Proceedings of the Fifth International Conference on Trends in Electronics and Informatics (ICOEI).
IEEE Xplore Part Number:CFP21J32-ART; ISBN:978-1-6654-1571-2

Hate Speech Detection Using Natural Language Processing: Applications and Challenges

Anil Singh Parihar[1], Surendrabikram Thapa[1,2], Sushruti Mishra[1,2]

[1]Machine Learning Research Lab, Department of CSE, Delhi Technological University, India
[2]Department of Software Engineering, Delhi Technological University, India
mishrasushruti99@gmail.com

2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) | 978-1-6654-1571-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICOEI51242.2021.9452882

ABSTRACT

The internet has become a common platform for everyone to share their ideas and opinions. The user has freedom to post whatever he/she likes in social networking and blogging sites. However, sometimes the content when directed towards certain group of individuals with an intention to incite hate or discrimination, causes a turmoil in the society. Such content is known as hate speech. Hate speech can be a serious problem to peace and harmony in the society. There are instances where hate speech have led to social unrest and extremism. Thus, hate speech in the internet needs to be monitored. In this paper, we discuss the relevant works done in the field of hate speech detection. Different types of hate speech like racism, sexism, religious hate speech, etc. and the various methods proposed to tackle them are discussed. Further, we identify the challenges and propose the solutions to challenges in hate speech detection in the public internet sphere.

Keywords— Hate Speech, Racism, Sexism, Machine Learning, Deep Learning, Natural Language Processing

1. INTRODUCTION

In the world full of information, trillions of bytes of data are processed every second [1]. Internet has become a common platform to share ideas and opinions. With the growing number of users in social media and blogging platforms, the diversity in content has become huge [41]. This leaves the user with a lot of choices to consume content from. The diversity in content has made people creative and open-minded. Social media has thus become a really powerful medium to share ideas and opinions. Despite all these benefits, there are some dark sides of social media. The kind of content one user may like might not be liked by all other users. With the growing variety of contents, the content relating to sarcasm, jibes and hate speech are increasing day by day. The content that is targeted to a particular race, religion or sexual orientation with an intention of threatening, abusing or provoking is generally called as a hate speech. The hate content in the internet is increasing because of a number of factors. The internet gives an option for users to be anonymous and this leaves users to adopt aggressive behaviors [2]. On the other hand, the people who would otherwise not involve in any discussions are also more likely to join the online discussions. This confidence to express leaves people with even more diverse opinions on same topic. Racism, online abuse, cyber bullying, etc. takes a greater form with prevalence of internet and social network sites. The attack in internet against a certain group of victims might even result into very serious protests and physical destruction of property and lives [3]. So, the content that incites certain group of people to hate the other group is also hate speech and such hate speech coming from influential people in the internet has very bad consequences. Thus, hate speech in social media should be controlled in order to maintain the peace and harmony in a society. This however requires a broad coalition of a lot of parties like government, research organizations, businesses, and so on [4]. The content relating to hate speech can be controlled in multiple ways. The content may be flagged or removed based on the impact that it can create. Manual annotation and removal of the hate speech isn't possible because of the huge amount of data processed every second. For example, as of 2020, there are more than 6000 tweets sent every second [5]. The number is staggering and same is the case with other social network sites like Facebook, reddit, etc. To tackle with huge volume of data in this virtual sphere, we need some intelligent systems that can automatically flag the content using various machine learning models. Machine learning has lately been used in wide variety of fields like intelligent healthcare, smart homes, cybersecurity and many more [6]. One of the most useful applications of machine learning is the management of hateful content. The management of hate speech will make internet more inclusive. This paper will discuss the ways in which machine learning and deep learning are used to control hate speech. Section 2 of the paper describes the related works in this domain. In section 3, the applications that are not covered in section 2 are discussed in brief. Section 4 discusses the challenges and presents possible solutions to tackle with the challenges. Section 5 which is also the conclusion section of the paper concludes the paper with

Authorized licensed use limited to: American University of Beirut. Downloaded on April 22,2022 at 04:18:15 UTC from IEEE Xplore. Restrictions apply.

necessary suggestions and the works that need to be carried out in the future.

2. RELATED WORKS

A lot of works has been done in the field of the detection of hate speech using machine learning models, deep learning architectures, language models, etc. Apart from the diversity in models and architectures, different works have different data which are annotated for different aspects or labels. Similarly, the dataset might be in different languages. Each language has their own lexical, morphological, and syntactic structures. The general methodology for building hate speech detection model is as shown in fig. 1. The process begins with dataset collection and goes through the process of annotation or labelling of the data, extraction of features, use of learning algorithms and evaluation of performance.

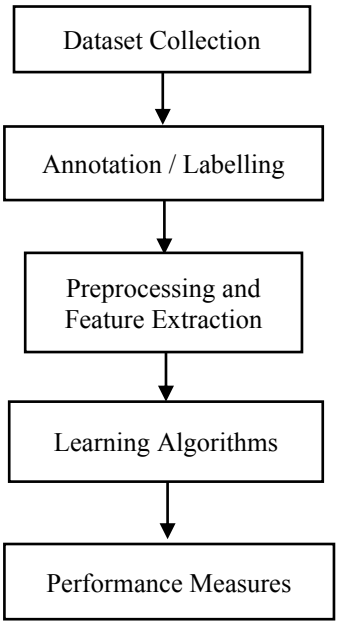

Fig. 1. General structure of hate-speech detection model

First of all, the dataset is prepared for the purpose of training the model. The dataset depends upon our purpose. The language of the data and platform from where the dataset can be extracted is selected first. After collection of the data, the annotations are done by annotators who have some background knowledge about the work that is being carried out. The annotators do the labelling of the dataset based on the criteria decided. The data is labelled into the categories for which models are being built. For example, to build a model that detects racism, the labels of "racist" and "non-racist" should be given. After annotation, the feature is extracted from the data. The text cannot be fed directly to the learning models. So, features like embeddings, linguistic features, etc. are extracted. After extraction of the features, various machine learning models and deep learning architectures are used. The model is then evaluated on the basis of various performance measures like accuracy, precision, recall and f1-score [7].

The past works relating to hate speech detection are discussed below. The works range from detection of general hate speech to detection of hate speech relating to targeted preserved categories.

A. Detection of General Hate Speech

The hate speech dataset in general are collected from tweets and posts from social media sites. The datasets are usually annotated with labels as whether the tweets/posts belong to hate speech or not. The label related to hate speech can further be extended to some specific categories. For instance, Malmasi et al. [8] took tweets annotated with three labels, two labels for hate category and one for non-hate category. 14509 tweets were annotated as hate (2,399), offensive (4,836) and OK (7,274). The character-4-grams feature is taken and then single classifier i.e., SVM classifier is used to train and test the model using 10-fold cross validation. They got a decent accuracy of 78% with the model. Upon further calculation from the confusion matrix, it can be seen that the model has precision of 79.7% and a recall of 77.6%.

Similarly, Vigna et al. [3] proposed a LSTM and SVM based models for the classification of hate speech from Facebook comments. For this purpose, Facebook comments in Italian language were used as the dataset. The total of 17,567 comments were collected and annotated by 5 annotators. Among the total comments, 1,687 of them had annotations from all five annotators whereas the remaining had annotations from one to four annotators. The comments were annotated for no hate, weak hate and strong hate categories. The hate messages were further divided into other sub-categories. The Fleiss' kappa inter-annotator agreement metric for the comments that received annotations from all five annotators over the three major classes (weak-hate, strong hate and non-hate) is 0.19. Such low inter-annotator metric shows that it is really difficult for annotating contents relating to hate as each of the annotators have their own beliefs and the perception of hate is different in all. Despite the work being first of its kind in Italian language, the proposed SVM and LSTM models were able to give an accuracy of 72.95% (hate and non-hate) and 75.23% (hate and non-hate) respectively. Similarly, when all three categories were taken, the accuracy for SVM and LSTM models were 64.61% and 60.50% respectively.

Cao et al. [9] proposed a DeepHate architecture to classify the tweets as normal and inappropriate. The class inappropriate has hate speech relating to racism, sexism and

Authorized licensed use limited to: American University of Beirut. Downloaded on April 22,2022 at 04:18:15 UTC from IEEE Xplore. Restrictions apply.

other different aspects of hate speech. Unlike other methodologies in the literature that rely on single textual features, the proposed work heavily makes use of multi-faceted representations of the text such as word embeddings, sentiments and topical information for detecting the hate speeches in social networking sites. The architecture had really great performance on different standard datasets like WZ-LS [10], DT [11] and FOUNTA [12]. They also made their own dataset, a combined dataset of 114,120 tweets that had two categories viz. inappropriate and normal [9]. For the dataset that the authors had curated, the performance with DeepHate architecture was as high as 92.48 for precision, 92.45 for recall and 92.43 for f1-score.

B. Racism

Park et al. [10] proposed an abusive language detection model which was trained on a corpus of 20K English tweets. The model distinguishes the tweets between sexist tweets, racist tweets and normal tweets. A two-step classification scheme is proposed where the tweets are classified as abusive or not by the model and further the model classifies abusive tweets as sexist tweets or racist tweets. The performance measures of two-step classification scheme are slightly better as compared to one-step multi-class classification problem where classification is done for sexist, racist and non-hate tweets as a multi-class classification problem in a single step. With one-step classification, f1-score of 0.824 was achieved using logistic regression whereas with two-step scheme, f1-score of 0.827 was achieved using HybridCNN. Gamback et al. [13] proposed different CNN based architectures for detection of hate speech relating to four broad categories namely racism, sexism, both (racism and sexism) and non-hate tweets. For CNN architecture based on word2vec embeddings, the f1-score of 0.7829 was achieved. Similarly, Gibert et al. [14] prepared a dataset from white supremacy forum with 1,119 sentences relating to "Hate" category and 8,537 sentences relating to "no hate" category. With LSTM based classifier, the accuracy of 78% was obtained.

C. Sexism

The prejudice or discrimination, usually against women, on the basis of the sex is known as sexism. Most of the societies had been a patriarchal in the past. For that reason, prejudice against women has become a problem. The problem is so widespread that it is not just limited to uneducated societies and rural villages. Big corporates, institutes are often alleged to have done unfair treatments against women. To tackle sexism, there had been a lot of efforts using computational intelligence. Badjatiya et al. [15] used 16K annotated tweets prepared by Waseem and Hovy [16] for detection of sexism. Apart from sexism, they had also used the models for detection of racism in the internet. The dataset had 3383 sexist tweets, 1972 racist tweets and remaining neither sexist nor racist tweets. The tweets were initialized to random vectors and the classification was one using LSTM and Gradient Boosted Decision Trees (GBDTs) with f1-score of 0.930. Pitsilis et al. [17] used an ensemble of RNNs with the same 16K dataset to get an f1-score of 0.932. The work incorporates user's tendency towards posting particular type of contents. The incorporation of user-related information proved to be useful in increasing the performance measures.

D. Prejudice Towards Migrants

Today, due to unrest in different parts of the world, people are bound to leave their homes and families to settle in other country. Apart from the migrations of displaced population due to war or other harsh situations, people also migrate from one country to another in search of better opportunities and career. The migrants are not necessarily always welcomed by the society in which they are migrating. In many cases, the migrants are hated by the native people citing various reasons [18]. The hate speech against migrants is increasingly common in social networking sites and microblogging platforms. Calderon et al. [19] did a detailed analysis of hate speech directed towards immigrants in Spain. The 1977 tweets relating to hate speech were used for the purpose of topic modelling and identifying the distinct linguistic characteristics of hate speech. Similarly, Warner et al. [20] made an effort to tackle with hate speech relating to wide range of prejudices including prejudice against immigrants. The dataset was taken from Yahoo! and American Jewish Congress and the annotations were done for seven labels viz. anti-Semitic, anti-black, anti-Asian, anti-woman, anti-muslim, anti-immigrant or other-hate. The xenophobic speeches in European region and USA were labelled as anti-immigrant. The corpus originally had an inter-annotator agreement metric of 0.63 when done by three annotators. Later, gold corpus was created by correcting the errors in annotations. With the gold positive unigram features, the SVM classifier was able to give an accuracy of 94%. SemEval-2019 Task-5 had the datasets relating to hate speech directed towards women and immigrants [21]. Out of the 19,600 tweets made available, 9091 tweets had the hate speech relating to immigrants and 10,509 tweets relating to women. With SVM based on RBF kernel, f-score of 0.651 was obtained.

E. Religion Hate Speech

In today's modern world, it is generally considered as fundamental right in most of the countries to practice the religion according to one's choice. Religious hate speech can be defined as the speech that is humiliating, offending and insulting a group of people on the basis of their religious

Authorized licensed use limited to: American University of Beirut. Downloaded on April 22,2022 at 04:18:15 UTC from IEEE Xplore. Restrictions apply.

TABLE I. DIFFERENT TYPES OF HATE SPEECH DETECTION, ALGORITHMS AND PERFORMANCE

Authors	Year	Language	Dataset	Use Case	Algorithm	Acc	Pre	Rec	F1-Score
Badjatiya et al. [15]	2017	English	16K tweets	Sexism & Racism	LSTM + GBDTs	-	93.0	93.0	93.0
Pitsilis et al. [17]	2018	English	16K tweets	Sexism & Racism	Ensemble of RNN Classifiers	-	93.05	93.34	93.2
Warner et al. [20]	2012	English	Yahoo + AJC	Different Categories	SVM Classifier	94.0	68.0	60.0	63.0
Cao et al. [9]	2020	English	WZ-LS [10]	Racism & Sexism	DeepHate [9]	-	77.95	79.48	78.19
			DT [11]	Hate and Offensive		-	89.97	90.39	89.92
			FOUNTA [12]	Normal, Abusive, Spam		-	78.95	80.43	79.09
			Combined [9]	General Hate		-	92.48	92.45	92.43
Park et al. [10]	2017	English	WZ-LS [10]	Racism, sexism	HybridCNN	-	82.70	82.70	82.70
Malmasi et al. [8]	2017	English	14509 Tweets	Hate and Offensive	Linear SVM	78.0	79.70	77.60	-
Davidson et al. [11]	2017	English	24802 Tweets	Hate and Offensive	Logistic Regression	-	91.0	90.0	90.0
Zhang et al. [28]	2018	English	DT [11]	General Hate	CNN + skippedCNN	-	-	-	92.0
			WZ [16]	Racism & Sexism	CNN + GRU	-	-	-	83.0
Vigna et al. [3]	2017	Italian	Facebook Comment	General Hate	SVM	64.6	-	-	-
					LSTM	60.5	-	-	-
Albadi et al. [22]	2018	Arabic	Tweets	Religious Hate Speech	GRU based RNN	79.0	76.0	78.0	77.0
Smedt et al. [24]	2018	Multiple	100K Tweets	Religious Hate Speech	SVM Classifier	82.0	82.3	82.26	82.0
Chowdhury et al. [23]	2019	Arabic	Tweets	Religious Hate Speech	LSTM + CNN + NODE2VEC (ARHNet)	79.0	69.0	89.0	78.0
Gibert et al. [14]	2018	English	White Supremacy Forum	Racism	LSTM based Classifier	78.0	-	-	-
Alshalan et al. [29]	2020	Arabic	Tweets	General Hate Speech	CNN	0.83	0.81	0.78	0.79

beliefs. In the Twittersphere or other social networking sites, a lot of content relating to religious hate speech is found. The religious hate speech is not just targeted to people who practice certain religion but also to the group of people who aren't following any religion (atheists). Such religious hate speech can incite discrimination and violence in the society. Albadi et al. [22] did the work of classification of hate speech in Arabic Twittersphere. The total of 6000 tweets belonging to six religious groups (1000 each) were collected in November of 2017 and 600 more tweets with 100 belonging to each religion were collected. Among the tweets belonging to Jews, 60% of them were hate speech. Similarly, for Atheists and Shias, the percentage of hate speech among given tweets are 56% and 50% respectively. Similarly, for

Christians, Sunnis and Muslims, the percentage of hateful content are 36%, 12% and 2% respectively. The total of 2,526 tweets relating to hate speech were prepared with an inter-annotator agreement of 81%. The model which is GRU based

Authorized licensed use limited to: American University of Beirut. Downloaded on April 22,2022 at 04:18:15 UTC from IEEE Xplore. Restrictions apply.

RNN was prepared and the accuracy of 79% was obtained. The performance measures of the work was further improved

by Chowdhury et al. [23] with their LSTM + CNN + NODE2VEC (ARHNet) model. The f-score of 0.78 was achieved as compared to the score of 0.77 by Albadi et al. [22, 23]. Similarly, Smedt et al. [24] proposed a system that does the work of automatic identification of Jihadist Hate speech. The dataset was prepared during the span of October 2014 to December 2016 and the tweets were collected aftermath 10 major terrorist attacks. With 49,311 tweets belonging to hate corpus and 50,166 tweets belonging to safe corpus, the accuracy of 82% was achieved with SVM model. Further in their work, they do a comprehensive qualitative and quantitative analysis of Jihadi Rhetoric.

F. Hate Speech in Regional Languages

Hate speech is a problem not only in microblogging and social networking sites that use English language. With people being able to tweet, post and blog in their own mother tongue and regional languages, the hate speech in regional languages also need to be tackled. Ranasinghe et al. [25] proposed a BERT based architecture for identification of hateful speeches. The social media dataset with multiple languages like English, German and Hindi including code-switched languages were used. The BERT-based architecture was able to achieve weighted F1-scores of 0.8379, 0.7870, 0.8030 for English, German and Hindi languages respectively. Hate Speech Detection (HSD) task at VLSP Workshop 2019 had 25,431 tweets relating to various classes like “hate”, “offensive” and “clean” [26]. With ensemble based Logistic Regression approach, f1-score of 0.678 was achieved. Similarly, Alshalan et al. [27] built a model with tweets relating to hate speech collected from Saudi twitter users. The total of 9316 tweets were annotated and then a CNN based classifier was used. The CNN-based classifier was able to achieve an accuracy of around 83%.

3. FURTHER APPLICATIONS AND DISCUSSION

The above-mentioned works majorly deal in classification of hate speech. There are some of the more important problems like identification of hate instigators and their targets which can help platforms to regulate contents more effectively. ElSherief et al. [30] did a detailed analysis on the personal traits and online behavior of the profiles that instigate the violence/hate and the profiles which are in high target. The research found out that the hate instigators target popular and active profiles so that with the involvement of popular and active users, the hate can be spread to a larger audience. Similarly, the works should also be done in code-mixed or code-switched languages. Chopra et al. [31] suggested a methodology based on profanity modeling, deep graph embeddings, and author profiling to retrieve instances of hateful content in Hindi-English code-switched language (Hinglish) on social media platforms like Twitter. Similarly, Ombui et al. [32] proposed a study for code-switched language that used English, Swahili, and some other native languages. Various machine learning and deep learning models for detecting hate speech in text messages of the code-switched language were used. Out of the eight learning algorithms, SVM was the best performing algorithm with character-level Term Frequency Inverse Document Frequency (TF-IDF) as the feature. The accuracy of 0.825 was achieved with the SVM model. Likewise, Rizwan et al. [33] proposed a CNN-based deep learning architecture for detection of hate speech in Roman Urdu language.

4. CHALLENGES AND SOLUTIONS

With a lot of research happening in the field of hate speech detection, there are challenges impending the research. First of all, there are problems related with the preparation of the dataset [34]. The data collection is not always an easy task because of data usage and distribution policies of different platforms. Some of them have easier policies whereas policies of some platforms make it really hard to collect and use data. Twitter has been a bit lenient in case of the data distribution and usage policy. For this very reason, most of the researchers use twitter as the source of data. The problem doesn’t get solved with twitter providing the data for training models. The data provided by twitter, usually tweets, have character limit of 280. This is an obstruction especially when we want to build models that need to distinguish between longer paragraphs and context rich content. Also, most of the datasets are not well-balanced [35]. It is mostly an observed phenomenon that there are very few tweets relating to hate speech and many more tweets that do not relate to hate speech. For 100 non-hate speech tweets, there are hardly one or two tweets relating to hate speech. The unbalanced datasets can make a model favor one class over the other. For this, data balancing techniques need to be used [36]. Another common problem is that the text data is often not clean and needs a lot of preprocessing. While preprocessing solves most of the problems related with text data, the problems like unwanted items in text that may interfere with the models still exist. Also, there is a trend of using code-mixed language like Hinglish (Hindi+English). Hinglish language has content spoken in Hindi but written in roman script instead of Devanagari script [37]. With code-mixed languages, it is hard to detect hate speech since the grammatical rules in such languages aren’t governed by grammatical rules of a single language [38]. For tackling this, models need to be made robust by training with datasets that are in code-switched languages. Also, there are problems when hate speech is being tackled in regional languages [40]. The unavailability

Authorized licensed use limited to: American University of Beirut. Downloaded on April 22,2022 at 04:18:15 UTC from IEEE Xplore. Restrictions apply.

of proper embeddings, language models and related literature makes it really hard when hate speech detection models are prepared with regional languages. This can be solved by increasing the research for building language models and proper corpus [39]. On a larger context, there is a problem of uniformity in annotation of dataset. The annotations done by one annotator may not match with the annotations done by the other. Vigna et al. [3] had inter-annotator agreement of 0.19 which shows that the annotation task is very hard. A high inter-annotator agreement metric is expected for the dataset. This problem of high variance among the annotators can be tackled by narrowing the criteria set for annotating the text.

5. CONCLUSION AND FUTURE SCOPE

Hate speech detection is a very difficult task and continues to be a societal problem. There is a very fine line between what is a hate speech and what is not. For example, a satire might also be considered as a possible threat but it is not actually a hate speech. The annotation and collection of data for building a model for hate speech detection is thus a very troublesome task. As discussed, this problem can be solved by narrowing down the criteria for annotations. Similarly, there is a need to focus research on code-mixed languages and regional languages as well. Language models and deep learning models have shown promising results in hate speech classifications. For tackling with unbalanced data, the upsampling or downsampling techniques based on language models should be researched upon. The challenges discussed above must be tackled with more research in the domain so that the internet becomes more inclusive, welcoming and free from hate.

REFERENCES

[1] A. Ghimire, S. Thapa, A. K. Jha, S. Adhikari, and A. Kumar, "Accelerating Business Growth with Big Data and Artificial Intelligence," in 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2020: IEEE, pp. 441-448.

[2] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," Policy & Internet, vol. 7, no. 2, pp. 223-242, 2015.

[3] F. Del Vigna12, A. Cimino23, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook," in Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), 2017, pp. 86-95.

[4] J. Banks, "Regulating hate speech online," International Review of Law, Computers & Technology, vol. 24, no. 3, pp. 233-239, 2010.

[5] "The Number of tweets per day in 2020." https://www.dsayce.com/social-media/tweets-day (accessed December 02, 2020).

[6] S. Adhikari, S. Thapa, P. Singh, A. Huo, G. Bharathy, and M. Prasad, "A Comparative Study of Machine Learning and NLP Techniques for Uses of Stop Words by Patients in Diagnosis of Alzheimer's Disease," in 2021 International Joint Conference on Neural Networks (IJCNN), 2021: IEEE, pp. 1-8.

[7] S. Thapa, S. Adhikari, A. Ghimire, and A. Aditya, "Feature Selection Based Twin-Support Vector Machine for the diagnosis of Parkinson's Disease," in 2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC), 2020.

[8] S. Malmasi and M. Zampieri, "Detecting hate speech in social media," arXiv preprint arXiv:1712.06427, 2017.

[9] R. Cao, R. K.-W. Lee, and T.-A. Hoang, "DeepHate: Hate speech detection via multi-faceted text representations," in 12th ACM Conference on Web Science, 2020, pp. 11-20.

[10] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on twitter," arXiv preprint arXiv:1706.01206, 2017.

[11] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proceedings of the International AAAI Conference on Web and Social Media, 2017, vol. 11, no. 1.

[12] A. Founta et al., "Large scale crowdsourcing and characterization of twitter abusive behavior," in Proceedings of the International AAAI Conference on Web and Social Media, 2018, vol. 12, no. 1.

[13] B. Gambäck and 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.

[14] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, "Hate speech dataset from a white supremacy forum," arXiv preprint arXiv:1809.04444, 2018.

[15] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 759-760.

[16] Z. Waseem and 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.

[17] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Effective hate-speech detection in Twitter data using recurrent neural networks," Applied Intelligence, vol. 48, no. 12, pp. 4730-4742, 2018.

[18] J. Bartlett, J. Reffin, N. Rumball, and S. Williamson, "Anti-social media," Demos, no. 2014, pp. 1-51, 2014.

[19] C. A. Calderón, G. de la Vega, and D. B. Herrero, "Topic Modeling and Characterization of Hate Speech against Immigrants on Twitter around the Emergence of a Far-Right Party in Spain," Social Sciences, vol. 9, no. 11, p. 188, 2020.

[20] W. Warner and 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.

[21] V. Basile et al., "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in 13th International Workshop on Semantic Evaluation, 2019: Association for Computational Linguistics, pp. 54-63.

[22] N. Albadi, M. Kurdi, and S. Mishra, "Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere," in 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018: IEEE, pp. 69-76.

[23] A. G. Chowdhury, A. Didolkar, R. Sawhney, and R. Shah, "ARHNet-Leveraging Community Interaction for Detection of Religious Hate Speech in Arabic," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2019, pp. 273-280.

[24] T. De Smedt, G. De Pauw, and P. Van Ostaeyen, "Automatic detection of online jihadist hate speech," arXiv preprint arXiv:1803.04596, 2018.

[25] T. Ranasinghe, M. Zampieri, and H. Hettiarachchi, "BRUMS at HASOC 2019: Deep Learning Models for Multilingual Hate Speech and Offensive Language Identification," in FIRE (Working Notes), 2019, pp. 199-207.

Authorized licensed use limited to: American University of Beirut. Downloaded on April 22,2022 at 04:18:15 UTC from IEEE Xplore. Restrictions apply.

[26] X.-S. Vu, T. Vu, M.-V. Tran, T. Le-Cong, and H. Nguyen, "HSD shared task in VLSP campaign 2019: Hate speech detection for social good," arXiv preprint arXiv:2007.06493, 2020.

[27] R. Alshaalan and H. Al-Khalifa, "Hate Speech Detection in Saudi Twittersphere: A Deep Learning Approach," in Proceedings of the Fifth Arabic Natural Language Processing Workshop, 2020, pp. 12-23.

[28] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? the challenging case of long tail on twitter," Semantic Web, vol. 10, no. 5, pp. 925-945, 2019.

[29] R. Alshalan and H. Al-Khalifa, "A Deep Learning Approach for Automatic Hate Speech Detection in the Saudi Twittersphere," Applied Sciences, vol. 10, no. 23, p. 8614, 2020.

[30] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding, "Peer to peer hate: Hate speech instigators and their targets," in Proceedings of the International AAAI Conference on Web and Social Media, 2018, vol. 12, no. 1.

[31] S. Chopra, R. Sawhney, P. Mathur, and R. R. Shah, "Hindi-English Hate Speech Detection: Author Profiling, Debiasing, and Practical Perspectives," in Proceedings of the AAAI Conference on Artificial Intelligence, 2020, vol. 34, no. 01, pp. 386-393.

[32] E. Ombui, L. Muchemi, and P. Wagacha, "Hate Speech Detection in Code-switched Text Messages," in 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2019: IEEE, pp. 1-6.

[33] H. Rizwan, M. H. Shakeel, and A. Karim, "Hate-speech and offensive language detection in roman Urdu," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 2512-2522.

[34] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in social networks: a survey on multilingual corpus," in 6th International Conference on Computer Science and Information Technology, 2019.

[35] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions," Computer Science Review, vol. 38, p. 100311, 2020.

[36] S. Thapa, P. Singh, D. K. Jain, N. Bharill, A. Gupta, and M. Prasad, "Data-driven approach based on feature selection technique for early diagnosis of Alzheimer's disease," in 2020 International Joint Conference on Neural Networks (IJCNN), 2020: IEEE, pp. 1-8.

[37] G. Sreeram and R. Sinha, "Language modeling for code-switched data: Challenges and approaches," arXiv preprint arXiv:1711.03541, 2017.

[38] P. Mathur, R. Shah, R. Sawhney, and D. Mahata, "Detecting offensive tweets in hindi-english code-switched language," in Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, 2018, pp. 18-26.

[39] S. Thapa, S. Adhikari, and S. Mishra, "Review of Text Summarization in Indian Regional Languages," in 2020 International Conference on Computing Informatics & Networks (ICCIN), 2020, pp. 23-32.

[40] S. Thapa, S. Adhikari, U. Naseem, P. Singh, G. Bharathy, and M. Prasad, "Detecting Alzheimer's Disease by Exploiting Linguistic Information from Nepali Transcript," in International Conference on Neural Information Processing, 2020: Springer, pp. 176-184.

[41] A. Ghimire, S. Thapa, A. K. Jha, A. Kumar, A. Kumar, and S. Adhikari, "AI and IoT Solutions for Tackling COVID-19 Pandemic", in 2020 International Conference on Electronics, Communication and Aerospace Technology, 2020: IEEE.

Authorized licensed use limited to: American University of Beirut. Downloaded on April 22,2022 at 04:18:15 UTC from IEEE Xplore. Restrictions apply.