Association for Information Systems
AIS Electronic Library (AISeL)

Digit 2023 Proceedings

Diffusion Interest Group In Information Technology

12-1-2023

A Review of Hate Speech Detection: Challenges and Innovations

Hetiao (Slim) Xie
The University of Queensland, hetiao.xie@uq.edu.au

Morteza Namvar
The University of Queensland, m.namvar@business.uq.edu.au

Marten Risius
The University of Queensland, m.risius@business.uq.edu.au

Follow this and additional works at: https://aisel.aisnet.org/digit2023

Recommended Citation
Xie, Hetiao (Slim); Namvar, Morteza; and Risius, Marten, "A Review of Hate Speech Detection: Challenges and Innovations" (2023). *Digit 2023 Proceedings*. 15.
https://aisel.aisnet.org/digit2023/15

This material is brought to you by the Diffusion Interest Group In Information Technology at AIS Electronic Library (AISeL). It has been accepted for inclusion in Digit 2023 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

A Review of Hate Speech Detection: Challenges and Innovations

Research-in-Progress Paper

Hetiao (Slim) Xie
Business School, The University of Queensland
Brisbane, QLD, Australia
hetiao.xie@uq.edu.au

Morteza Namvar
Business School, The University of Queensland
Brisbane, QLD, Australia
m.namvar@business.uq.edu.au

Marten Risius
Business School, The University of Queensland
Brisbane, QLD, Australia
m.risius@business.uq.edu.au

Abstract

Hate speech on social media platforms has severe impacts on individuals, online communities, and society. Platforms are criticized for shirking their responsibilities to effectively moderate hate speech on their platforms. However, Various challenges, including implicit expressions, complicate the task of detecting hate speech. Consequently, developing and tuning algorithms for improving the automated detection of hate speech has emerged as a crucial research topic. This paper aims to contribute to this rapidly emerging field by outlining how the adoption of natural language processing and *machine learning technologies has helped hate speech detection, delving into the latest mainstream detection techniques and their performance, and offering a comprehensive review of the literature on hate speech detection online including the notable challenges and respective mitigating efforts. This paper proposes the integration of interdisciplinary perspectives into deep learning models to enhance the generalization of models, providing a new agenda for future research.*

Keywords: hate speech detection, natural language processing, social media, text classification, deep learning, machine learning

A Review of Hate Speech Detection: Challenges and Innovations

Research-in-Progress Paper

Introduction

The democratization of internet enables users to freely express their opinions (Kane et al., 2014). The liberated communication has created a convenient channel for instantaneously sharing information and ideas. However, this freed communication has also facilitated the spread of hate speech. Hate speech is typically characterized by language that attacks, degrades, or incites violence based on protected attributes such as ethnicity, religion, gender, sexual orientation, and age (Lee et al., 2018).

Hate speech is essentially harmful targeting of socially vulnerable or minority groups and is banned by law in most countries (Vidgen et al., 2019). Well-known social platforms like Facebook, Twitter, and Reddit have faced criticism due to the widespread occurrence of hate speech on their platforms (Gunarathne et al., 2022). The negative impact of hate speech on social platforms is well documented (Lee & Ram, 2020). For users, it cultivates a violent atmosphere, leading to a loss of trust in fellow community members and, in some cases, prompting users to leave these online communities. For the platforms, the presence of hate speech raises concerns for advertisers who worry about their ads appearing alongside content that uses hate speech, reducing their willingness to advertise on such platforms (Fortuna & Nunes, 2019). Social platforms are making efforts to respond by implementing policies to block posts or comments identified as hate speech. However, the large volume of text makes manual detection nearly impossible. As a result, social platforms and researchers are actively involved in developing automated methods for detecting hate speech.

Automated hate speech detection has become a prominent research topic in recent years. On Google Scholar, approximately 400 papers on review of automated hate speech detection, published between 2017 and 2021, can be found. In contrast, prior to 2010, there were fewer than 10 papers on this topic. (Jahan & Oussalah, 2023). This surge is primarily attributed to the rapid advancements in machine learning (ML), natural language processing (NLP), and Artificial Intelligence (AI) techniques (Benbya et al., 2021). Simultaneously, the mainstream approaches to automated hate speech detection have evolved from using simple statistical methods to identify basic surface linguistic features (such as bag-of-words, dictionary-based methods) to employing deep learning models with more complex architectures, enabling them to learn in-depth information from unstructured textual (Schmidt & Wiegand, 2017). However, automated hate speech detection, as an emerging field, faces numerous unresolved challenges, and even state-of-the-art models exhibit noticeable limitations (e.g. difficult to generalize to new datasets) (Mathew et al., 2019). Many previous literature reviews have focused on the development of automated hate speech detection and identified potential issues (Jahan & Oussalah, 2023; Fortuna & Nunes, 2019; Govers et al., 2023; Schmidt & Wiegand, 2017). We expand previous efforts to summarise the development of automated hate speech detection techniques. This paper also explores innovated perspectives for future research direction by an in-depth discussion of persistent challenges.

Related Works

In previous literature, numerous researchers have reviewed methods for the automatic detection of hate speech and discovered a multitude of approaches. Therefore, in this section, we will primarily focus on the most mainstream and essential approaches for automated hate speech detection.

Rules-based Approaches

The rules-based approach involves determining whether a target text constitutes hate speech by referencing a dictionary (or hate lexicon). As one of the most crucial techniques used in NLP, dictionaries were among the earliest and simplest methods for hate speech detection. Typically, a dictionary is manually crafted and comprises a substantial number of hateful terms (Gitari et al., 2015). The assessment of the hate level in each text often involves a straightforward count of the occurrences of hate terms. This method usually relies

on statistical tools to calculate probabilities and can also be combined with supervised learning classifiers. In general, if Text 1 contains more hate terms than Text 2, Text 1 is considered more likely to be hate speech. This method boasts an acceptable interpretability and is operationally straightforward and cost-effective, making it widely applicable in various business settings. Besides, hate lexicons or dictionaries may vary based on the specific requirements of different scenarios and can be composed of distinct terms or even different languages. For instance, Tulkens et al. (2016) created three dictionaries containing Dutch hate terms, utilized for categorizing text into racial discrimination and non-racial discrimination categories.

Rules-based approaches can be improved in various regards. Firstly, collecting different hate speech vocabularies for various scenarios is expensive and time-consuming. Given the globally pervasive challenge of hate speech, developing and maintaining dictionaries for all languages is extremely resource intensive. Secondly, the performance of Automated Hate Speech highly depends on the quality of the dictionary. Combined with the first drawback, building and maintaining a high-quality dictionary is an error-prone task that lead to capricious model performances. Thirdly, models based on individual words cannot learn semantic information from context (Fortuna & Nunes, 2019). Their overly simplistic structure makes it challenging to adapt to more complex contexts, yet understanding context is crucial for accurately capturing the true intent of speech. Lastly, models built on dictionaries lack universality, for instance, a detection model developed for an anti-LGBT forum may not be suitable for an anti-immigrant forum. In summary, rules-based approaches lack flexibility in different situations and the ability to extract contextual semantics, and they have gradually been replaced by more advanced machine learning and deep learning methods.

Machine Learning Approaches

Machine learning approaches mitigate some of the issues present in rules-based approaches. As a form of automated algorithm, ML focuses on uncovering the information hidden in the data and simulates human problem-solving. In hate speech detection, ML models do not calculate probabilities based on individual words, as in dictionary-based methods. Instead, they often employ NLP technics first such as word frequency or distributional similarity to process text. This way of representing text can better capture contextual information, and it has been proven to effectively enhance the performance of classifiers (Davidson et al., 2017). Here are some mainstream ML algorithms applied in hate speech detection.

Logistic Regression (LR): Logistic regression is a supervised machine learning algorithm used to solve classification problems and is one of the simplest machine learning algorithms. It employs a Sigmoid function (also known as the logistic function) to map input data, treated as vectors, to a probability value between 0 and 1, determining whether the input is hate speech. Multiple logistic regression was applied to 1067 comments on Instagram for hate speech detection, resulting in an average precision of 80.02%, recall of 82%, and accuracy of 87.68% (Br Ginting et al., 2019). However, considering the small size of the dataset and other factors, the excellent performance achieved in a simple task often cannot be replicated in more complex datasets (Ayo et al., 2020).

Support Vector Machine (SVM): As the most widely used ML algorithm in hate speech detection (Mullah & Zainon, 2021), SVM was introduced by Cortes & Vapnik (1995) to address binary classification problems that lacked appropriate statistical tools. SVM maps nonlinear vectors to a high-dimensional feature space and constructs a linear decision boundary to perform classification. SVM has demonstrated relatively stable performance across different language datasets. Florio et al. (2020) employed SVM with TF-IDF (Term Frequency- Inverse Document Frequency) for hate speech detection in an Italian language dataset. The study revealed that linear SVM performs better when there is a significant disparity in language features between the testing set and the training set.

Ensemble Approach: Ensemble is a ML approach that combines multiple estimators to enhance generalization and robustness. Ensemble integrates various estimators with different advantages to improve the overall performance of the model, and aggregating many classifiers has consistently proven to be superior to the best individual classifier (Hosni et al., 2019). Therefore, state-of-the-art ML classifiers are almost all based on the ensemble concept, with widely used methods including random forests and boosting (Sagi & Rokach, 2018). Agarwal & Chowdary (2021) proposed an adaptive ensemble model for automatic hate speech detection by combining various classifiers, including Gradient Boosting Decision Trees, Multi-Layer Perceptron Classifier, and others. The model was tested across different public datasets, demonstrating that the ensemble method can help overcome the issue of user overfitting.

Deep Learning Approaches

Deep learning is a subset of ML in the realm of AI. The rapid advancement of deep learning in recent years has led it to outperform traditional machine learning models in various application domains. In the context of hate speech detection, deep learning's main advantage over traditional machine learning lies in its ability to extract more information from inseparable nonlinear text data. Moreover, deep learning models often perform better as the volume of training data increases. In the following we introduce some of the popular deep learning algorithms used in hate speech detection.

Convolutional Neural Network (CNN): CNN is a deep learning algorithm primarily used for processing and analysing data with grid structures. While this algorithm is commonly employed in computer vision and audio domains, considering that text data shares similar unstructured features, CNN has also been one of the earliest technologies applied to hate speech detection (Poria et al., 2016). CNN uses convolutional layers and pooling layers to extract important features from text, then it performs convolution operations on input data, applying filters to extract high-level feature maps. Subsequently, pooling operations are applied to these feature maps to extract more crucial features. However, in more complex hate speech scenarios, CNNs are required to undergo structured changes to approach optimal performance. Models based on CNN have demonstrated outstanding performance in handling hate speech. Khan et al. (2022) designed a model based on deep convolutional layer with hierarchical attention, has shown significant improvements of over 10% in accuracy, recall, and F-score compared to using ordinary CNN model, and surpasses the performance achieved by ML methods. However, some studies indicate that while CNN can extract important features from sentences, they may lack the ability to capture local features (Zhang et al., 2018). In the case of long texts, this limitation can result in a decrease in model performance due to information loss.

Long Short-Term Memory Networks (LSTM): LSTM is a specialized type of recurrent neural network (RNN) designed to address the drawback of traditional RNNs, which tend to lose information from the early part of a sequence when processing long texts. Traditional RNNs perform poorly on lengthy textual data because, during backpropagation, the weights associated with initial words continually diminish. LSTM tackles this issue by introducing memory units and three gating mechanisms—forget gate, input gate, and output gate—to control which information should be retained and which should be discarded. Consequently, a well-trained LSTM network can better capture long-term dependencies within text sequences (Bisht et al., 2020). Before Transformer became widely used, LSTM emerged as one of the most popular algorithms in hate speech detection. Pitsilis et al. (2018) incorporated user-related features into LSTM and evaluated this approach on a publicly available corpus of 16,000 tweets. The results demonstrated its effectiveness compared to existing state-of-the-art solutions at that time. Additionally, LSTM performed well on cross-language datasets. An LSTM model using word embeddings generated by the genism word2vec model achieved a maximum recall of 0.7504 in a Hinglish (Indian English) dataset (Varade & Pathak, 2020).

Bidirectional Encoder Representations from Transformers (BERT): Since the advent of Transformer (Vaswani et al., 2017), this latest deep learning innovation that has swept through the field of NLP and has triumphed over past algorithms in almost every aspect. Similar with LSTM, Transformer can also handle long-term dependencies, while Transformer doesn't process data sequentially. Instead, it adds the position of each word to the embeddings. Transformer introduces an attention mechanism to enhance the attention weights on important textual features, leading to improved performance (Anjum & Katarya, 2023). In hate speech detection, BERT (Devlin et al., 2019), a pre-trained Transformer models, have gradually become the most popular deep learning algorithm. BERT considers the contextual information of text data by employing bidirectional encoding and incorporates ideas from reinforcement learning. In the training process, BERT utilizes the Masked Language Model (MLM), wherein some words in the input sequence are randomly masked, and the model is trained to predict these masked words. This approach compels the model to comprehend missing information in the context, enhancing its understanding of contextual nuances. These enhancements enable BERT to exhibit superior performance in handling complex language structures (Plaza-del-Arco et al., 2021). In latest research on hate speech, BERT has become the most popular deep learning algorithms in hate speech detections (Jahan & Oussalah, 2023), and a substantial number of studies have employed models based on BERT (Valle-Cano et al., 2023; Su et al., 2023).

State-of-art (SOTA) Performance

A consensus among NLP researchers has emerged: enhancing model architecture complexity and integrating diverse feature representations are key strategies for achieving superior performance (Kang et al., 2020). The emergence of openly available high-quality datasets allows researchers to evaluate their hate speech detection algorithms more intuitively on a public platform. SemEval, an international NLP research workshop dedicated to advancing the state-of-the-art in semantic analysis, has organized competitions released datasets related hate speech and corresponding tasks in the past few years. Table 1 summarizes the algorithms and performances used by outstanding participants in recent three years (Zampieri et al., 2020; Meaney et al., 2021; Abu Farha et al., 2022), serving as a reference standard for future researchers.

Task	Author	Algorithm	F1 Scores
SemEval-20 Task 12	Wiedemann et al. (2020)	MLM-RoBERTa	0.920
SemEval-20 Task 12	Wang et al. (2020)	Multi-lingual method using ERNIE and XLM-RoBERTa	0.919
SemEval-20 Task 12	Pant & Dadu (2020)	XLM-RoBERTa	0.918
SemEval-21 Task 7	Song et al. (2021)	Average prediction from RoBERTa and ALBERT	0.662
SemEval-21 Task 7	Pang et al. (2021)	ERNIE 2.0	0.660
SemEval-21 Task 7	Gupta et al. (2021)	Weighted average of BERT, RoBERTa, ERNIE 2.0, DeBERTa and XLNET	0.657
SemEval-22 Task 6	Yuan et al. (2022)	Weighted average of RoBERTa, RoBERTa -large and XLM-RoBERTa	0.605
SemEval-22 Task 6	Han et al. (2022)	ERNIEM and DeBERTa	0.569
SemEval-22 Task 6	Angel et al. (2022)	Fine-tuned BERT with BERTweet checkpoints	0.530

Table 1. The performance of outstanding participants in tasks related to hate speech detection in the 2020 to 2022 SemEval competitions

It can be observed that excellent performance was achieved in specific datasets by employing SOTA algorithms and appropriate fine-tuning techniques. Almost all outstanding participants utilized models based on BERT, particularly RoBERTa (Liu et al., 2019). However, when tasks became more complex, as in SemEval-22 Task 6, which was about sarcasm detection, even SOTA models struggled to achieve high F1 scores. There were also instances where teams using advanced and complex models failed to outperform baseline models. Moreover, during cross-dataset testing, the excellent performance of these competition models often cannot be replicated due to issues such as bias in text feature extraction and overfitting (Arango et al., 2022).

Discussions

Recent research consistently indicates that detecting hate speech is not an easy task (Arango et al., 2022; Yin & Zubiaga, 2021). The complexity of language and variable real-world application scenarios poses numerous challenges to this classification task. In the following, we elaborate on some crucial challenges

that have been identified in the current work on hate speech detection. Subsequently, we will present our approach to address these research gaps.

Implicit and rapidly changing expressions of hate speech: Hate speech users on social media are aware of the existence of hate speech detection systems and attempt to avoid the detection. Typically, they employ implicit expressions, such as metaphors, stereotypes, homophones (Yin & Zubiaga, 2021). For example, a post reads 'We shouldn't lower our standards just to hire more women' implies women are less qualified, while usually it will not be recognized as hate speech (Sap et al., 2020). Most of the implicitly expressed words appear to be normal vocabulary, making dictionary-based methods challenging to work. Pre-trained deep learning models can help with this problem, but simply increasing the model's complexity does not contribute significant improvements and cannot effectively address this issue (Yin & Zubiaga, 2021).

More critically, the rapidly changing language on social media greatly expands the vocabulary that needs to be detected, posing substantial challenges for pre-trained models. It is difficult to identify implicitly expressed hate using generic models because they require a profound understanding of internet language, often relying on specific contexts or relevant real-world knowledge. To the best of our knowledge, only Sap et al. (2020) attempted experiments using the dataset with independent implicit labels, and the results still indicated that predicting implicit hate expressions remained a significant challenge in the future. While many researchers acknowledge this, there is limited engagement in developing implicit-driven hate speech approaches.

Bias in the limited labelled data: Even though there are many publicly available datasets for hate speech detection, compared to deep learning models with millions or even billions of parameters, this still introduces potential overfitting risks. Constructing labeled datasets is expensive due to the complexity of language, which dictates higher training and labor costs for annotators than typical datasets. A potential improvement method is to leverage unsupervised large language models for automatic algorithmic annotation (Röttger et al., 2021), but the biases inherent in these large models will also be introduced.

The datasets also contain biases. The proportion of hate speech on social media is approximately 3%, and researchers often employ boosted sampling methods to address this sparse issue (Fortuna & Nunes, 2019). However, excessively learning from sampled data leads to a decline in the model's generalization ability. Moreover, publicly available large datasets lack data from minority groups, causing models to potentially make errors in topics related to minority races, minority sexual orientations, etc. This could even cause more serious harm of hate speech towards minority groups (Kim et al., 2020). Metadata helps increase the data dimensions and reduce bias by embedding user demographic information. Nonetheless, collecting sufficient metadata is a challenge and involves legal, ethical, and privacy risks.

Cross context detection: Hate speech exists in different countries, regions, and cultures, but public datasets for research are primarily available in English. The concrete expressions of hate speech are different in various language and cultural environments, while there are many abstract logics that are similar. It would be truly exciting if we could capture these underlying logical patterns to train a model that is universally applicable across languages. Recent research has introduced datasets for different languages, but unfortunately, most studies have remained at the level of scrape-and-report (Wu, 2023). Many explorations prioritize simply testing and algorithmic structure improvements to adapt to datasets in different languages, rather than emphasizing a comprehensive exploration generalizing models trained on one language to another.

The way ahead: The advancement of deep learning technology has significantly enhanced hate speech detection. However, an excessive focus on algorithmic structures often proves effective only on specific datasets. In many cases, simpler algorithms outperform better than the complex ones. Hate speech detection constitutes a complex interdisciplinary field involving computer science, psychology, linguistics, and political science. The existing research has been overly centred on computer science. The adverse effects of most challenges in hate speech detection are introducing bias and reducing generalizability of models. This often results in models performing well in training sets but exhibiting subpar performance in the real world or becoming less reliable with the rapid evolution of online social language. Mitigating model bias and enhancing generalizability not only benefits the practical application of detection models but also fosters the development of potential universally applicable hate speech detection models in the futures.

A research framework based on interdisciplinary perspectives may effectively address the challenges present in the field of hate speech detection. To the best of our knowledge, Lee & Ram (2020) has set a commendable example by being the first to integrate psychological theoretical models into deep learning models. They applied psychological models to compute personality scores based on text and input these scores into a deep learning model. The cross-dataset testing has demonstrated the superior generalization of this approach over SOTA models.

Introducing an interdisciplinary perspective is essentially aimed at better understanding text and assisting models in extracting more semantic features. For instance, using the N-word can be a racially supercharged offensive slur or an expression of cultural identity depending on the messenger and context (Holt, 2018), and these issues translate into hate speech detection algorithm biases (e.g. lexical, dialectical, identity-mentioned). If we introduce perspectives from political science or sociology, gaining a deeper understanding of the cultural context of African American English, we expect to establish a model structure that can effectively reduce algorithm biases.

In summary, we recommend adopting a multidisciplinary approach to enrich the comprehensiveness of the research. Evaluating the possibility of incorporating additional dimensions of feature extraction for hate speech detection is essential. In our future work, we aim to further explore potential interdisciplinary perspectives, delving into the underlying logic of hate speech to propose a hate speech detection method with less bias and better generalizability.

Acknowledgements

Marten Risius is the recipient of an Australian Research Council Australian Discovery Early Career Award (project number DE220101597) funded by the Australian Government.

References

Abu Farha, I., Oprea, S. V., Wilson, S., & Magdy, W. (2022). SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 802–814.

Agarwal, S., & Chowdary, C. R. (2021). Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19. *Expert Systems with Applications*, *185*, 115632.

Angel, J., Aroyehun, S., & Gelbukh, A. (2022). TUG-CIC at SemEval-2021 Task 6: Two-stage Fine-tuning for Intended Sarcasm Detection. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 951–955.

Anjum, & Katarya, R. (2023). Hate speech, toxicity detection in online social media: A recent survey of state of the art and opportunities. *International Journal of Information Security*.

Arango, A., Pérez, J., & Poblete, B. (2022). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, *105*, 101584.

Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, *38*, 100311.

Benbya, H., Pachidi, S., & Jarvenpaa, S. (2021). Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research. *Journal of the Association for Information Systems*, *22*(2).

Bisht, A., Singh, A., Bhadauria, H. S., Virmani, J., & Kriti. (2020). Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model. In S. Jain & S. Paul (Eds.), *Recent Trends in Image and Signal Processing in Computer Vision* (pp. 243–264).

Br Ginting, P. S., Irawan, B., & Setianingsih, C. (2019). Hate Speech Detection on Twitter Using Multinomial Logistic Regression Classification Method. *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, 105–111.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), Article 1.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Florio, K., Basile, V., Polignano, M., Basile, P., & Patti, V. (2020). Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media. *Applied Sciences*, *10*(12), Article 12.

Fortuna, P., & Nunes, S. (2019). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, *51*(4), 1–30.

Gitari, N. D., Zhang, Z., Damien, H., & Long, J. (2015). A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215–230.

Govers, J., Feldman, P., Dant, A., & Patros, P. (2023). Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech. *ACM Computing Surveys*, *55*(14s), 319:1-319:35.

Gunarathne, P., Rui, H., & Seidmann, A. (2022). Racial Bias in Customer Service: Evidence from Twitter. *Information Systems Research*, *33*(1), 43–54.

Gupta, A., Pal, A., Khurana, B., Tyagi, L., & Modi, A. (2021). *Humor@IITK at SemEval-2021 Task 7: Large Language Models for Quantifying Humor and Offensiveness* (arXiv:2104.00933).

Han, Y., Chai, Y., Wang, S., Sun, Y., Huang, H., Chen, G., Xu, Y., & Yang, Y. (2022). *X-PuDu at SemEval-2022 Task 6: Multilingual Learning for English and Arabic Sarcasm Detection* (arXiv:2211.16883).

Holt, L. F. (2018). Dropping the 'N-Word': Examining How a Victim-Centered Approach Could Curtail the Use of America's Most Opprobrious Term. *Journal of Black Studies*, *49*(5), 411–426.

Hosni, M., Abnane, I., Idri, A., Carrillo de Gea, J. M., & Fernández Alemán, J. L. (2019). Reviewing ensemble classification methods in breast cancer. *Computer Methods and Programs in Biomedicine*, *177*, 89–112.

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*, 126232.

Kane, G. C., Alavi, M., Labianca, G., & Borgatti, S. (2014). What' different about social media networks? A framework and research agenda. *MIS Quarterly*, *38*, 274–304.

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, *7*(2), 139–172.

Khan, S., Fazil, M., Sejwal, V. K., Alshara, M. A., Alotaibi, R. M., Kamal, A., & Baig, A. R. (2022). BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. *Journal of King Saud University - Computer and Information Sciences*, *34*(7), 4335–4344.

Kim, J. Y., Ortiz, C., Nam, S., Santiago, S., & Datta, V. (2020). *Intersectional Bias in Hate Speech and Abusive Language Datasets* (arXiv:2005.05921).

Lee, H.-S., Lee, H.-R., Park, J.-U., & Han, Y.-S. (2018). An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decision Support Systems*, *113*, 22–31.

Lee, K., & Ram, S. (2020). PERSONA: Personality-Based Deep Learning for Detecting Hate Speech. *ICIS 2020 Proceedings*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692).

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherjee, A. (2019). Thou Shalt Not Hate: Countering Online Hate Speech. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 369–380.

Meaney, J. A., Wilson, S., Chiruzzo, L., Lopez, A., & Magdy, W. (2021). SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 105–119.

Mullah, N. S., & Zainon, W. M. N. W. (2021). Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. *IEEE Access*, *9*, 88364–88376.

Pang, C., Fan, X., Su, W., Chen, X., Wang, S., Liu, J., Ouyang, X., Feng, S., & Sun, Y. (2021). abcbpc at SemEval-2021 Task 7: ERNIE-based Multi-task Model for Detecting and Rating Humor and Offense. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 286–289.

Pant, K., & Dadu, T. (2020). *Cross-lingual Inductive Transfer to Detect Offensive Language*.(arXiv:2007.03771).

Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in Twitter data using recurrent neural networks. *Applied Intelligence*, *48*(12), 4730–4742.

Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications*, *166*.

Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, *108*, 42–49.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional Tests for Hate Speech Detection Models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 41–58.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4), e1249.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020). Social Bias Frames: Reasoning about Social and Power Implications of Language. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5477–5490.

Schmidt, A., & Wiegand, M. (2017a). A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10.

Song, B., Pan, C., Wang, S., & Luo, Z. (2021). DeepBlueAI at WANLP-EACL2021 task 2: A Deep Ensemble-based Method for Sarcasm and Sentiment Detection in Arabic. *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 390–394.

Su, X., Li, Y., Branco, P., & Inkpen, D. (2023). SSL-GAN-RoBERTa: A robust semi-supervised model for detecting Anti-Asian COVID-19 hate speech on social media. *Natural Language Engineering*.

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., & Daelemans, W. (2016). *A Dictionary-based Approach to Racism Detection in Dutch Social Media* (arXiv:1608.08738).

Valle-Cano, G. D., Quijano-Sánchez, L., Liberatore, F., & Gómez, J. (2023). SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles. *Expert Systems with Applications*, 216.

Varade, R. S., & Pathak, V. B. (2020). Detection of Hate Speech in Hinglish Language. *Machine Learning and Information Processing* (pp. 265–276).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*.

Vidgen, B., Margetts, H., & Harris, A. (2019). *How much online abuse is there? A systematic review of evidence for the UK* (pp. 1–53). Alan Turing Institute.

Wang, S., Liu, J., Ouyang, X., & Sun, Y. (2020). *Galileo at SemEval-2020 Task 12: Multi-lingual Learning for Offensive Language Identification using Pre-trained Language Models* (arXiv:2010.03542).

Wiedemann, G., Yimam, S. M., & Biemann, C. (2020). *UHH-LT at SemEval-2020 Task 12: Fine-Tuning of Pre-Trained Transformer Networks for Offensive Language Detection* (arXiv:2004.11493).

Wu, P. F. (2023). Veni, vidi, vici? On the rise of scrape-and-report scholarship in online reviews research. *Journal of the Association for Information Science and Technology*, *74*(2), 145–149.

Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, *7*, e598.

Yuan, M., Mengyuan, Z., Jiang, L., Mo, Y., & Shi, X. (2022). stce at SemEval-2022 Task 6: Sarcasm Detection in English Tweets. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 820–826.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç. (2020). SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1425–1447.

Zhang, Y., Wang, Q., Li, Y., & Wu, X. (2018). Sentiment Classification Based on Piecewise Pooling Convolutional Neural Network. *Computers, Materials & Continua*, *56*(2), 285–297.