EM-26@LT-EDI 2025: Detecting Racial Hoaxes in Code-Mixed Social Media Data

Tewodros Achamaleh[1], Fatima Uroosa[1], Nida Hafeez[1], Abiola T. O.[1]
Mikiyas Mebiratu[2], Sara Getachew[2], Grigori Sidorov[1,], Rolando Quintero[1]
[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico
[2]Wolkite University, Department of Information Technology, Wolkite, Ethiopia
[2]Jimma University, Institute of Technology, Jimma, Ethiopia
Corr. email: sidorov@cic.ipn.mx

Abstract

Social media platforms and user-generated content, such as tweets, comments, and blog posts often contain offensive language, including racial hate speech, personal attacks, and sexual harassment. Detecting such inappropriate language is essential to ensure user safety and to prevent the spread of hateful behavior and online aggression. Approaches base on conventional machine learning and deep learning have shown robust results for high-resource languages like English and find it hard to deal with code-mixed text, which is common in bilingual communication. We participated in the shared task "LT-EDI@LDK 2025" organized by DravidianLangTech, applying the BERT-base multilingual cased model and achieving an F1 score of 0.63. These results demonstrate how our model effectively processes and interprets the unique linguistic features of code-mixed content. The source code is available on GitHub.[1]

1 Introduction

In recent years, the rise of smartphones and the affordable internet has made social networks a central part of everyday life (Aichner et al., 2021). Platforms like Twitter, Instagram, and Facebook have allowed users to communicate and share ideas widely. Although these platforms offer improved communications and networking benefits, they also pose risks, especially with regard to privacy, misinformation, and hate speech. Such issues have been significantly affected during crises (Chhabra and Vishwakarma, 2023).

Racial hoaxes, a form of information disorder (Hatta, 2020), involve spreading false or misleading content that targets individuals based on ethnicity or nationality (Corazza et al., 2020; Biradar et al., 2024). Although the relationship between disinformation and hate speech is complex, the two often overlap and can destabilize public opinion. Researchers have categorized information disorders into three main types: misinformation, disinformation, and malinformation all of which disrupt trust and communication (Fallis, 2015; Frau-Meigs, 2019; Tsang, 2024). These narratives can intensify hostility, polarize groups, and fuel stereotypes or threats against communities base on race, religion, or other attributes (Joshi et al., 2020). Such content may also cause lasting psychological harm, including anxiety and depression. During emergencies, it can mislead the public and result in harmful decisions (Talat and Hovy, 2016). The COVID-19 pandemic revealed the scale of racial hoaxes, where particular groups were unjustly blamed, often resulting in discrimination and violence (Pérez et al., 2023).

The prevalence of code-mixed language adds further complexity, as current NLP systems struggle to handle informal and linguistically diverse expressions. This underscores the need for improved hate speech detection techniques in multilingual contexts. Researchers aim to address these issues by participating in shared tasks and contributing to safer, more inclusive digital spaces.

2 Related work

Many researchers carried out important early work on the detection of hate- or fake-generated content. Disinformation, in particular, relies on identity-base controversies and adversarial narratives. It uses a variety of rhetorical techniques and forms of knowing, including truths, half-truths, and judgments laden with value, to exploit and amplifies identity-driven controversies (Diaz Ruiz and Nilsson, 2023). Because it can use the truth or portions of the truth to misinform, the concept of disinformation extends much beyond what is true or not (Brisola and Doyle, 2019). The deliberate creation of deceptive or inaccurate content has led to the

[1]https://github.com/teddymas95/Detecting-Racial-Hoaxes.git

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 146–151

emergence of what is commonly referred to as fake news (Lazer et al., 2018; Achamaleh et al., 2025b, 2024; Eyob et al., 2024).

The term fake news typically describes fully fabricated stories (Imhoff and Lamberty, 2020) that are knowingly false yet often presented with enough realism to appear credible. While defining fake news precisely remains challenging, scholars generally agree that it involves the intentional misleading of large audiences by individuals or groups outside of traditional media, using sensationalist and seemingly trustworthy formats crafted to deceive (Finneman and Thomas, 2018). What makes fake news particularly damaging is its ability to imitate and exploit legitimate news sources, drawing on their credibility while simultaneously eroding it. One key feature that distinguishes fake news from conventional journalism is its emotional appeal it tends to use surprising and emotionally charged content to increase user engagement, sharing, and memory retention (Scardigno et al., 2023). Hence, it is necessary to address hateful and fake narratives by considering both their targets and severity (Zhou and Zafarani, 2020).

(Yin et al., 2009) made the first step for using supervised learning methods to identify harassment in online platforms. Researchers used a support-vector machine (SVM) to group social media posts base on local contextual and sentiment cues (Yin et al., 2009; Si et al., 2019). Researchers investigated the effectiveness of character n-grams, word n-grams, and skip-grams in detecting hoax speech in social media content. Their system, trained on an English dataset with three class labels, achieved a classification accuracy of 78% (Malmasi and Zampieri, 2017). Researchers introduced a convolutional neural network (CNN) model, which was a system to detect offensive tweets in Hindi-English code-switched language (Zampieri et al., 2019b). Researchers curated Hindi-English code-mixed tweets to aid the development of methods to identify hate speech (Bohra et al., 2018; Mathur et al., 2018; Kapil and Ekbal, 2024). The dataset consists exclusively of Twitter data written in the Roman script. The authors used character and word n-grams, punctuation, lexicon, and negation features for their classification method, using either SVM or random forest classifiers. Any combination of all features with SVM achieved the best performance accuracy, up to 71.7% to detect hate speech (Ullah et al., 2024; Nagpal et al.).

Although the automatic detection of offensive language has been extensively studied in resource-rich languages such as English (Waseem and Hovy, 2016; Davidson et al., 2017; de Gibert et al., 2018; Zampieri et al., 2019a), research in the resource-poor Hindi language remains extremely limited. As a contribution to the initiative on online hate and societal harmony, this work advances the current state of research by addressing the detection of offensive content in code-mixed text using a BERT-base multilingual cased model, demonstrating its effectiveness in the context of the "LT-EDI@LDK 2025" task organized by DravidianLangTech. Related efforts by the CIC-NLP team has also shown the applicability of multilingual transformer models for detecting AI-generated and deceptive content across languages, including English and Dravidian code-mixed text (Abiola et al., 2025a,b; Achamaleh et al., 2025a). These approaches collectively emphasize the growing potential of transformer-based models in handling complex, multilingual, and socially sensitive NLP tasks.

3 Methodology

This study employed the BERT-base-multilingual-cased model from the Hugging Face Transformers library. It was chosen for its strong contextual understanding across 100+ languages, crucial for handling code-mixed social media text. The model was fine-tuned for binary classification to distinguish racial hoaxes from non-hoax content. While it is well known that pre-trained transformers outperform shallow models, we included CNN and Transformer-FFNN as baselines to quantify performance differences and highlight trade-offs in low-resource scenarios. PyTorch was used for training and evaluation with GPU support.

3.1 Task Overview

The aim of this shared task is to identify instances of racial hoaxes in Hindi-English code-mixed social media content, tackling one type of misinformation that unwisely ascribes to an individual or group behavior against the law or ethical standards (Chakravarthi et al., 2025). Such hoaxes usually rely on deceitful stories, stereotypes, and groundless accusations against social, ethnic, or marginalized groups that lead to the spread of false information and instability in society. The complexity in code-mixed content stems from mixing several languages with colloquial structures and unconventional spellings, which contribute to a lot of

difficulty in analyzing the content.

3.2 Dataset Description

The dataset provided by the LT-EDI@LDK 2025 Shared Task, known as the HoaxMixPlus dataset, consists of "5,105" code-mixed Hindi-English YouTube comments annotated for detecting racial hoaxes, a harmful form of misinformation that falsely associates individuals or communities with crimes or incidents. The training set and the validation set comprise a total of "3,060" and "1,021" samples, respectively. Both sets demonstrate a class imbalance that includes approximately 75.8% labels as Racial Hoax (Label 0) and 24.2% labels as Not Racial Hoax (Label 1). Although this imbalance reflects real-world events, it creates difficulties in training and evaluating models. We rely on the BERT-base-multilingual-cased model to deal with the code-mixed nature of the data. Its ability to multilingually and subword tokenise makes it appropriate for handling noisy social media text in the form of mixed Hindi-English text. The stable distribution of split labels enables reliable evaluation. This task addresses the urgent need to fight racially motivated misinformation in resource-constrained environments and drives the emergence of strong models for code-mixed social media contexts.

3.3 System Setup

The model was fine-tuned for three epochs with a batch size 16 and a learning rate of 2e-5, following standard practices for transformer models on moderately sized datasets. These hyperparameters were chosen to balance training efficiency and generalization, though further tuning could improve performance. The AdamW optimizer was employed to update model weights effectively, incorporating weight decay to reduce overfitting. Training was conducted on GPU hardware when available to accelerate computation. During each step, the model received tokenized input batches, computed the loss against ground-truth labels, and updated its parameters via backpropagation. Performance was evaluated on a validation set after each epoch using accuracy, precision, recall, F1-score, and a confusion matrix to identify misclassification patterns. The checkpoint with the highest validation accuracy was retained for inference. A custom prediction pipeline was also implemented to classify unseen text and return the predicted label and its confidence score.

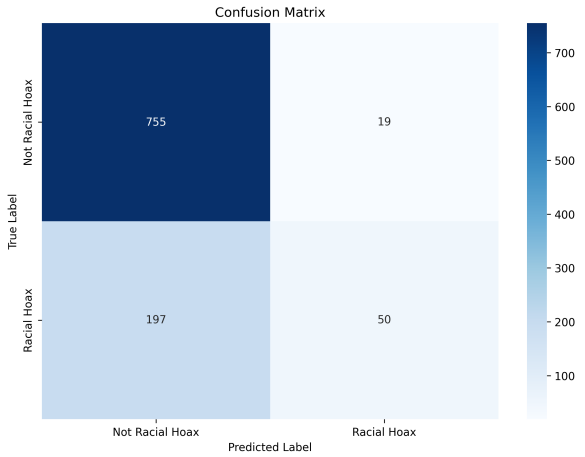

Figure 1: Confusion Matrix

4 Results

We compared several models for detecting hoax speech in code-mixed content. As shown in Table 1, our BERT-base model demonstrated the best overall performance on the development set, achieving an accuracy of 0.8000 and a macro-averaged F1-score of **0.6700**, outperforming XLM-RoBERTa (F1-score 0.5584), CNN (F1-score 0.5071), and Transformer-FFNN (F1-score 0.0863). Although XLM-RoBERTa achieved a slightly higher AUC score of 0.7781, BERT provided a better balance of precision (0.7300) and recall (0.6500), as well as a superior F1-score. On the official test set, our BERT model obtained an F1-score of **0.63**. These results emphasize the strength of multilingual-BERT for processing noisy, code-mixed data and demonstrate its real usefulness for multilingual hoax-speech detection

5 Discussion

Due to the linguistic complexity and social sensitivity involved, detecting racial hoaxes in code-mixed Hindi-English social media posts remains a challenging task. Our results demonstrate that multilingual transformer models, particularly BERT, perform well in this context. BERT achieved an F1-score of 0.6700 and an accuracy of 0.8000, reflecting strong generalization capabilities and effective contextual understanding, even when handling informal and noisy data. XLM-RoBERTa achieved the highest AUC (0.7781), reflecting good class separation, but its lower F1-score shows an imbalance between precision and recall. CNN, though faster and more efficient, lacked the depth to capture the nuanced meaning in racial hoax texts.

Model	Accuracy	Precision	Recall	F1-Score	AUC
mBERT	0.8000	0.7300	0.6500	**0.6700**	0.7720
XLM-RoBERTa	0.7818	0.5465	0.5709	0.5584	0.7781
CNN	0.7281	0.4511	0.5789	0.5071	0.7512
Transformer-FFNN	0.7508	0.3871	0.0486	0.0863	0.5692

Table 1: Model Comparison on the Development Dataset

Similarly, the Transformer-FFNN model underperformed, suggesting that shallow architectures struggle with the ambiguity and language mixing common in such posts. These results highlight the importance of deep contextual modeling for identifying deceptive narratives in multilingual environments. While BERT demonstrated the best overall performance, all models were challenged by code-switching and subtle sarcasm, highlighting the need for more diverse and culturally annotated training data. Table 1 illustrates how model depth and multilingual architecture influence performance.

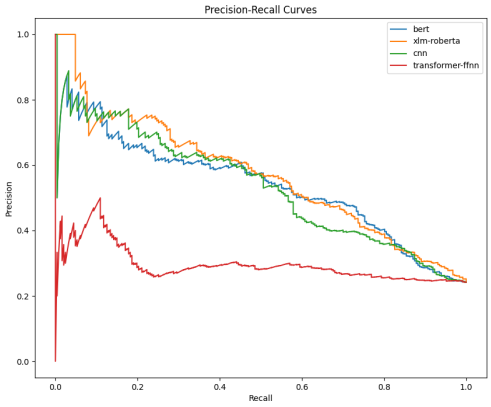

Figure 2: Precision and Recall plot on validation

6 Error Analysis

The confusion matrices in Figures 1 and 2 highlight a repeated pattern of misclassification, particularly for the minority class "Racial Hoax." Out of 247 actual "Racial Hoax" instances, 197 were misclassified as "Not Racial Hoax" indicating a strong bias toward the majority class. In contrast, the model performed well on the "Not Racial Hoax" class, correctly classifying 755 out of 774 cases. This imbalance indicates that the model has difficulty identifying the small linguistic or contextual signals that identify racial hoaxes. The results highlight the need to consider methods such as class balancing, deep semantic understanding, and advanced feature engineering. Increasing the sensitivity of the model about minority class characteristics would enhance the overall classification rate and reduce false negatives in critical categories such as racial hoaxes.

Conclusion

This work investigated the detection of racial hoaxes in Hindi-English code-mixed social media content using deep learning models. BERT outperformed other models in terms of F1-score, further demonstrating its ability to capture the contextual and linguistic nuances of bilingual, informal text. Despite its strong overall performance, the model struggled to correctly classify the minority class labeled as "Racial Hoax" showing a pronounced bias toward predicting the majority class. This indicates the persistent issue of class imbalance and the detection of delicate hints in minority classes. Remediation of this problem through class-aware methods and better representation of the features of the minority class will be the main way forward for future enhancements. Our work highlights the capabilities of multilingual transformers in code-mixed NLP, especially on socially sensitive tasks. Future studies should focus on tuning models with balanced datasets and incorporating richer semantic knowledge to improve the accurate identification of harmful or deceptive online content.

Limitations

While our work using BERT and other transformer-based models produced promising results in identifying racial hoaxes within code-mixed Hindi-English social media data, several limitations were observed. A major concern was the issue of class imbalance, which led the model to misclassify instances of the minority class and adversely affected its accuracy in detecting racial hoaxes. Besides, the data's mixed-code nature, usually involving informal language, transliteration, and non-uniform grammar, required more effort from the models, which were not adapted to such patterns. The ab-

sence of targeted pre-processing or code-mixed language modelling may have led to lower overall performance. The dataset used was relatively small and highly task-specific, limiting the generalizability of the results to broader, real-world scenarios. Furthermore, we have not yet used more sophisticated techniques like ensemble methods, data augmentation, or external knowledge integration, which could only increase the understanding of the model regarding complex, socially charged language.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Oluwatobi Joseph Abiola, Temitope Olasunkanmi Oladepo, Olumide Ebenezer Ojo, Grigori Sidorov, and Olga Kolesnikova. 2025a. Cic-nlp at genai detection task 1: Leveraging distilbert for detecting machine-generated text in english. pages 271–277.

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Olumide Ebenezer Ojo. 2025b. Cic-nlp at genai detection task 1: Advancing multilingual machine-generated text detection. pages 262–270.

Tewodros Achamaleh, Tolulope Olalekan Abiola, Lemlem Eyob Kawo, Mikiyas Mebraihtu, and Grigori Sidorov. 2025a. Cic-nlp@ dravidianlangtech 2025: Detecting ai-generated product reviews in dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 502–507.

Tewodros Achamaleh, Nida Hafeez, Mikiyas Mebraihtu, Fatima Uroosa, and Grigori Sidorov. 2025b. Cic-nlp@ dravidianlangtech 2025: Fake news detection in dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 647–654.

Tewodros Achamaleh, Lemlem Kawo, Ildar Batyrshini, and Grigori Sidorov. 2024. Tewodros@ dravidianlangtech 2024: Hate speech recognition in telugu codemixed text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 96–100.

Thomas Aichner, Matthias Grünfelder, Oswin Maurer, and Deni Jegeni. 2021. Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, behavior, and social networking*, 24(4):215–222.

Shankar Biradar, Kasu Sai Kartheek Reddy, Sunil Saumya, and Md Shad Akhtar. 2024. Proceedings of the 21st international conference on natural language processing (icon): Shared task on decoding fake narratives in spreading hateful stories (faux-hate). In *Proceedings of the 21st International Conference on Natural Language Processing (ICON): Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)*, pages 1–5.

Aashiq Bohra, Deepanway Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41.

Adriana C. Brisola and Ann Doyle. 2019. Critical information literacy as a path to resist "fake news": Understanding disinformation as the root problem. *Open Information Science*, 3(1):274–286.

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Shanu Dhawale, Saranya Rajiakodi, Sajeetha Thavareesan, Subalalitha Chinnaudayar Navaneethakrishnan, and Durairaj Thenmozhi. 2025. Overview of the shared task on detecting racial hoaxes in code-mixed hindi-english social media data. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, 29(3):1203–1230.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Carlos Diaz Ruiz and Tomas Nilsson. 2023. Disinformation and echo chambers: how disinformation circulates on social media through identity-driven controversies. *Journal of public policy & marketing*, 42(1):18–35.

Lemlem Eyob, Tewodros Achamaleh, Muhammad Tayyab, Grigori Sidorov, and Ildar Batyrshin. 2024. Stress recognition in code-mixed social media texts using machine learning. *International Journal of Combinatorial Optimization Problems and Informatics*, 15(1):32.

Don Fallis. 2015. What is disinformation? *Library trends*, 63(3):401–426.

Teri Finneman and Ryan J. Thomas. 2018. A family of falsehoods: Deception, media hoaxes and fake news. *Newspaper Research Journal*, 39(3):350–361.

Divina Frau-Meigs. 2019. Information disorders: Risks and opportunities for digital media and information literacy? *Medijske studije*, 10(19):10–28.

Muhammad Hatta. 2020. The spread of hoaxes and its legal consequences. *International Journal of Psychosocial Rehabilitation*, 24(03):1750–60.

Roland Imhoff and Pia Lamberty. 2020. A bioweapon or a hoax? the link between distinct conspiracy beliefs about the coronavirus disease (covid-19) outbreak and pandemic behavior. *Social Psychological and Personality Science*, 11(8):1110–1118.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Prashant Kapil and Asif Ekbal. 2024. A corpus of Hindi-English code-mixed posts to hate speech detection. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 79–85. NLP Association of India (NLPAI).

David M. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2017)*, pages 467–472, Varna, Bulgaria. INCOMA Ltd.

Puneet Mathur, Ramit Sawhney, Maitreya Ayyar, and Rajiv Ratn Shah. 2018. Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 11–17.

Sargun Nagpal, Sharad Dargan, Harsha Koneru, and Shikhar Rastogi. Innovations in code-mixed hate speech detection: The llm perspective.

Juan Manuel Pérez, Franco M Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, and 1 others. 2023. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, 11:30575–30590.

Rocco Scardigno, Adele Paparella, and Francesca D'Errico. 2023. Faking and conspiring about covid-19: A discursive approach. *The Qualitative Report*, 28:49–68.

Suman Si, Anupam Datta, Soumya Banerjee, and Sudip Kumar Naskar. 2019. Aggression detection on multilingual social media text. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE.

Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Stephanie Jean Tsang. 2024. Misinformation, disinformation, and fake news? proposing a typology framework of false information. *Journalism*, page 14648849241304380.

Fida Ullah, Muhammad Zamir, Muhammad Arif, M. Ahmad, E. Felipe-Riveron, and Alexander Gelbukh. 2024. Fida@dravidianlangtech 2024: A novel approach to hate speech detection using distilbert-base-multilingual-cased. pages 85–90.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

Dawei Yin, Zhiting Xue, Liangjie Hong, Brian D. Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. In *Proceedings of the Content Analysis in the Web 2.0 (CAW2.0)*, pages 1–7.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.