Appendix G: Bibliography for appendices A–F

Abro, S., Shaikh, S., Ali, Z., Khan, S., Mujtaba, G., & Khand, Z. H. (2020). Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, *11*(8), 484–491. https://doi.org/10.14569/IJACSA.2020.0110861

Aggarwal, A., Maurya, K., & Chaudhary, A. (2020). Comparative study for predicting the severity of cyberbullying across multiple social media platforms. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 871–877). IEEE. https://doi.org/10.1109/ICICCS48265.2020.9121046

Agnes, S. A., Solomon, A. A., & Tamilmaran, D. J. C. (2023). Abusive comment detection in social media with bidirectional LSTM model. In *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1368-1373). IEEE. https://doi.org/10.1109/ICSSIT55814.2023.10060887

Agrafiotis, I., Bada, M., Cornish, P., Creese, S., Goldsmith, M., Ignatuschtschenko, E., Roberts, T., & Upton, D. M. (2016). *Cyber harm: Concepts, taxonomy and measurement.* SSRN. https://dx.doi.org/10.2139/ssrn.2828646

Ahmed, T., Ivan, S., Kabir, M., Mahmud, H., & Hasan, K. (2022). Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying. *Social Network Analysis and Mining*, *12*(1), 99. https://doi.org/10.1007/s13278-022-00934-4

Ahmed, T., Kabir, M., Ivan, S., Mahmud, H., & Hasan, K. (2021). Am I being bullied on social media? An ensemble approach to categorize cyberbullying. In *2021 IEEE International Conference on Big Data (Big Data)* (pp. 2442–2453). IEEE. https://doi.org/10.1109/BigData52589.2021.9671594

Aind, A. T., Ramnaney, A., & Sethia, D. (2020). Q-bully: A reinforcement learning based cyberbullying detection framework. In *2020 International Conference for Emerging Technology (INCET)* (pp. 1–6). IEEE. https://doi.org/10.1109/INCET49848.2020.9154092

Akinyemi, J. D., Ibitoye, A. O., Oyewale, C. T., & Onifade, O. F. (2023). Cyberbullying detection and classification in social media texts using machine learning techniques. In Z. Hu, I. Dychka, & M. He (Eds.), *International Conference on Computer Science, Engineering and Education Applications* (pp. 440–449). Springer. https://doi.org/10.1007/978-3-031-36118-0_40

Al-Garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, *63*, 433–443. https://doi.org/10.1016/j.chb.2016.05.051

Ali, W. N. H. W., Mohd, M., & Fauzi, F. (2018). Cyberbullying detection: An overview. In *2018 Cyber Resilience Conference (CRC)* (pp. 1–3). IEEE. https://doi.org/10.1109/CR.2018.8626869

ALjohani, E. J., Yafooz, W. M., & Alsaeedi, A. (2023). Cyberbullying detection approaches: A review. In *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 1310–1316). IEEE. https://doi.org/10.1109/ICIRCA57980.2023.10220688

Alkomah, F., & Ma, X. (2022). A literature review of textual hate speech detection methods and datasets. *Information*, *13*(6), 273. https://doi.org/10.3390/info13060273

Alksasbeh, M. Z., Alqaralleh, B. A., Abukhalil, T., Abukaraki, A., Al Rawashdeh, T., & Al-Jaafreh, M. (2021). Smart detection of offensive words in social media using the soundex algorithm and permuterm index. *International Journal of Electrical and Computer Engineering (IJECE)*, *11*(5), 4431–4438. http://doi.org/10.11591/ijece.v11i5.pp4431-4438

Alonso, C., & Romero, E. (2017). Aggressors and victims in bullying and cyberbullying: A study of personality profiles using the five-factor model. *The Spanish Journal of Psychology*, *20*, E76. https://doi.org/10.1017/sjp.2017.73

Alonso, P., Saini, R., & Kovács, G. (2019). The North at HASOC 2019: Hate speech detection in social media data. In P. Mehta, P. Rosso, P. Majumder, & M. Mitra (Eds.), *Working notes of FIRE 2019: Forum for Information Retrieval Evaluation, Kolkata, India, December 12–15, 2019* (pp. 293–299). CEUR-WS.org. https://ceur-ws.org/Vol-2517/T3-15.pdf

Alonso, P., Saini, R., & Kovács, G. (2020). Hate speech detection using transformer ensembles on the HASOC dataset. In A. Karpov & R. Potapova (Eds.), *Speech and Computer: 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 7–9, 2020, Proceedings* (pp. 13–21). Springer. https://doi.org/10.1007/978-3-030-60276-5_2

Alotaibi, M., Alotaibi, B., & Razaque, A. (2021). A multichannel deep learning framework for cyberbullying detection on social media. *Electronics*, *10*(21), 2664. https://doi.org/10.3390/electronics10212664

Alrashidi, B., Jamal, A., Khan, I., & Alkhathlan, A. (2022). A review on abusive content automatic detection: Approaches, challenges and opportunities. *PeerJ Computer Science*, *8*. https://doi.org/10.7717/PEERJ-CS.1142

Alrehili, A. (2019). Automatic hate speech detection on social media: A brief survey. *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)* (pp. 1–6). IEEE. https://doi.org/10.1109/AICCSA47632.2019.9035228

Ambareen, K., & Meenakshi Sundaram, S. (2023). A survey of cyberbullying detection and performance: Its impact in social media using artificial intelligence. *SN Computer Science*, *4*(6), 859. https://doi.org/10.1007/s42979-023-02301-2

Anjum, & Katarya, R. (2022). Analysis of online toxicity detection using machine learning approaches. In G. Sanyal, C. M. Travieso-González, S. Awasthi, C. M. A. Pinto, & B. R. Purushothama (Eds.), *International Conference on Artificial Intelligence and Sustainable Engineering: Select Proceedings of AISE 2020* (Vol. 836, pp. 381–392). Springer. https://doi.org/10.1007/978-981-16-8542-2_29

Anjum, & Katarya, R. (2024). Hate speech, toxicity detection in online social media: A recent survey of state of the art and opportunities. *International Journal of Information Security*, *23*, 577–608. https://doi.org/10.1007/s10207-023-00755-2

Antypas, D., & Camacho-Collados, J. (2023). Robust hate speech detection in social media: A cross-dataset empirical evaluation. In Y. Chung, P. Röttger, D. Nozza, Z. Talat, & A. Mostafazadeh Davani (Eds.), *The 7th Workshop on Online Abuse and Harms (WOAH)* (pp. 231–242). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.woah-1.25

Anzovino, M., Fersini, E., & Rosso, P. (2018). Automatic identification and classification of misogynistic language on Twitter. In M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, & F. Meziane (Eds.), *Natural Language Processing and Information Systems 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings* (pp. 57–64). Springer. https://doi.org/10.1007/978-3-319-91947-8_6

Ashraf, N., Zubiaga, A., & Gelbukh, A. (2021). Abusive language detection in YouTube comments leveraging replies as conversational context. *PeerJ Computer Science*, *7*. https://doi.org/10.7717/peerj-cs.742

Awal, M. R., Cao, R., Lee, R. K. W., & Mitrović, S. (2021, May). Angrybert: Joint learning target and emotion for hate speech detection. In K. Karlapalem, H. Cheng, N. Ramakrishnan, R. K. Agrawal, P. K. Reddy, J. Srivastava, & T. Chakraborty (Eds.), *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 701–713). Springer. https://doi.org/10.1007/978-3-030-75762-5_55

Babaeianjelodar, M., Poorna Prudhvi, G., Lorenz, S., Chen, K., Mondal, S., Dey, S., & Kumar, N. (2022). Interpretable and high-performance hate and offensive speech detection. In J. Y. C. Chen, G. Fragomeni, H. Degen, & S. Ntoa (Eds.), *International Conference on Human-Computer Interaction* (pp. 233–244). Springer. https://doi.org/10.1007/978-3-031-21707-4_18

Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers & Security*, *90*, 101710. https://doi.org/10.1016/j.cose.2019.101710

Banko, M., MacKeen, B., & Ray, L. (2020). A unified taxonomy of harmful content. In S. Akiwowo, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 125–137). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.alw-1.16

Baydogan, C., & Alatas, B. (2021). Metaheuristic ant lion and moth flame optimization-based novel approach for automatic detection of hate speech in online social networks. *IEEE Access*, *9*, 110047–110062. https://doi.org/10.1109/ACCESS.2021.3102277

Baydogan, C., & Alatas, B. (2022). Deep-Cov19-Hate: A textual-based novel approach for automatic detection of hate speech in online social networks throughout COVID-19 with shallow and deep learning models. *Tehnicki Vjesnik*, *29*(1), 149–156. https://doi.org/10.17559/TV-20210708143535

Beddiar, D., Jahan, M., & Oussalah, M. (2021). *Data expansion using back translation and paraphrasing for hate speech detection. Online Social Networks and Media, 24*, 100153. https://doi.org/10.1016/j.osnem.2021.100153

Bhagya, J., & Deepthi, P. (2021). Cyberbullying detection on social media using SVM. In V. Suma, J. I. Chen, Z. Baig, & H. Wang (Eds.), *Inventive Systems and Control: Proceedings of ICISC 2021* (pp. 17–27). Springer. https://doi.org/10.1007/978-981-16-1395-1_2

Bhatt, C., Saini, N., Chauhan, R., & Sahoo, A. K. (2023). Machine learning techniques for hate speech detection on social media. In *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)* (pp. 1–5). IEEE. https://doi.org/10.1109/CISCT57197.2023.10351228

Bilen, A. (2023). A review: Detection of discrimination and hate speech shared on social media platforms using artificial intelligence methods. In M. Kılıç & S. Bozkuş Kahyaoğlu (Eds.), *Algorithmic discrimination and ethical perspective of artificial intelligence* (pp. 171–181). Springer. https://doi.org/10.1007/978-981-99-6327-0_12

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(2003), 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Bokolo, B. G., & Liu, Q. (2023). Cyberbullying detection on social media using machine learning. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 1–6). IEEE. https://doi.org/10.1109/INFOCOMWKSHPS57453.2023.10226114

Bozyiğit, A., Utku, S., & Nasibov, E. (2021). Cyberbullying detection: Utilizing social media features. *Expert Systems with Applications*, *179,* 115001. https://doi.org/10.1016/j.eswa.2021.115001

Buan, T. A., & Ramachandra, R. (2020). Automated cyberbullying detection in social media using an SVM activated stacked convolution LSTM network. In *Proceedings of the 2020 4th International Conference on Compute and Data Analysis* (pp. 170–174). Association for Computing Machinery. https://doi.org/10.1145/3388142.3388147

Bunde, E. (2021). AI-assisted and explainable hate speech detection for social media moderators–A design science approach. In *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS)*. (pp. 1264–1273). https://doi.org/10.24251/HICSS.2021.154

Calabrese, A., Ross, B., & Lapata, M. (2022). Explainable abuse detection as intent classification and slot filling. *Transactions of the Association for Computational Linguistics*, *10*, 1440–1454. https://doi.org/10.1162/tacl_a_00527

Cécillon, N., Labatut, V., Dufour, R., & Linares, G. (2021). Graph embeddings for abusive language detection. *SN Computer Science*, *2*, 37. https://doi.org/10.1007/s42979-020-00413-7

Neelakandan, S., Sridevi, M., Chandrasekaran, S., Murugeswari, K., Singh Pundir, A. K., Sridevi, R., & Lingaiah, T. B. (2022). Deep learning approaches for cyberbullying detection and classification on social media. *Computational Intelligence and Neuroscience*, *2022*(1), 163458. https://doi.org/10.1155/2022/2163458

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., & 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

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on Twitter. In *Proceedings of the 2017 ACM on Web Science Conference* (pp. 13–22). Association for Computing Machinery. https://doi.org/10.1145/3091478.3091487

Chelmis, C., & Zois, D.-S. (2021). Dynamic, incremental, and continuous detection of cyberbullying in online social media. *ACM Transactions on the Web (TWEB)*, *15*(3), 1–33. https://doi.org/10.1145/3448014

Chen, H., Mckeever, S., & Delany, S. J. (2017). Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media. In: P. Angelov, A. Gegov, C. Jayne, & Q. Shen (Eds.), *Advances in Computational Intelligence Systems. Contributions presented at the 16th UK Workshop on Computational Intelligence, September 7–9, 2016, Lancaster, UK* (pp. 187–205). Springer. https://doi.org/10.1007/978-3-319-46562-3_12

Chen, H., McKeever, S., & Delany, S. J. (2018). A comparison of classical versus deep learning techniques for abusive content detection on social media sites. In S. Staab, O. Koltsova, & D. I. Ignatov (Eds.), *Social Informatics: 10th International Conference, SocInfo 2018, St. Petersburg, Russia, September 25-28, 2018, Proceedings, Part I* (pp. 117–133). Springer. https://doi.org/10.1007/978-3-030-01129-1_8

Chen, H.-Y., & Li, C.-T. (2020). HENIN: Learning heterogeneous neural interaction networks for explainable cyberbullying detection on social media. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2543–2552). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.200

Cheng, L., Li, J., Silva, Y., Hall, D., & Liu, H. (2019a). PI-bully: Personalized cyberbullying detection with peer influence. In S. Kraus (Ed.), *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 5829–5835). International Joint Conferences on Artificial Intelligence. https://doi.org/10.24963/ijcai.2019/808

Cheng, L., Li, J., Silva, Y. N., Hall, D. L., & Liu, H. (2019b). Xbully: Cyberbullying detection within a multi-modal context. In *WSDM '19: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (pp. 339–347). Association for Computing Machinery. https://doi.org/10.1145/3289600.3291037

Dadvar, M., Trieschnigg, D., & De Jong, F. (2014). Experts and machines against bullies: A hybrid approach to detect cyberbullies. In M. Sokolova & P. Beek (Eds.), *Advances in Artificial Intelligence: Proceedings of the 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, Montréal, QC, Canada, May 6-9, 2014* (pp. 275–281). Springer. https://doi.org/10.1007/978-3-319-06483-3_25

Dadvar, M., Trieschnigg, D., Ordelman, R., & De Jong, F. (2013). Improving cyberbullying detection with user context. In P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, & E. Yilmaz (Eds.), *Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings* (pp. 693–696). Springer. https://doi.org/10.1007/978-3-642-36973-5_62

Daniel, R., Murthy, T. S., Kumari, C. D. V. P., Lydia, E. L., Ishak, M. K., Hadjouni, M., & Mostafa, S. M. (2023). Ensemble learning with tournament selected glowworm swarm optimization algorithm for cyberbullying detection on social media. *IEEE Access*, *11*, 123392–123400. https://doi.org/10.1109/ACCESS.2023.3326948

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), 512–515. https://doi.org/10.1609/icwsm.v11i1.14955

De Souza, G. A., & Da Costa-Abreu, M. (2020). Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–6). IEEE. https://doi.org/10.1109/IJCNN48605.2020.9207652

Elsafoury, F., Katsigiannis, S., Pervez, Z., & Ramzan, N. (2021). When the timeline meets the pipeline: A survey on automated cyberbullying detection. *IEEE Access*, *9*, 103541–103563. https://doi.org/10.1109/ACCESS.2021.3098979

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1), 52–61. https://doi.org/10.1609/icwsm.v12i1.15038

Escalante, H. J., Villatoro-Tello, E., Garza, S. E., López-Monroy, A. P., Montes-y-Gómez, M., & Villaseñor-Pineda, L. (2017). Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications*, *89*, 99–111. https://doi.org/10.1016/j.eswa.2017.07.040

Fale, P. N., Goyal, K. K., & Shivani, S. (2023). A hybrid deep learning approach for abusive text detection. *AIP Conference Proceedings*, *2753*(1). https://doi.org/10.1063/5.0128071

Fersini, E., Rosso, P., & Anzovino, M. (2018). Overview of the task on automatic misogyny identification at IberEval 2018. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, & J. Carrillo de Albornoz (Eds.), *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, *CEUR Workshop Proceedings 2150* (pp. 214–228). https://boa.unimib.it/handle/10281/219328

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). https://doi.org/10.1609/icwsm.v12i1.14991

Gandhi, A., Ahir, P., Adhvaryu, K., Shah, P., Lohiya, R., Cambria, E., Poria, S., & Hussain, A. (2024). Hate speech detection: A comprehensive review of recent works. *Expert Systems*, *41*(8), e13562. https://doi.org/10.1111/exsy.13562

Gangurde, A., Mankar, P., Chaudhari, D., & Pawar, A. (2022). A systematic bibliometric analysis of hate speech detection on social media sites. *Journal of Scientometric Research*, *11*(1), 100–111. https://dx.doi.org/10.5530/jscires.11.1.10

Gao, L., Kuppersmith, A., & Huang, R. (2017). Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In G. Kondrak & T. Watanabe (Eds.), *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (Vol 1, pp. 774–782). Asian Federation of Natural Language Processing. https://aclanthology.org/I17-1078

Geetha, R., Karthika, S., Sowmika, C. J., & Janani, B. M. (2021). Auto-Off ID: Automatic detection of offensive language in social media. *Journal of Physics: Conference Series*, *1911*(1), 012012. https://doi.org/10.1088/1742-6596/1911/1/012012

Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gnanasekaran, R. K., Gunasekaran, R. R., Hoffman, K. M., Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, M. J., Naik, S., Nixon, H. L., Ramachandran, P., … D. M. Wu. (2017). A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference* (pp. 229–233). Association for Computing Machinery. https://doi.org/10.1145/3091478.3091509

Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2020). Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1470–1478). IEEE. https://doi.ieeecomputersociety.org/10.1109/WACV45572.2020.9093414

Gongane, V. U., Munot, M. V., & Anuse, A. (2022). Feature representation techniques for hate speech detection on social media: A comparative study. In *2022 International Conference on Signal and Information Processing (IConSIP)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICoNSIP49665.2022.10007458

Gongane, V. U., Munot, M. V., & Anuse, A. D. (2024). A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms. *Journal of Computational Social Science, 7*, 587–623. https://doi.org/10.1007/s42001-024-00248-9

Gopalan, A., Mohanavel, V., Geo, A. A., Rajkumar, G. V., Kavitha, T., & Pooja, P. (2023). Experimental evaluation of robust cyberbullying detection over social media using intelligent learning scheme. In *2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)* (pp. 1–7). IEEE. https://doi.org/10.1109/RMKMATE59243.2023.10368747

Gudumotu, C. E., Nukala, S. R., Reddy, K., Konduri, A., & Gireesh, C. (2023). A survey on deep learning models to detect hate speech and bullying in social media. In A. Biswas, V. B. Semwal, & D. Singh (Eds.), *Artificial Intelligence for Societal Issues* (pp. 27–44). Springer. https://doi.org/10.1007/978-3-031-12419-8_2

Haider, F., Dipty, I., Rahman, F., Assaduzzaman, M., & Sohel, A. (2023). Social media hate speech detection using machine learning approach. In S. C. K. R., N. Sujaudeen, A. Beulah, & H. S. Hamead (Eds.), *Computational Intelligence in Data Science: 6th IFIP TC 12 International Conference, ICCIDS 2023, Chennai, India, February 23–25, 2023, revised selected papers* (pp. 218–229). Springer. https://doi.org/10.1007/978-3-031-38296-3_17

Hamdy, E., Mitrović, J., & Granitzer, M. (2020). nlpUP at SemEval-2020 Task 12: A blazing fast system for offensive language detection. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 2098–2104). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.278

Hani, J., Nashaat, M., Ahmed, M., Emad, Z., Amer, E., & Mohammed, A. (2019). Social media cyberbullying detection using machine learning. *International Journal of Advanced Computer Science and Applications*, *10*(5), 703–707. https://dx.doi.org/10.14569/IJACSA.2019.0100587

Harish, D., Alamelu, M., Manimaran, M., & Jayashakthi, V. P. (2023). Automatic detection of cyberbullying on social media using machine learning. In *2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICAECA56562.2023.10201149

He, B., Ziems, C., Soni, S., Ramakrishnan, N., Yang, D., & Kumar, S. (2021). Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis. In M. Coscia, A. Cuzzocrea, & K. Shu (Ed.), *ASONAM '21: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 90–94). Association for Computing Machinery. https://doi.org/10.1145/3487351.3488324

Herath, M., Atapattu, T., Dung, H. A., Treude, C., & Falkner, K. (2020). AdelaideCyC at SemEval-2020 task 12: Ensemble of classifiers for offensive language detection in social media. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1516–1523). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.198

Huang, Q., Singh, V. K., & Atrey, P. K. (2014). Cyber bullying detection using social and textual analysis. In *SAM '14: Proceedings of the 3rd International Workshop on Socially-Aware Multimedia* (pp. 3–6). Association for Computing Machinery. https://doi.org/10.1145/2661126.2661133

Huang, Y., Song, R., Giunchiglia, F., & Xu, H. (2022). A multitask learning framework for abuse detection and emotion classification. *Algorithms*, *15*(4). https://doi.org/10.3390/a15040116

Hussein, F. N. A., & Aleqabie, H. J. (2023). Cyberbullying detection on social media: A brief survey. In *2023 Second International Conference on Advanced Computer Applications (ACA)* (pp. 1–6). IEEE. https://doi.org/10.1109/ACA57612.2023.10346758

Ibrahim, M., Torki, M., & El-Makky, N. (2020). AlexU-BackTranslation-TL at SemEval-2020 Task 12: Improving offensive language detection using data augmentation and transfer learning. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1881–1890). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.248

Istaiteh, O., Al-Omoush, R., & Tedmori, S. (2020). Racist and sexist hate speech detection: Literature review. In *2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)* (pp. 95–99). https://doi.org/10.1109/IDSTA50958.2020.9264052

Iwendi, C., Srivastava, G., Khan, S., & Maddikunta, P. K. R. (2023). *Cyberbullying detection solutions based on deep learning architectures*. *29*(3), 1839–1852. https://doi.org/10.1007/s00530-020-00701-5

Jahan, M. S., Beddiar, D. R., Oussalah, M., & Mohamed, M. (2022). Data expansion using wordnet-based semantic expansion and word disambiguation for cyberbullying detection. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 1761–1770). European Language Resources Association. https://aclanthology.org/2022.lrec-1.187

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546,* 126232. https://doi.org/10.1016/j.neucom.2023.126232

Jahn, L., Rendsvig, R. K., Flammini, A., Menczer, F., & Hendricks, V. F. (2023). *Friction interventions to curb the spread of misinformation on social media*. arXiv. https://doi.org/10.48550/arXiv.2307.11498

Jain, V., Kumar, V., Pal, V., & Vishwakarma, D. K. (2021). Detection of cyberbullying on social media using machine learning. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1091–1096). IEEE. https://doi.org/10.1109/ICCMC51019.2021.9418254

Karatsalos, C., & Panagiotakis, Y. (2020). Attention-based method for categorizing different types of online harassment language. In P. Cellier & K. Driessens (Eds.), *Machine learning and knowledge discovery in databases: International workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II* (pp. 321–330). Springer. https://doi.org/10.1007/978-3-030-43887-6_26

Kaur, S., Singh, S., & Kaushal, S. (2021). Abusive content detection in online user-generated data: A survey. *Procedia Computer Science, 189*, 274–281. https://doi.org/10.1016/j.procs.2021.05.098

Kavatagi, S., & Rachh, R. (2021). A context aware embedding for the detection of hate speech in social media networks. In *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)* (pp. 1–4). IEEE. https://doi.org/10.1109/SMARTGENCON51891.2021.9645877

Kavitha, S., Anchitaalagammai, J., Murali, S., Deepalakshmi, R., Himal, L., & Suryakanth, M. (2023). Smart language checker: A machine learning solution for offensive language detection in social media. In *2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICDSAAI59313.2023.10452454

Kazbekova, G., Ismagulova, Z., Kemelbekova, Z., Tileubay, S., Baimurzayev, B., & Bazarbayeva, A. (2023). Offensive language detection on online social networks using hybrid deep learning architecture. *International Journal of Advanced Computer Science and Applications*, *14*(11), 793–805. https://doi.org/10.14569/IJACSA.2023.0141180

Kennedy, B., Atari, M., Mostafazadeh Davani, A., Yeh, L., Omrani, A., Kim, Y., Coombs, K. Jr., Havaldar, S., Portillo-Wightman, G., Gonzalez, E., Hoover, J., Azatian, A., Hussain, A., Lara, A., Cardenas, G., Omary, A., Park, C., Wang, X., Wijaya, C., … M. Dehghani. (2022). Introducing the Gab hate corpus: Defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, *56,* 79–108. https://doi.org/10.1007/s10579-021-09569-x

Kennedy, C. J., Bacon, G., Sahn, A., & von Vacano, C. (2020). *Constructing interval variables via faceted Rasch measurement and multitask deep learning: A hate speech application*. arXiv. https://doi.org/10.48550/arXiv.2009.10277

Kogilavani, S. V., Malliga, S., Jaiabinaya, K., Malini, M., & Kokila, M. M. (2023). Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings*, *81.2.,* 630–633. https://doi.org/10.1016/j.matpr.2021.04.102

Kovács, G., Alonso, P., & Saini, R. (2021). Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. *SN Computer Science*, *2*(2), 95. https://doi.org/10.1007/s42979-021-00457-3

Kovács, G., Alonso, P., Saini, R., & Liwicki, M. (2022). Leveraging external resources for offensive content detection in social media. *AI Communications*, *35*(2), 87–109. https://doi.org/10.3233/AIC-210138

Kumar, A., & Kumar, S. (2023). Hate speech detection in multi-social media using deep learning. In R. N. Shaw, M. Paprzycki, & A. Ghosh (Eds.), *Advanced Communication and Intelligent Systems Second International Conference, ICACIS 2023, Warsaw, Poland, June 16–17, 2023, Revised Selected Papers, Part I* (pp. 59–70). Springer. https://doi.org/10.1007/978-3-031-45121-8_6

Kumar, A., & Sachdeva, N. (2022). A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media. *World Wide Web*, *25*(4), 1537–1550. https://doi.org/10.1007/s11280-021-00920-4

Kumar, A., Tyagi, V., & Das, S. (2021). Deep learning for hate speech detection in social media. In *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)* (pp. 1–4). IEEE. https://doi.org/10.1109/GUCON50781.2021.9573687

Kumar, R., & Bhat, A. (2022). A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media. *International Journal of Information Security*, *21*(6), 1409–1431. https://doi.org/10.1007/s10207-022-00600-y

Li, Z., & Shimada, K. (2022). Combining pre-trained language models and features for offensive language detection. In *2022 13th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)* (pp. 5–10). IEEE. https://doi.org/10.1109/IIAI-AAI-Winter58034.2022.00012

Liu, P., Li, W., & Zou, L. (2019). NULI at SemEval-2019 Task 6: Transfer learning for offensive language detection using bidirectional transformers. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 87–91). Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2011

Liu, Y., Zavarsky, P., & Malik, Y. (2019). Non-linguistic features for cyberbullying detection on a social media platform using machine learning. In J. Vaidya, X. Zhang, & J. Li (Eds.), *Cyberspace Safety and Security: 11th International Symposium, CSS 2019, Guangzhou, China, December 1–3, 2019, Proceedings, Part I* (pp. 391–406). Springer. https://doi.org/10.1007/978-3-030-37337-5_31

López-Vizcaíno, M., Nóvoa, F. J., Artieres, T., & Cacheda, F. (2023). Site agnostic approach to early detection of cyberbullying on social media networks. *Sensors*, *23*(10), 4788. https://doi.org/10.3390/s23104788

Lu, N., Wu, G., Zhang, Z., Zheng, Y., Ren, Y., & Choo, K.-K. R. (2020). Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts. *Concurrency and Computation: Practice and Experience*, *32*(23), e5627. https://doi.org/10.1002/cpe.5627

Malik, P., Aggrawal, A., & Vishwakarma, D. K. (2021). Toxic speech detection using traditional machine learning models and bert and fasttext embedding with deep neural networks. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1254–1259). IEEE. https://doi.org/10.1109/ICCMC51019.2021.9418395

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019). Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In P. Majumder, M. Mitra, S. Gangopadhyay, & P. Mehta (Eds.), *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation* (pp. 14–17). Association for Computing Machinery. https://doi.org/10.1145/3368567.3368584

Mansur, Z., Omar, N., & Tiun, S. (2023). Twitter hate speech detection: A systematic review of methods, taxonomy analysis, challenges, and opportunities. *IEEE Access*, *11*, 16226–16249. https://doi.org/10.1109/ACCESS.2023.3239375

Mathur, S. A., Isarka, S., Dharmasivam, B., & Jaidhar, C. (2023). Analysis of tweets for cyberbullying detection. In *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)* (pp. 269–274). IEEE. https://doi.org/10.1109/ICSCCC58608.2023.10176416

Mehta, H., & Passi, K. (2022). Social media hate speech detection using explainable artificial intelligence (XAI). *Algorithms*, *15*(8), 291. https://doi.org/10.3390/a15080291

Mercan, V., Jamil, A., Hameed, A. A., Magsi, I. A., Bazai, S., & Shah, S. A. (2021). Hate speech and offensive language detection from social media. In *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICECube53880.2021.9628255

Miran, A. Z., & Yahia, H. S. (2023). Hate speech detection in social media (Twitter) using neural network. *Journal of Mobile Multimedia*, *19*(3), 765–798. http://dx.doi.org/10.13052/jmm1550-4646.1936

Miró-Llinares, F., & Rodríguez-Sala, J. J. (2016). Cyber hate speech on Twitter: Analyzing disruptive events from social media to build a violent communication and hate speech taxonomy. *International Journal of Design & Nature and Ecodynamics*, *11*(3), 406–415. https://doi.org/10.2495/DNE-V11-N3-406-415

Modi, S. (2018). AHTDT- Automatic hate text detection techniques in social media. In *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)* (pp. 1–3). IEEE. https://doi.org/10.1109/ICCSDET.2018.8821128

Mohtaj, S., & Möller, S. (2022). On the importance of word embedding in automated harmful information detection. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, Proceedings* (pp. 251–262). Springer. https://doi.org/10.1007/978-3-031-16270-1_21

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020a). A BERT-based transfer learning approach for hate speech detection in online social media. In H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, & L. M. Rocha (Eds.), *Complex networks and their applications VIII: Volume 1, proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019* (pp. 928–940). Springer. https://doi.org/10.1007/978-3-030-36687-2_77

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020b). Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS One*, *15*(8), e0237861. https://doi.org/10.1371/journal.pone.0237861

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. https://doi.org/10.1109/ACCESS.2021.3089515

Muneer, A., Alwadain, A., Ragab, M. G., & Alqushaibi, A. (2023). Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT. *Information (Switzerland)*, *14*(8), 467. https://doi.org/10.3390/info14080467

Murshed, B. A. H., Abawajy, J., Mallappa, S., Saif, M. A. N., & Al-Ariki, H. D. E. (2022). DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform. *IEEE Access*, *10*, 25857–25871. https://doi.org/10.1109/ACCESS.2022.3153675

Murshed, B. A. H., Suresha, Abawajy, J., Saif, M. A. N., Abdulwahab, H. M., & Ghanem, F. A. (2023). FAEO-ECNN: Cyberbullying detection in social media platforms using topic modelling and deep learning. *Multimedia Tools and Applications*, *82*(30), 46611–46650. https://doi.org/10.1007/s11042-023-15372-3

Muzakir, A., Adi, K., & Kusumaningrum, R. (2023). Classification of hate speech language detection on social media: Preliminary study for improvement. In M. B. Ahmed, B. A. Abdelhakim, B. K. Ane, & D. Rosiyadi (Eds.), *Emerging trends in intelligent systems & network security* (pp. 146–156). Springer. https://doi.org/10.1007/978-3-031-15191-0_14

Nagar, S., Gupta, S., Bahushruth, C., Barbhuiya, F. A., & Dey, K. (2022). Hate speech detection on social media using graph convolutional networks. In R. M. Benito, C. Cherifi, H. Cherifi, E. Moro, L. M. Rocha, & M. Sales-Pardo (Eds.), *Complex networks & their applications X: Volume 2, proceedings of the Tenth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2021 10* (pp. 3–14). Springer. https://doi.org/10.1007/978-3-030-93413-2_1

Nascimento, F. R. S., Cavalcanti, G. D. C., & Da Costa-Abreu, M. (2022). Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning. *Expert Systems with Applications*, *201,* 117032. https://doi.org/10.1016/j.eswa.2022.117032

Nath, N., George, J. P., Kesan, A., & Rodrigues, A. (2022). An efficient deep learning-based hybrid architecture for hate speech detection in social media. In S. Shukla, X. Gao, J. V. Kureethara, & D. Mishra (Eds.), *Data science and security, proceedings of IDSCS 2022* (pp. 347–355). Springer. https://doi.org/10.1007/978-981-19-2211-4_30

Nisha, M., & Jebathangam, J. (2022). Detection and classification of cyberbullying in social media using text mining. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology* (pp. 856–861). IEEE. https://doi.org/10.1109/ICECA55336.2022.10009445

Nitya Harshitha, T., Prabu, M., Suganya, E., Sountharrajan, S., Bavirisetti, D. P., Gadde, N., & Uppu, L. S. (2024). ProTect: A hybrid deep learning model for proactive detection of cyberbullying on social media. *Frontiers in Artificial Intelligence*, *7*, 1269366. https://doi.org/10.3389/frai.2024.1269366

Nocentini, A., Calmaestra, J., Schultze-Krumbholz, A., Scheithauer, H., Ortega, R., & Menesini, E. (2010). Cyberbullying: Labels, behaviours and definition in three European countries. *Australian Journal of Guidance and Counselling*, *20*(2), 129–142. https://doi.org/10.1375/ajgc.20.2.129

Omran, E., Al Tararwah, E., & Al Qundus, J. (2023). A comparative analysis of machine learning algorithms for hate speech detection in social media. *Online Journal of Communication and Media Technologies*, *13*(4), e202348. https://doi.org/10.30935/ojcmt/13603

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4675–4684). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1474

Pahuja, V., Neema, S., & Dubey, R. (2023). Securing social spaces: Cyberbullying detection with ML and DL on social media platforms. In *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)* (pp. 1471–1476). IEEE. https://doi.org/10.1109/icscna58489.2023.10370114

Pariyani, B., Shah, K., Shah, M., Vyas, T., & Degadwala, S. (2021). Hate speech detection in Twitter using natural language processing. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 1146–1152). IEEE. https://doi.org/10.1109/ICICV50876.2021.9388496

Pavlopoulos, J., Thain, N., Dixon, L., & Androutsopoulos, I. (2019). ConvAI at SemEval-2019 Task 6: Offensive language identification and categorization with perspective and BERT. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 571–576). Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-2102

Perera, A., & Fernando, P. (2021). Accurate cyberbullying detection and prevention on social media. *Procedia Computer Science*, *181*, 605–611. https://doi.org/10.1016/j.procs.2021.01.207

Phung, H. T., Dang, H. K. L., & Pham, M. T. (2020). Cyberbullying detection based on word curve representations using B-spline interpolation. In *Proceedings of the 4th International Conference on Future Networks and Distributed Systems*, *50*, 1–7. https://doi.org/10.1145/3440749.3442657

Pradhan, A., Yatam, V. M., & Bera, P. (2020). Self-attention for cyberbullying detection. In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)* (pp. 1–6). IEEE. https://doi.org/10.1109/CyberSA49311.2020.9139711

Preetham, J., & Anitha, J. (2023). Offensive language detection in social media using ensemble techniques. In *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)* (pp. 805–808). IEEE. https://doi.org/10.1109/ICCPCT58313.2023.10245673

Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4755–4764). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1482

Qian, W., Yu, S., Nie, Z., Lu, X. S., Liu, H., & Huang, B. (2023). Improved hierarchical attention networks for cyberbullying detection via social media data. In *2023 IEEE International Conference on Networking, Sensing and Control (ICNSC)* (pp. 407–409). IEEE. http://dx.doi.org/10.1109/ICNSC58704.2023.10319023

Qiu, J., Hegde, N., Moh, M., & Moh, T.-S. (2022). Investigating user information and social media features in cyberbullying detection. In *2022 IEEE International Conference on Big Data (Big Data)* (pp. 3063–3070). IEEE. https://doi.org/10.1109/BigData55660.2022.10020305

Qureshi, K. A., & Sabih, M. (2021). Un-compromised credibility: Social media based multi-class hate speech classification for text. *IEEE Access*, *9*, 109465–109477. https://doi.org/10.1109/ACCESS.2021.3101977

Ramiandrisoa, F. (2022). Multi-task learning for hate speech and aggression detection. In L. Tamine, E. Amigó, & J. Mothe (Eds.), *Proceedings of the 2nd Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2022) Samatan, Gers, France, July 4-7, 2022*. https://ceur-ws.org/Vol-3178/CIRCLE_2022_paper_31.pdf

Rawat, A., Kumar, S., & Samant, S. S. (2024). Hate speech detection in social media: Techniques, recent trends, and future challenges. *Wiley Interdisciplinary Reviews: Computational Statistics*, *16*(2), e1648. https://doi.org/10.1002/wics.1648

Reddy, B. A. C., Chandra, G. K., Sisodia, D. S., & Anuragi, A. (2023). Balancing techniques for improving automated detection of hate speech and offensive language on social media. In *2023 2nd International Conference for Innovation in Technology (INOCON)* (pp. 1–8). IEEE. https://doi.org/10.1109/INOCON57975.2023.10101157

Rezvan, M., Shekarpour, S., Balasuriya, L., Thirunarayan, K., Shalin, V. L., & Sheth, A. (2018). A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 33–36). Association for Computing Machinery. https://doi.org/10.1145/3201064.3201103

Ribeiro, M., Calais, P., Santos, Y., Almeida, V., & Meira Jr, W. (2018). Characterizing and detecting hateful users on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). https://doi.org/10.1609/icwsm.v12i1.15057

Sachdeva, J., Chaudhary, K. K., Madaan, H., & Meel, P. (2021). Text based hate-speech analysis. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)* (pp. 661–668). IEEE. https://doi.org/10.1109/ICAIS50930.2021.9396013

Saha, P., Mathew, B., Goyal, P., & Mukherjee, A. (2019). *Hatemonitors: Language agnostic abuse detection in social media.* arXiv.

Sahana, V., Anil Kumar, K. M., & Darem, A. A. (2023). A comparative analysis of machine learning techniques for cyberbullying detection on FormSpring in textual modality. *International Journal of Computer Network and Information Security*, *15*(4), 36–47. https://doi.org/10.5815/ijcnis.2023.04.04

Sajadi Ansari, F., Barhamgi, M., Khelifi, A., & Benslimane, D. (2021). An approach to detect cyberbullying on social media. In C. Attiogbé & S. Ben Yahia (Eds.), *Model and Data Engineering: 10th International Conference, MEDI 2021, Tallinn, Estonia, June 21–23, 2021, proceedings* (pp. 53–66). Springer. https://doi.org/10.1007/978-3-030-78428-7_5

Salawu, S., Lumsden, J., & He, Y. (2021). A large-scale English multi-label Twitter dataset for cyberbullying and online abuse detection. In A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (pp. 146–156). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.woah-1.16

Salehgohari, A., Mirhosseini, M., Tabrizchi, H., & Koczy, A. V. (2022). Abusive language detection on social media using bidirectional long-short term memory. In *2022 IEEE 26th International Conference on Intelligent Engineering Systems (INES)* (pp. 000243–000248). IEEE. https://doi.org/10.1109/INES56734.2022.9922628

Samory, M., Sen, I., Kohne, J., Flöck, F., & Wagner, C. (2021). "Call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 573–584. https://doi.org/10.1609/icwsm.v15i1.18085

Sathishkumar, R., Karthikeyan, T., Shamsundar, S., & Shamsundar, S. M. (2023). Ensemble text classification with TF-IDF vectorization for hate speech detection in social media. In *2023 International Conference on System, Computation, Automation and Networking (ICSCAN)* (pp. 1–7). IEEE. http://dx.doi.org/10.1109/ICSCAN58655.2023.10395354

Shakeel, N., & Dwivedi, R. K. (2022). A survey on detection of cyberbullying in social media using machine learning techniques. In *Intelligent Communication Technologies and Virtual Mobile Networks, proceedings of ICICV 2022* (pp. 323–340). Springer. https://doi.org/10.1007/978-981-19-1844-5_25

Shankar, K., Abirami, A., Indira, K., Angeline, C. N., & Shubhavya, K. (2022). Cyberbullying detection in social media using supervised ML and NLP techniques. In *Communication and Intelligent Systems, proceedings of ICCIS 2021* (pp. 817–828). Springer. https://doi.org/10.1007/978-981-19-2130-8_63

Sharif, O., Hossain, E., & Hoque, M. M. (2021). NLP-CUET@DravidianLangTech-EACL2021: Offensive language detection from multilingual code-mixed text using Transformers. In B. R. Chakravarthi, R. Priyadharshini, A. Kumar M, P. Krishnamurthy, & E. Sherly (Eds.), *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages* (pp. 255–261). Association for Computational Linguistics. https://aclanthology.org/2021.dravidianlangtech-1.35

Sharma, A., & Bhalla, R. (2022). Automatic and advance techniques for hate speech detection on social media: A review. In *2022 Algorithms, Computing and Mathematics Conference (ACM)* (pp. 54–61). IEEE. https://doi.org/10.1109/ACM57404.2022.00017

Sharma, G., Brar, G. S., Singh, P., Gupta, N., Kalra, N., & Parashar, A. (2022). An exploration of machine learning and deep learning techniques for offensive text detection in social media—A systematic review. In D. Gupta, A. Khanna, A. E. Hassanien, S. Anand, & A. Jaiswal (Eds.), *International Conference on Innovative Computing and Communications, proceedings of ICICC 2022* (Vol. 3, pp. 541–559). Springer. http://dx.doi.org/10.1007/978-981-19-3679-1_45

Sharma, P., & Tiwari, R. K. (2023). Deep learning approach for hate and non hate speech detection in online social media. In *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)* (pp. 492–496). IEEE. https://doi.org/10.1109/ICTACS59847.2023.10390417

Singh, N. K., Singh, P., & Chand, S. (2022). Deep learning-based methods for cyberbullying detection on social media. In P. Nand, M. Singh, M. Kaur, & V. Jain (Eds.), *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 521–525). IEEE. https://doi.org/10.3390/fi15050179

Singh, N. M., & Sharma, S. K. (2024). An efficient automated multi-modal cyberbullying detection using decision fusion classifier on social media platforms. *Multimedia Tools and Applications*, *83*(7), 20507–20535. https://doi.org/10.1007/s11042-023-16402-w

Singh, R. K., Sanjay, H., SA, P. J., Rishi, H., & Bhardwaj, S. (2023). NLP based hate speech detection and moderation. In *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)* (pp. 1–5). IEEE. https://doi.org/10.1109/CSITSS60515.2023.10333320

Singh, V. K., Ghosh, S., & Jose, C. (2017). Toward multimodal cyberbullying detection. In *CHI EA '17: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2090–2099). Association for Computing Machinery. https://doi.org/10.1145/3027063.3053169

Singh, V. K., Huang, Q., & Atrey, P. K. (2016). Cyberbullying detection using probabilistic socio-textual information fusion. In R. Kumar, J. Caverlee, & H. Tong (Eds.), *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 884–887). IEEE. https://doi.org/10.1109/ASONAM.2016.7752342

Sookarah, D., & Ramwodin, L. S. (2022). Combatting online harassment by using transformer language models for the detection of emotions, hate speech and offensive language on social media. In *2022 4th International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM)* (pp. 1–6). IEEE. https://doi.org/10.1109/ELECOM54934.2022.9965237

Suhas Bharadwaj, R., Kuzhalvaimozhi, S., & Vedavathi, N. (2022). A novel multimodal hybrid classifier based cyberbullying detection for social media platform. In R. Šilhavý, P. Šilhavý, & Z. Prokopová (Eds.), *Proceedings of the Computational Methods in Systems and Software* (pp. 689–699). Springer. http://dx.doi.org/10.1007/978-3-031-21438-7_57

Sultan, D., Mendes, M., Kassenkhan, A., & Akylbekov, O. (2023). Hybrid CNN-LSTM network for cyberbullying detection on social networks using textual contents. *International Journal of Advanced Computer Science and Applications*, *14*(9). https://dx.doi.org/10.14569/IJACSA.2023.0140978

Suryawanshi, S., Chakravarthi, B. R., Arcan, M., & Buitelaar, P. (2020). Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In R. Kumar, A. Kr. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, & D. Kadar (Eds.), *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying* (pp. 32–41). European Language Resources Association (ELRA). https://aclanthology.org/2020.trac-1.6

Tanase, M.-A., Cercel, D.-C., & Chiru, C. (2020). UPB at SemEval-2020 Task 12: Multilingual offensive language detection on social media by fine-tuning a variety of BERT-based models. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 2222–2231). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.296

Thapa, S., Shah, A., Jafri, F., Naseem, U., & Razzak, I. (2022). A multi-modal dataset for hate speech detection on social media: Case-study of Russia-Ukraine conflict. In A. Hürriyetoğlu, H. Tanev, V. Zavarella, & E. Yörük (Eds.), *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)* (pp. 1–6). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.case-1.1

Themeli, C., Giannakopoulos, G., & Pittaras, N. (2019). A study of text representations for hate speech detection. In A. Gelbukh (Ed.), *International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 424–437). Springer. https://doi.org/10.1007/978-3-031-24340-0_32

Thenmozhi, D., Pr, N., Arunima, S., & Sengupta, A. (2020). Ssn_nlp at SemEval 2020 Task 12: Offense target identification in social media using traditional and deep machine learning approaches. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 2155–2160). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.286

Thenmozhi, D., Sharavanan, S., Chandrabose, A., & Chandrabose, A. (2019). SSN_NLP at SemEval-2019 Task 6: Offensive language identification in social media using traditional and deep machine learning approaches. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 739–744). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/S19-2130

Van Bruwaene, D., Huang, Q., & Inkpen, D. (2020). A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation*, *54*(4), 851–874. https://doi.org/10.1007/s10579-020-09488-3

Vidgen, B., Hale, S., Guest, E., Margetts, H., Broniatowski, D., Waseem, Z., Botelho, A., Hall, M., & Tromble, R. (2020). Detecting East Asian prejudice on social media. In S. Akiwowo, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 162–172). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.alw-1.19

Vora, D., Mukherjee, A., Repaka, S., Das, S., & Ingle, S. (2023). Multimodal cyberbullying detection on social media: Review and challenges. In *2023 International Conference on Integration of Computational Intelligent System (ICICIS)* (pp. 1–8). IEEE. http://dx.doi.org/10.1109/ICICIS56802.2023.10430250

Wang, K., Xiong, Q., Wu, C., Gao, M., & Yu, Y. (2020). Multi-modal cyberbullying detection on social networks. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN48605.2020.9206663

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In D. Bamman, A. S. Doğruöz, J. Eisenstein, D. Hovy, D. Jurgens, B. O'Connor, A. Oh, O. Tsur, & S. Volkova (Eds.), *Proceedings of the First Workshop on NLP and Computational Social Science* (pp. 138–142). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-5618

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In J. Andreas, E. Choi, & A. Lazaridou (Eds.), *Proceedings of the NAACL Student Research Workshop* (pp. 88–93). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-2013

Wijesiriwardene, T., Inan, H., Kursuncu, U., Gaur, M., Shalin, V. L., Thirunarayan, K., Sheth, A., & Arpinar, I. B. (2020). Alone: A dataset for toxic behavior among adolescents on Twitter. In S. Aref, K. Bontcheva, M. Braghieri, F. Dignum, F. Giannotti, F. Grisolia, & D. Pedreschi (Eds.), *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, proceedings* (pp. 427–439). Springer. https://doi.org/10.1007/978-3-030-60975-7_31

Wright, M. F. (2021). Cyberbullying: Definition, behaviors, correlates, and adjustment problems. In *Encyclopedia of information science and technology* (5th ed., pp. 356–373). IGI Global. http://dx.doi.org/10.4018/978-1-7998-3479-3.ch026

Xiang, T., MacAvaney, S., Yang, E., & Goharian, N. (2021). ToxCCIn: Toxic content classification with interpretability. In O. De Clercq, A. Balahur, J. Sedoc, V. Barriere, S. Tafreshi, S. Buechel, & V. Hoste (Eds.), *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 1–12). Association for Computational Linguistics. https://aclanthology.org/2021.wassa-1.1

Xingyi, G., & Adnan, H. (2024). Potential cyberbullying detection in social media platforms based on a multi-task learning framework. *International Journal of Data and Network Science*, *8*(1), 25–34. https://doi.org/10.5267/j.ijdns.2023.10.021

Yao, M. (2019). Robust detection of cyberbullying in social media. In L. Liu & R. White (Eds.), *WWW '19: Companion proceedings of the 2019 World Wide Web Conference* (pp. 61–66). Association for Computing Machinery. https://doi.org/10.1145/3308560.3314196

Yao, M., Chelmis, C., & Zois, D.-S. (2019). Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In L. Liu & R. White (Eds.), *WWW '19: The World Wide Web Conference* (pp. 3427–3433). Association for Computing Machinery. https://doi.org/10.1145/3308558.3313462

Yi, P., & Zubiaga, A. (2023a). Learning like human annotators: Cyberbullying detection in lengthy social media sessions. In Y. Ding, J. Tang, J. Sequeda, L. Aroyo, C. Castillo, & G. Houben (Eds.), *WWW '23: Proceedings of the ACM Web Conference 2023* (pp. 4095–4103). Association for Computing Machinery. https://doi.org/10.1145/3543507.3583873

Yi, P., & Zubiaga, A. (2023b). Session-based cyberbullying detection in social media: A survey. *Online Social Networks and Media*, *36,* 100250. https://doi.org/10.1016/j.osnem.2023.100250

Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, *7*, 1–38. https://doi.org/10.7717/PEERJ-CS.598

Yuan, L., Wang, T., Ferraro, G., Suominen, H., & Rizoiu, M.-A. (2023). Transfer learning for hate speech detection in social media. *Journal of Computational Social Science*, *6*(2), 1081–1101. https://doi.org/10.1007/s42001-023-00224-9

Zampieri, N., Illina, I., & Fohr, D. (2021). Multiword expression features for automatic hate speech detection. In E. Métais, F. Meziane, H. Horacek, & E. Kapetanios (Eds.), *Natural Language Processing and Information Systems*. 26th *International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23–25, 2021, proceedings* (pp. 156–164). Springer. https://doi.org/10.1007/978-3-030-80599-9_14