Hate Speech Detection beyond plain Natural Language Processing

Notebook for FDIA at ESSIR 2023

Kwabena Odame Akomeah

University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

Abstract

The usage of Social Media Networks has increased over the years, having become a common space where users are open to relate their opinions and sentiments which may lead to a wide spread of hatred or abusive messages, misinformation, fake news and claims among others. It is in the best interest of stakeholders that an attempt to moderate these issues is made as it infringes on human rights, incites people and connotes violence. This paper is a synopsis describing how certain promising ideas and directions for a doctoral thesis in the field of Hate Speech Detection that shall employ state-of-the-art machine learning algorithms and artificial intelligence into the detection of hate speech on social media platforms.

Keywords

Hate Speech Detection, Transformers, Embeddings, Data Augmentation, Social Network Analysis

1. Introduction

The field of NLP has seen much development in recent years[1, 2]. The development of Transformer based algorithms and large language models have benefited many research problems such as sentiment analysis and embedding generation of texts which is currently being applied in Social Media Analysis tasks[3, 4]. This include the use of BERT-based algorithms in tasks such as automated Fake News and Hate Speech Detection, taking into consideration news contents, social media replies, and external knowledge[5, 6, 7, 8]. The development of text generation and augmentation that captures semantic similarity through the use of word and sentence embeddings produced by deep language models into extractive summarization techniques based on graph centrality have as well in recent times been beneficial[9, 10, 11].

Hate Speech Detection is a very active field specifically in the broader field of Natural Language Processing that has particularly gained attention and vibrance over the last decade[5, 12]. Due to the increasing use of social media and access to the internet, unfortunately, the flip-side of this increasing connectivity is cyber-bullying and hatred among other hurtful and anti-social behaviours which include the spread of fake news and claims[13, 8, 6, 7, 12].

There is a thin line between hate crimes and hate speech[14, 15]. In effect criminalizing the latter has been ongoing by most liberal democracies since the 1960s, significantly in Western Europe due to the historical factor[16]. Nevertheless, there has been a big discussion on its

The 14th European Summer School on Information Retrieval ESSIR 2023

kwabena-odame.akomeah@ur.de (K. O. Akomeah)

significant limitations on freedom of speech and core liberal values[17]. One can argue for freedom of speech but one cannot deny the fact that when this freedom has the potential to ultimately incite the masses in committing crimes against humanity, then laws have to take their course[16, 15].

An interesting argument on the topic is that restrictive laws are most easily justified if they punish hateful expressions and speech when they inflict significant harm to individuals, incite violence or stir up extreme hatred but not when it is merely offensive or hurtful[16]. This perspective focuses on assessing the level of harm of particular events in a way that has seldom been systematically done before[16]. Moreover, as regulations of hate speech are on the books, enforcement has become a critical component of the equation[17].

It is therefore imperative that hate speech is checked in such a manner where human rights of users are also not infringed upon as spelt out boldly by the European Commission's statement on Justice and fundamental rights[1] and the US government[2][18, 16, 5] .

Annual renowned shared-tasks competitions and conferences such as the SemEval[10, 19], Evalita[20], CLEF[21] among others have provided a common platform and a direction for researchers to discuss and find solutions to emerging questions in Hate Speech Detection[22].

Having participated in a number of these shared tasks myself and contributed to this field in my current doctoral studies[23], I have identified emerging directions the research into Hate Speech Detection can potentially go[24].

This paper proposes a research synopsis of my current doctoral thesis to study systems of Hate Speech Detection employing machine learning algorithms and artificial intelligence.

1.1. What is Hate Speech?

Hate speech has been defined by several people in this field of research[25, 26, 12, 17, 5]. However, they all seem to converge on a common path, which is the characteristics of people who are on the receiving end of hate speech[5, 27, 16]. These include race, gender[25, 28, 29], religion, sexual orientation, ethnicity and nationality[13, 19, 30, 31].

It is important to notice that the understanding and definition of what hate speech is affects the technique and algorithm employed for its detection[25, 5]. This is quite evident in the non-agreeing and non-homogeneity in the annotation and labelling of data perceived to be hate speech[32, 16]. Whether or not speech that merely disparages people or the speech content should contain evidence of violence or insults towards a person or group should be a standard remains quite unstandardized among various research works[33, 28, 32].

There is a general concurrence about the effort to reduce or control hate speech on social media due to its effects[16]. These effects include political insurgence, physical attacks, xenophobia, misogyny and severe riots to the person or group targeted especially women[19, 34, 35]. Also another effect worth mentioning is the mental toll it brings upon human moderators tasked to detect and repress such speech content on Social Media by Non-Governmental Organizations (NGOs) and media platforms[32, 12, 27].

[1] https://commission.europa.eu/strategyandpolicy/policies/justiceandfundamentalrights/combatting-discrimination/racismandxenophobia/combating-hatespeechand-hatecrime_en

[2] https://www.justice.gov/hatecrimes

1.2. Research Objectives

The purpose of this study is to extensively analyze hate speech on social media platforms and additionally improve on the already existing research that propose technical methods of detecting and defusing hate speech on social media platforms using machine learning tools and artificial intelligence.

The main questions I shall be seeking to address with the research are:

1. Does the incorporation of social network analysis provide measurable improvements in Hate Speech Detection?
2. What social network signals are most effective in Hate Speech Detection?
3. How effective have the already developed detection algorithms been and what hinders performances?

Further I shall be seeking to improve performance of hate speech classification models based on the investigations and knowledge build-up on related works.

2. Related work

Early works on Hate Speech Detection started with collection, sampling and annotation of dataset from Social Media platforms such as Twitter[29, 25, 28], Facebook[36]. Also, available are datasets such as twitter corpus in other languages[37, 34, 22, 21]. There have been other works on code switching algorithms also seeking to detect Hate Speech in text written in a mixture of 2 languages such as English-Hindi code switched Hate Speech Detection data[38].

Most of the early research on Hate Speech Detection employed features such as lexical resources[39, 40], sentiment polarity, multi-modal information[41, 42, 43, 40], external information, user information and tweet metadata[26] with mainly logistic regression and Support Vector Machines(SVM) in the models for classification[44, 45, 19]. In the SemEval 2019 task 5 (Hateval)[19], the winning team for English task A had remarkable results with training an SVM model with RBF Kernels, exploiting sentence embedding with Google's Universal Sentence Encoder[46] and beating other models that employed neural network models[19]. In recent works, Transformer based architectures have taken over due to its ability to generate a more semantic and representative embeddings for texts and sentences[3].

There is a striking trend that can be seen in most recent works on Hate Speech Detection research which proves the popularity of Transformers, most significantly BERT[5, 6, 21]. Over 38% deep-learning models in the past five years since the publishing of the BERT paper have in one way or the other employed a variant of this important model[5]. This demonstrates the importance of the model as a key state-of-the-art method in the field of Hate Speech Detection[47, 21]. Studies who compared BERT model to other deep learning models concluded on the superiority of BERT architecture in that BERT-based models achieved top performance in various multilingual Hate Speech Detection tasks[6, 22, 21, 4, 48].

Algorithms leveraging on Artificial Neural Networks (ANNs) without BERT have also been applied in classification algorithms in much earlier benchmarks in a significant fashion[5, 18, 49, 19]. The traditional deep CNN model has also been applied in some studies[50, 51, 19] as well as the Long Short Term Memory (LSTM)[19, 45, 52]. LSTM is employed in word

and character encoding-decoding algorithms that learn words and sentences for input in a classification model[45, 3, 51]. Word and sentence embedding algorithms proficiently the google BERT[3] have been significant in a number of models that encode hate speech data as input data for a neural network employing the use BiLSTM and other transformer models layers as classifiers[6, 5, 21].

Development in hate speech research grows significantly as researchers keep incorporating model structures and architectures proven to be successful in other fields of Artificial Neural Networks and Machine Learning in the classification of Hate Speech[10, 12].Such application which is fundamental is the use of embedding-based transformer models like BERT which may only be required to be pre-trained due to its large hyper-parameter setup to fit current dataset[3, 53]. Application of ensemble based models which can employ a number of separate deep neural networks of different architectures, few shot learners, data augmentation among others has been in spotted in the projects[18, 54, 55, 23, 24]. Ensemble learning has had much success in other works performing arguably better than benchmarks which uses mainstream networks as a classifier development method[18, 56]. Evaluating how a network with layers of pre-trained transformer models applied in an ensemble model was a task I took part in my early papers[24, 23].

In GermEval 2021, I employed the use of similar multilingual transformer-based ensemble architectures using BERT, RNN-based embeddings (BiLSTM) and sentence encoders on three different levels to learn, compare and analyse how models behave on subtasks[24]. In PAN@CLEF Task 3, for the English subtask I fine-tuned a BERT model while for Spanish I used a language-agnostic BERT-based sentence embedding model without fine-tuning[23]. The results I gathered in both experiments [3] showed that although transformer models are key to solving the problem of semantic representation, an increase in the training data has a measurable positive effect on the overall performance across all metrics[23, 24]. Transformer architectures have been key to the development of this field but word and sentence embedding biases generated from the use of pre-trained transformer models can still exist.[57, 58]

In exploring the discussion pertaining to the incorporation of social networks and signals with regards to transfer learning, the study of social network analysis to investigate the graph-like connection behind hate spreaders and their posts at different levels is seemingly growing and I intend to study further [59, 60, 61, 62]. As a general understanding, social network analysis can be modeled as a graph $G = (V, E)$ where V represents people or entities present and E denotes edges[59]. An edge connects two nodes if both the nodes have a social connection like friendship, follow–followee, co-authorship, comments, likes, shares among other social signals[59, 61]. Influence Maximization(IM) is a kind of optimization problem which focuses on the task of identifying a constant number k of seed users with a high spreading capability known as "influential nodes" such that if there is any form of dispatch from them, the information can reach an optimal number of nodes in the network through a cascading effect[60]. Hate spreaders can easily be profiled or tracked through this analysis because of the connection they share[61, 62]. Another connection worth mentioning, is the multi-modal kind which integrates both texts and images embeddings and connects them with a graph-like transformer architecture while exploring similarity and dissimilarity between embeddings for easy detection

[3] https://github.com/kaodamie/PAN-CLEF—Profiling-Hate-Spreaders-on-Twitter

Table 1
Hate Speech Annotated datasets

Name	Year	Source	Size	Graph Data	Reference
multi-Modal Discussion	2023	Reddit	18,000	Yes	[62]
PAN@CLEF-Task 3	2021	Twitter	40,000	No	[21]
GermEval-2021	2021	Facebook	3,000	No	[22]
HS Detection in Multimodal Publications	2020	Twitter	149,823	No	[65]
SemEval-2019 Task 6	2019	Twitter	14,100	No	[19]
Online Hate Speech	2019	Reddit	22,324	No	[66]
Online Hate Speech	2019	Gab	33,776	No	[66]
Peer to Peer Hate	2019	Twitter	27,330	No	[67]
HASOC-2019	2019	Twitter,Facebook	7,005	No	[68]
Language	2017	Twitter	24,802	No	[26]

of hate speech[62]. By leveraging on graph transformers to capture the contextual relationships in the entire discussion that surrounds a comment be it text or images, one looks at the problem holistically and not out-of-context as done by researches in a close related filed like Fake News Detection[62, 63].

3. Data Sources

The primary mode of data collection of Hate Speech data has been from Social media sources where data has been collected and annotated by different studies. These sources include mainly English and on some occasions multi-lingual hate speech data from comments and posts alike including those from popular platforms. Generally, it is to be noted that the majority of these sources are collected by querying social media platforms (Twitter, Facebook, etc) APIs with keywords that do not necessarily connote hate speech[64]. This is because in order to balance the dataset and get them properly annotated for learning, speech that do not contain abusive or offensive terms have to be included. A well balanced dataset selected rather not a heavily biased one is much beneficial to the research of this field.

Table 1 is a summary of key datasets pertaining to Hate Speech Detection published over the years. One of the problems of Hate Speech Detection is its dataset collection and annotation. The method of annotating data for hate speech research uses crowd-sourcing[32, 26, 69] and annotation experts[40] alike. Annotation is indeed a painstaking and expensive process with often contestably erroneous annotations when a crowd-sourcing method is employed[29]. For instance, certain words such as “gay”, “black”, "transgender", "homosexual" and African American English (AAE) receive often times a bias towards being classified as hate speech where in the actual sense they may be resulting as rather a casual use of jargon without necessarily denoting hate or citing violence[32]. This results in lots of false positives in many Hate Speech Detection algorithms which rely on the presence of keywords in speech from a dictionary as a method of classification as well as bidirectional encoder-decoder Transformers because of annotation problems[13, 32]. Transformer models have no strong inductive biases due to its encoder-decoder architecture, so they are flexible and more data intensive models[70]. This

implies that it can find better optimum if enough data is provided with the main drawback being such models perform worse in a low data setting and annotation bias[70, 71].

Even with this risk of annotator-bias at hand, there has been little to no significant research with application of active learning in annotation of hate speech data as it is done for other research areas like image processing[72].

Active learning is a particular sub-field of machine learning which studies how to employ a learning algorithm for annotation of data whereby the learning algorithm iteratively extends the annotation dataset by repetitively asking the user to label some patterns that only appear difficult to annotate[73].

Most recently with the development of Transformer models and Generative AI , the interest of data augmentation is spiking and this is also a direction worth exploring in the study of hate speech data annotation[74]. Transfer learning is a beneficial to Hate Speech research in how researchers keep incorporating model structures and architectures proven to be successful in other fields of AI and Machine Learning in the classification of Hate Speech[5]. One such remarkable application which is indeed promising is Generative Pre-trained Transformers (GPTs), which have been proven be perform remarkably well on data augmentation, machine translation, question-answer tasks, text sumarization in fact-checking tasks, cloze tasks and the like[75][76, 77]. The use of Generative Neural Networks and Transformers by way of data augmentation looks promising on how it is able to generate text data from a few shot examples for training and testing purposes[78, 79], especially with the development of GPT variants[77] over the years, most notably ChatGPT. [4]

4. Conclusion

In this paper I have generally demonstrated the development, motivation, direction of Hate Speech Detection over the years and the benefit of studying social network analysis in the field of hate speech detection, specifically the inclusion social contextualized data in a graph network. Looking beyond the bottleneck of detection classifiers, I have identified a key problem of dataset collection and annotation bias which can potentially be studied. Data augmentation of Hate Speech data is a possibility and can be leveraged upon for many benefits instead of relying solely on sourcing data from social media channels. Studies into how a classifier behaves with augmented data versus actual human annotated hate speech data and various combinations could significantly aid at solving the problems of annotator bias and data hungry algorithms developed with Transformer models.

References

[1] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, D. Klakow, A survey on recent approaches for natural language processing in low-resource scenarios, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Lin-

[4] https://openai.com/blog/chatgpt

guistics, Online, 2021, pp. 2545–2568. URL: https://aclanthology.org/2021.naacl-main.201. doi:10.18653/v1/2021.naacl-main.201.

[2] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable AI for natural language processing, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2020, pp. 447–459. URL: https://aclanthology.org/2020.aacl-main.46.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[4] G. Roccabruna, S. Azzolin, G. Riccardi, Multi-source multi-domain sentiment analysis with BERT-based models, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 581–589. URL: https://aclanthology.org/2022.lrec-1.62.

[5] M. S. Jahan, M. Oussalah, A systematic review of hate speech automatic detection using natural language processing., Neurocomputing (2023) 126232.

[6] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, HateBERT: Retraining BERT for abusive language detection in English, in: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Association for Computational Linguistics, Online, 2021, pp. 17–25. URL: https://aclanthology.org/2021.woah-1.3. doi:10.18653/v1/2021.woah-1.3.

[7] R. Oshikawa, J. Qian, W. Y. Wang, A survey on natural language processing for fake news detection, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6086–6093. URL: https://aclanthology.org/2020.lrec-1.747.

[8] S.-h. Yang, C.-c. Chen, H.-H. Huang, H.-H. Chen, Entity-aware dual co-attention network for fake news detection, in: Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 106–113. URL: https://aclanthology.org/2023.findings-eacl.7.

[9] J. Ramirez-Orta, E. Milios, Unsupervised document summarization using pre-trained sentence embeddings and graph centrality, in: Proceedings of the Second Workshop on Scholarly Document Processing, Association for Computational Linguistics, Online, 2021, pp. 110–115. URL: https://aclanthology.org/2021.sdp-1.14. doi:10.18653/v1/2021.sdp-1.14.

[10] P. Patwa, G. Aguilar, S. Kar, S. Pandey, S. Pykl, B. Gambäck, T. Chakraborty, T. Solorio, A. Das, Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets, arXiv preprint arXiv:2008.04277 (2020).

[11] T. Mickus, K. Van Deemter, M. Constant, D. Paperno, Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 1–14. URL: https://aclanthology.org/2022.semeval-1.1.

doi:10.18653/v1/2022.semeval-1.1.
[12] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10. URL: https://aclanthology.org/W17-1101. doi:10.18653/v1/W17-1101.
[13] W. Warner, J. Hirschberg, Detecting hate speech on the world wide web, in: Proceedings of the second workshop on language in social media, 2012, pp. 19–26.
[14] I. Turner, Criminalising (hateful) extremism in the uk: Critical reflections from free speech, Journal for Deradicalization 34 (2023) 145–175.
[15] N. Peršak, Criminalising hate crime and hate speech at eu level: Extending the list of eurocrimes under article 83(1) tfeu - criminal law forum, SpringerLink (2022). URL: https://link.springer.com/article/10.1007/s10609-022-09440-w.
[16] E. Bleich, The freedom to be racist?: How the United States and Europe struggle to preserve freedom and combat racism, Oxford University Press, 2011.
[17] E. Bleich, The rise of hate speech and hate crime laws in liberal democracies, Journal of Ethnic and Migration Studies (2011).
[18] S. Zimmerman, U. Kruschwitz, C. Fox, Improving hate speech detection with deep learning ensembles, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.
[19] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al., Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 54–63.
[20] S. Manuela, C. Gloria, E. Di Nuovo, S. Frenda, M. A. Stranisci, C. Bosco, C. Tommaso, V. Patti, R. Irene, et al., Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task, in: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), CEUR, 2020, pp. 1–9.
[21] J. Bevendorff, B. Chulvi, G. L. De La Peña Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, et al., Overview of pan 2021: Authorship verification, profiling hate speech spreaders on twitter, and style change detection, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12, Springer, 2021, pp. 419–431.
[22] J. Risch, A. Stoll, L. Wilms, M. Wiegand, Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments, in: Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments co-located with KONVENS, 2021, pp. 1–12.
[23] K. O. Akomeah, U. Kruschwitz, B. Ludwig, University of regensburg@ pan: Profiling hate speech spreaders on twitter., in: CLEF (Working Notes), 2021, pp. 2083–2089.
[24] K. O. Akomeah, U. Kruschwitz, B. Ludwig, Ur@ nlp_a_team@ germeval 2021: Ensemble-based classification of toxic, engaging and fact-claiming comments, in: Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments, 2021, pp. 95–99.

[25] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.
[26] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, arXiv preprint arXiv:1703.04009 (2017).
[27] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.
[28] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, M. Wojatzki, Measuring the reliability of hate speech annotations: The case of the european refugee crisis, arXiv preprint arXiv:1701.08118 (2017).
[29] Z. Waseem, Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter, in: Proceedings of the first workshop on NLP and computational social science, 2016, pp. 138–142.
[30] S. Stecklow, Why facebook is losing the war on hate speech in myanmar, URL: https://www. reuters. com/investigates/special-report/myanmar-facebook-hate (2018).
[31] J. T. Nockleby, Hate speech in context: The case of verbal threats, Buff. L. Rev. 42 (1994) 653.
[32] M. Xia, A. Field, Y. Tsvetkov, Demoting racial bias in hate speech detection, arXiv preprint arXiv:2005.12246 (2020).
[33] B. Kennedy, X. Jin, A. M. Davani, M. Dehghani, X. Ren, Contextualizing hate speech classifiers with post-hoc explanation, arXiv preprint arXiv:2005.02439 (2020).
[34] J. Moon, W. I. Cho, J. Lee, Beep! korean corpus of online news comments for toxic speech detection, arXiv preprint arXiv:2005.12503 (2020).
[35] J. Fortin, Sulli, south korean k-pop star and actress, is found dead, New York Times. October 14 (2019).
[36] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Benchmarking aggression identification in social media, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 1–11.
[37] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, M. Stranisci, An italian twitter corpus of hate speech against immigrants, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.
[38] S. Bansal, V. Garimella, A. Suhane, J. Patro, A. Mukherjee, Code-switching patterns can be an effective route to improve performance of downstream nlp applications: A case study of humour, sarcasm and hate speech detection, arXiv preprint arXiv:2005.02295 (2020).
[39] N. D. Gitari, Z. Zuping, H. Damien, J. Long, A lexicon-based approach for hate speech detection, International Journal of Multimedia and Ubiquitous Engineering 10 (2015) 215–230.
[40] S. S. Tekiroglu, Y.-L. Chung, M. Guerini, Generating counter narratives against online hate speech: Data and strategies, arXiv preprint arXiv:2004.04216 (2020).
[41] P. Burnap, M. L. Williams, Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making, Policy & Internet 7 (2015) 223–242.
[42] P. Burnap, M. L. Williams, Us and them: identifying cyber hate on twitter across multiple protected characteristics, EPJ Data science 5 (2016) 11.

[43] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, S. Mishra, Analyzing labeled cyberbullying incidents on the instagram social network, in: International conference on social informatics, Springer, 2015, pp. 49–66.
[44] M. Karan, J. Šnajder, Cross-domain detection of abusive language online, in: Proceedings of the 2nd workshop on abusive language online (ALW2), 2018, pp. 132–137.
[45] M.-A. Rizoiu, T. Wang, G. Ferraro, H. Suominen, Transfer learning for hate speech detection in social media, arXiv preprint arXiv:1906.03829 (2019).
[46] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al., Universal sentence encoder, arXiv preprint arXiv:1803.11175 (2018).
[47] P. Patwa, G. Aguilar, S. Kar, S. Pandey, S. PYKL, B. Gambäck, T. Chakraborty, T. Solorio, A. Das, SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 774–790. URL: https://aclanthology.org/2020.semeval-1.100. doi:10.18653/v1/2020.semeval-1.100.
[48] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020), in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1425–1447. URL: https://aclanthology.org/2020.semeval-1.188. doi:10.18653/v1/2020.semeval-1.188.
[49] J. H. Park, P. Fung, One-step and two-step classification for abusive language detection on twitter, arXiv preprint arXiv:1706.01206 (2017).
[50] B. Gambäck, U. K. Sikdar, Using convolutional neural networks to classify hate-speech, in: Proceedings of the first workshop on abusive language online, 2017, pp. 85–90.
[51] A. Severyn, A. Moschitti, Twitter sentiment analysis with deep convolutional neural networks, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2015, pp. 959–962.
[52] S. Vosoughi, P. Vijayaraghavan, D. Roy, Tweet2vec: Learning tweet embeddings using character-level cnn-lstm encoder-decoder, in: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, 2016, pp. 1041–1044.
[53] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, arXiv preprint arXiv:1909.11942 (2019).
[54] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for NLP, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 968–988. URL: https://aclanthology.org/2021.findings-acl.84. doi:10.18653/v1/2021.findings-acl.84.
[55] J. Zhou, Y. Zheng, J. Tang, L. Jian, Z. Yang, FlipDA: Effective and robust data augmentation for few-shot learning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 8646–8665. URL: https://aclanthology.org/2022.acl-long.

592. doi:10.18653/v1/2022.acl-long.592.
[56] H. T. Nguyen, M. L. Nguyen, An ensemble method with sentiment features and clustering support, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Asian Federation of Natural Language Processing, Taipei, Taiwan, 2017, pp. 644–653. URL: https://aclanthology.org/I17-1065.
[57] D. Kumar, O. Lesota, G. Zerveas, D. Cohen, C. Eickhoff, M. Schedl, N. Rekabsaz, Parameter-efficient modularised bias mitigation via adapterfusion, arXiv preprint arXiv:2302.06321 (2023).
[58] F.-e. Lagrari, Y. ElKettani, A comparative study of a new customized bert for sentiment analysis, in: Sentiment Analysis and Deep Learning: Proceedings of ICSADL 2022, Springer, 2023, pp. 315–322.
[59] M. E. Newman, D. J. Watts, S. H. Strogatz, Random graph models of social networks, Proceedings of the national academy of sciences 99 (2002) 2566–2572.
[60] S. Banerjee, M. Jenamani, D. K. Pratihar, A survey on influence maximization in a social network, Knowledge and Information Systems 62 (2020) 3417–3455.
[61] S. Kumar, A. Mallik, B. Panda, Influence maximization in social networks using transfer learning via graph-based lstm, Expert Systems with Applications 212 (2023) 118770. URL: https://www.sciencedirect.com/science/article/pii/S0957417422017882. doi:https://doi.org/10.1016/j.eswa.2022.118770.
[62] L. Hebert, G. Sahu, N. Sreenivas, L. Golab, R. Cohen, Multi-modal discussion transformer: Integrating text, images and graph transformers to detect hate speech on social media (2023).
[63] G. Donabauer, U. Kruschwitz, Exploring fake news detection with heterogeneous social media context graphs, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 396–405.
[64] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation (2020) 1–47.
[65] J. A. Gonzalez, L.-F. Hurtado, F. Pla, Twilbert: Pre-trained deep bidirectional transformers for spanish twitter, Neurocomputing 426 (2021) 58–69.
[66] J. Qian, A. Bethke, Y. Liu, E. Belding, W. Y. Wang, A benchmark dataset for learning to intervene in online hate speech, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4755–4764. URL: https://aclanthology.org/D19-1482. doi:10.18653/v1/D19-1482.
[67] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, E. Belding, Peer to peer hate: Hate speech instigators and their targets, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 12, 2018.
[68] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '19, Association for Computing Machinery, New York, NY, USA,

2019, p. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[69] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, Large scale crowdsourcing and characterization of twitter abusive behavior, arXiv preprint arXiv:1802.00393 (2018).

[70] E. Kharitonov, R. Chaabouni, What they do when in doubt: a study of inductive biases in seq2seq learners, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YmA86Zo-P_t.

[71] L. Liu, M. Hulden, Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 739–749. URL: https://aclanthology.org/2022.acl-short.84. doi:10.18653/v1/2022.acl-short.84.

[72] R. Caramalau, B. Bhattarai, D. Stoyanov, T.-K. Kim, Mobyv2al: Self-supervised active learning for image classification, arXiv preprint arXiv:2301.01531 (2023).

[73] B. Settles, Active learning literature survey, Technical Report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[74] H. Dai, Z. Liu, W. Liao, X. Huang, Z. Wu, L. Zhao, W. Liu, N. Liu, S. Li, D. Zhu, et al., Chataug: Leveraging chatgpt for text data augmentation, arXiv preprint arXiv:2302.13007 (2023).

[75] S. Shahriar, K. Hayawi, Let's have a chat! a conversation with chatgpt: Technology, applications, and limitations, arXiv preprint arXiv:2302.13817 (2023).

[76] B. D. Lund, T. Wang, Chatting about chatgpt: how may ai and gpt impact academia and libraries?, Library Hi Tech News (2023).

[77] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.

[78] K.-L. Chiu, A. Collins, R. Alexander, Detecting hate speech with gpt-3, arXiv preprint arXiv:2103.12407 (2021).

[79] C. Casula, S. Tonelli, Generation-based data augmentation for offensive language detection: Is it worth it?, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 3351–3369.