistilBERT (Sanh, 2019), and XLNet (Yang, 2019) to classify a comment to be flagging or not.

The labeled data was split into training and testing sets using an 8:2 ratio, ensuring class balance by stratifying based on the 'label.' As the labels of 'flagging' and 'non-flagging' are unbalanced, we experiment with sampling and data augmentation methods to investigate whether these techniques will benefit the prediction performance. We first use undersampling, which involves reducing the number of records in the majority class. Random undersampling has been empirically shown to be one of the most effective resampling methods. Few of the more sophisticated undersampling methods have outperformed random undersampling in empirical studies (Liu, 2004).

We then expand our dataset to evaluate whether implementing NLTK tools such as punkt, averaged_perceptron_tagger, and wordnet would enhance our models' performance. Punkt is a tokenizer that uses an unsupervised algorithm to divide a text into a list of sentences, taking into consideration abbreviation words, collocations, and words that start sentences (Natural Language Toolkit, n.d.).

The models were trained using the PyTorch framework and the Hugging Face Transformers library. Model evaluation metrics included accuracy, precision, recall, and F1-score. We generated classification reports to provide detailed insights into model performance across different classes.

Results

In this section, we present detailed findings of our experiments, encompassing both the insights gained from linguistic analysis techniques and the performance of various transformer-based models.

Linguistic analysis

This analysis offers insights into the linguistic characteristics that differentiate flagging from non-flagging messages, providing a deeper understanding of their nature.

Our SEANCE analysis revealed several significant linguistic differences between flagging and non-flagging comments, with a particularly strong emphasis on negative emotion words in flagging content. Statistically significant differences ($p < 0.001$) were observed in various measures. For example, anger (Lexicon) was considerably more prevalent in flagging comments, with a mean of 0.073 compared to 0.029 in non-flagging comments. Fear and disgust showed the largest difference, with means of 0.437 and 0.210 for flagging and non-flagging comments respectively. General negative emotions (Negative_EmoLex) were also more common in flagging comments (mean 0.104) compared to non-flagging ones (mean 0.052).

These findings indicate that flagging comments are characterized by a significantly higher emotional intensity, specifically negative emotions like anger, fear, and disgust. This is consistent with prior research on hate speech and offensive language, which frequently involve strong expressions of negative emotions (Yu et al., 2022). We also observed differences in other linguistic features. For example, the second-person pronouns (You_GI) were higher in flagging comments, potentially indicating more direct confrontational language. Table 1 shows these findings.

Variable Category	***Variable***	***Mean (Flagging)***	***Mean (Non-flagging)***	***p-Value***	***Bonferroni Correction***
Negative Emotion words	Anger (Lexicon)	0.073	0.029	$2.47E^{-41}$	$6.66E^{-39}$
	Fear and Disgust	0.437	0.210	$6.30E^{-39}$	$1.70E^{-36}$
	Negative Emotions (Negative_EmoLex)	0.104	0.052	$6.81E^{-38}$	$1.83E^{-35}$
	Fear (Lexicon)	0.072	0.029	$7.45E^{-36}$	$2.01E^{-33}$
	Negative Adjectives	0.733	0.364	$6.08E^{-10}$	$1.64E^{-07}$
	Negative Words	0.063	0.043	$8.39E^{-10}$	$2.27E^{-07}$
	Negative Sentiment (VADER)	0.144	0.105	$3.71E^{-09}$	$1.00E^{-06}$
	Hatred (Lexicon)	0.002	0.000	$1.93E^{-05}$	0.005
	Disgust (Lexicon)	0.027	0.020	$1.01E^{-04}$	0.027
	Negativity (Lexicon)	0.074	0.062	$3.19E^{-04}$	0.086
Reference	Second-Person Pronouns (You_GI)	0.041	0.031	$1.78E^{-06}$	0.000
Quality and quantity	Numerical Mentions	0.004	0.009	$9.64E^{-05}$	0.026
	Frequency References	0.010	0.005	$3.06E^{-04}$	0.083
Other affect	Sentiment Compound (VADER)	-0.113	-0.005	$1.02E^{-08}$	$2.75E^{-06}$
Emotion words	Positive Nouns	-0.177	-0.088	$1.90E^{-07}$	0.000

Table 1. The most significant linguistic differences are grouped by variable category. Only variables that pass the Bonferroni correction are included

NER analysis

This analysis assists in uncovering contextual elements and specific entity types that might indicate potentially problematic content. The results show a statistically significant difference between PERSON and CARDINAL. The presence of person names (PERSON) is strongly associated with non-flagging comments. This suggests that comments mentioning specific individuals are less likely to be a flagging message. The higher presence of cardinal numbers (CARDINAL) in non-flagging comments indicates that more factual or quantitative content is less likely to be the flagging message. These results are displayed in Table 2.

GPE (Geo-Political Entities) tends to appear more often in non-flagging comments. The trend with GPEs, while not significant after correction, hints that mentions of geographical or political entities might be more common in non-flagging comments.

Variable	***Mean (Flagging)***	***Mean (Non-flagging)***	***p-value***	***Bonferroni correction***
PERSON	0.039	0.095	0.000	0.002
CARDINAL	0.043	0.080	0.001	0.011
GPE	0.038	0.064	0.015	0.212

Table 2. Entities with more prevalent differences. Only variables that pass the Bonferroni correction are included

Model performance

We examine the effectiveness of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet models. We conduct experiments using various input configurations, including replies to hate speech alone (Child), and a combination of hate speech (Parent) and replies to it. Table 3 presents these results.

BERT and DistilBERT show the highest overall performance, both achieving an accuracy of 0.81 and balanced F1 scores across labels. BERT, using only Child, demonstrates the highest F1 score (0.82) for flagging comments.

There is a slight variation in how models handle different contexts (only Child vs. Parent and Child). Most models demonstrate balanced performance between labels of flagging and non-flagging. Including Parent as a context varies across models: BERT and XLNet perform best with only Child, suggesting they may be more sensitive to noise in Parent comments. In contrast, RoBERTa, ALBERT, and DistilBERT show strong performance with both Parent and Child, indicating better context integration capabilities.

DistilBERT's high performance is particularly noteworthy, as it is a compressed model designed for efficiency. The similar performance of BERT and DistilBERT suggests that DistilBERT might be more suitable in resource-constrained environments.

Model	***Input***	**Non-Flagging**			**Flagging**			**Weighted Average**			***Accuracy***
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	
BERT	Child only	0.84	0.76	0.80	0.78	0.85	**0.82**	0.81	0.81	**0.81**	**0.81**
RoBERTa	Parent and child	0.77	0.81	0.79	0.80	0.75	0.77	0.78	0.78	0.78	0.78
ALBERT	Parent and child	0.82	0.77	0.79	0.78	0.83	0.80	0.80	0.80	0.80	0.80
DistilBERT	Parent and child	0.82	0.80	0.81	0.80	0.82	0.81	0.81	0.81	**0.81**	**0.81**
XLNet	Child only	0.80	0.77	0.79	0.78	0.81	0.79	0.79	0.79	0.79	0.79

Table 3. Summary of best-performing model results, including the impact of different input configurations (Child only or Parent and child) on the model's performance

Conclusion

Our study provides insights into the linguistic characteristics of flagging comments. Our study combining Linguistic and NER analysis, and transformer-based models, reveals several key findings:

The Linguistic analysis demonstrated a strong association between negative emotional language and flagging comments. The NER analysis revealed that certain types of named entities, particularly PERSON and CARDINAL, are differentially associated with flagging and non-flagging comments. These findings could be instrumental in identifying flagging messages, which could be used to develop more sophisticated content moderation tools. The evaluation of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet models demonstrated their effectiveness in detecting flagging messages. The context should be carefully considered when designing flagging systems.

Limitations and future directions

We have identified flagging messages; how these could enhance the detection of implicit or varied forms of hate speech will be further investigated in the next step, as the work is still in progress. While our study gives us valuable insights, it's important to recognize its limitations. The binary classification of comments as flagging or not may simplify the complex nature of online discourse. Future research will explore more granular categorizations, considering the nuances of different types and intensities of negative emotions and their relationship to various forms of online harm like implicit hate. The trending large language models may have better performance in identifying flagging messages, which will be experimented in the future.

This study contributes to the ongoing efforts to create safer and healthier online discussions and communities. Yet, the scope of our research can be expanded and address a broader range of online harm indicators.

Acknowledgements

The authors gratefully acknowledge the financial support from the Institute of Museum and Library Services under Grant LG-256661-OLS-24 and LG-256666-OLS-24. Sharon Lisseth Perez also acknowledges the support of the Fulbright Program for funding her PhD studies at the University of North Texas. The authors would like to thank the collaborators who contributed to this research.

About the authors

Sharon Lisseth Perez is a PhD student in the College of Information at the University of North Texas, Texas, USA. Her research interests are on natural language processing, artificial intelligence, and their applications in education and social justice. She can be reached via email at sharonperez@my.unt.edu

Xiaoying Song is a PhD student in College of Information, University of North Texas, Texas, USA. Her research focuses on evaluating generative counter-speech and addressing health misinformation. She can be reached via email at xiaoyingsong@my.unt.edu

Lingzi Hong is an Assistant Professor of Data Science in the College of Information at the University of North Texas. She holds a Ph.D. in Information Science from the University of Maryland, College Park. Her research focuses on human-centered computing and artificial intelligence. She can be reached at lingzi.hong@unt.edu

References

Ahn, H., Kim, Y., Kim, J., & Han, Y. S. (2024, August). SharedCon: Implicit hate speech detection using shared semantics. In *Findings of the Association for Computational Linguistics ACL 2024* (pp. 10444-10455). https://aclanthology.org/2024.findings-acl.622/

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., ... & Gilbert, E. (2018). The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1-25. https://doi.org/10.1145/3274301

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3), 410-428. https://doi.org/10.1177/1461444814543163

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, 49, 803-821. https://doi.org/10.3758/s13428-016-0743-z

De la Peña Sarracén, G. L., & Rosso, P. (2023). Systematic keyword and bias analyses in hate speech detection. *Information Processing & Management*, 60(5), 103433. https://doi.org/10.1016/j.ipm.2023.103433

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021, November). Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 345-363). https://aclanthology.org/2021.emnlp-main.29/

Ghosh, S., Suri, M., Chiniya, P., Tyagi, U., Kumar, S., & Manocha, D. (2023). CoSyn: Detecting Implicit Hate Speech in Online Conversations Using a Context Synergized Hyperbolic Network. *arXiv preprint arXiv*:2303.03387. https://doi.org/10.48550/arXiv.2303.03387

Jafari, A. R., Li, G., Rajapaksha, P., Farahbakhsh, R., & Crespi, N. (2023). Fine-grained emotions influence on implicit hate speech detection. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3318863

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (Vol. 1, p. 2). https://doi.org/10.48550/arXiv.1810.04805

Lan, Z. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942. https://doi.org/10.48550/arXiv.1909.11942

Liu, A. Y. C. (2004). The effect of oversampling and undersampling on classifying imbalanced text datasets. http://dx.doi.org/10.26153/tsw/12300

Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv*:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS one*, 14(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Naseer, S., Ghafoor, M. M., bin Khalid Alvi, S., Kiran, A., Rahmand, S. U., Murtazae, G., & Murtaza, G. (2021). Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance. *Pakistan Journal of Multidisciplinary Research*, 2(2), 293-308. https://pjmr.org/pjmr/article/view/150

Natural Language Toolkit. (n.d.). *NLTK 3.0 documentation*. https://www.nltk.org/index.html

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10, 1-34. https://doi.org/10.1186/s13673-019-0205-6

Sanh, V. (2019). DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv preprint arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108

Vidgen, B., Nguyen, D., Margetts, H., Rossini, P., & Tromble, R. (2021). Introducing CAD: the contextual abuse dataset. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 2289-2303) https://doi.org/10.18653/v1/2021.naacl-main.182

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5), 360-363. https://api.semanticscholar.org/CorpusID:38150955

Yang, Z. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint arXiv:1906.08237. https://doi.org/10.48550/arXiv.1906.08237

Yu, X., Blanco, E., & Hong, L. (2022). Hate Speech and Counter Speech Detection: Conversational Context Does Matter. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 5918-5930). https://doi.org/10.48550/arXiv.2206.06423

Zhang, A. Q., Montague, K., & Jhaver, S. (2023). Cleaning Up the Streets: Understanding Motivations, Mental Models, and Concerns of Users Flagging Social Media Posts. *arXiv preprint arXiv*:2309.06688. https://doi.org/10.48550/arXiv.2309.06688