A DEEP LEARNING FRAMEWORK FOR AUTO DETECTION OF HATE SPEECH ON SOCIAL MEDIA

[1] Varun Gupta, [2] Dr Umesh Sehgal

Research Scholar, Sant Baba Bhag Singh University
Associate Professor, Sant Baba Bhag Singh University
jsrtechvarun@gmail.com,umeshsehgalind@gmail.com

ABSTRACT:

Hate speech on social media may spread quickly through online users and subsequently, may even escalate into local vile violence and heinous crimes. This paper proposes a hate speech detection model by means of machine learning and text mining feature extraction techniques. In this study, the authors collected the hate speech of English-Odia code mixed data from a Face book public page and manually organized them into three classes. In order to build binary and ternary datasets, the data are further converted into binary classes. The modeling of hate speech employs the combination of a machine learning algorithm and features extraction. Support vector machine (SVM), naïve Bayes (NB) and random forest (RF) models were trained using the whole dataset, with the extracted feature based on word unigram, bigram, trigram, combined n-grams, term frequency-inverse document frequency (TF-IDF), combined n-grams weighted by TF-IDF and word2vec for both the datasets. Using the two datasets, we developed two kinds of models with each feature—binary models and ternary models.

KEYWORDS:

Hate speech; social media; English-Odia; machine learning; feature extraction; TF-IDF

1. INTRODUCTION

Social media is changing the face of communication and culture of societies around the world [1]. Numbers of social media users in India have grown substantially in recent years, despite the low quality of internet services and the occasional interruptions or blocking of social media sites in the country. Multifarious populations in the country have been using online social media to communicate, express opinions, engage with friends, and share information [2,3,4]. However, the anonymity and mobility of online social media enable the netizens behind the screen to easily spread hateful content [5,6]. There are now 4.48 billion social media users around the world, which is equal to almost 57 percent of the world’s total population.

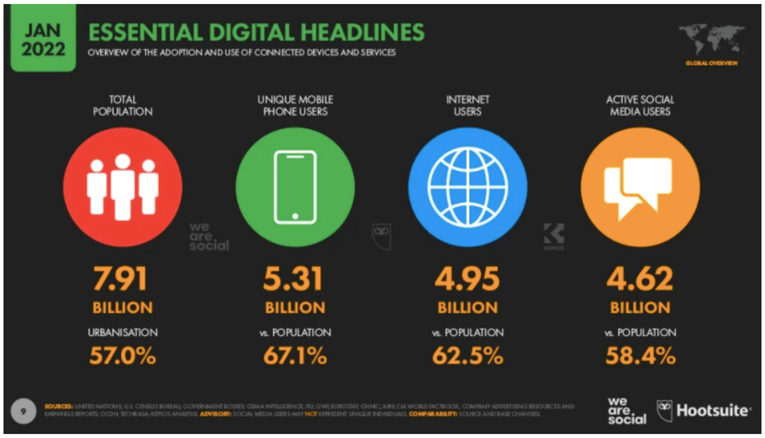

Figure 1: Number of active users on social media around the world.

Social media has lot of advantages & disadvantages like it helps in sharing messages, posts, text, video with our friends and family and with the world but it also has some serious dis-advantage like hate speech. As social media is free platform with not much control over content shared by others. Social media sites are the primary media for perpetrating hate speeches nowadays [3]. Even some political and racist organisations have exploited social media platforms to propagate toxic contact such as hate speeches to spread anger or fear among some societies and can may trigger violence and raise public safety concerns [4].

2. RELATED WORK

This section presents a comprehensive review of the general techniques, methods, and results of existing research about automatic hate speech detection on social media. Mossie and Wang [22] investigated hate speech detection for the Amharic language. They created a dataset of 6120 instances of Amharic posts from Facebook, and classified the speech as “hate” and “not hate” using word2vec and term frequency-inverse document frequency (TF-IDF) feature extraction. They used the machine learning classifier algorithms, naïve Bayes (NB) and random forest (RF), to detect the features of “hate” and “not hate” speech. The NB model achieved 73.02% and 79.83% accuracy, while the RF model achieved 63.55 and 65.34% accuracy, respectively, for both of the features. The authors conclude that the result is promising for computing a large volume of data for a social network. Ibrohim et al. [23] studied hate speech for the Indonesian language on social media. The authors collected tweets and created a binary class dataset comprising HS and Non-HS (NHS), and classified them using a different combination of feature and machine learning classifier algorithms, which included a BOW model, word n-gram, character n-gram and negative sentiment with NB, support vector machine (SVM), beacon-less routing (BLR), and random forest decision tree (RFDT). They achieved a 93.5% F-measure; the best performance with the combination of word n-gram with RFDT than other combined models. For the problem of differentiating hate from offensive speech, Davidson et al. [24] studied the characterization of hate for other instances of speech, like offensive speech, for automatic hate speech detection using 33,458 English tweets. They used hate speech lexicon from hatebase.org to label the hate speech dataset into three categories: hate, offensive, and neither. The authors then employed bigram, unigram, and trigram features with TF-IDF, and used part-of-speech, sentiment lexicon for social media. Logistic regression with Linear SVM yielded an overall precision of 0.91, recall of 0.90 and F1 score of 0.90. They concluded that high accuracy detection can be achieved by differentiating between these two classes of speech. Gambäck et al.

[25] presented a deep-learning-based hate speech text classification system for Twitter. They used a dataset prepared by Benikova et al. [26], which was comprised of four categories: racism, sexism, both (racism and sexism), and NHS. They used four features for embedding, namely word2vector, random vector, character n-grams, and word vectors, combined with the deep learning of convolutional neural network (CNN). The model that was based on word2vec embedding turned out to be the best, with a 78.3% F-score.

Del Vigna et al. [27] studied an Italian online hate campaign on social network sites using the textual content of comments that appeared on a public Italian Facebook page as a source. The datasets are labeled as no hate, weak hate, and strong hate, and by merging weak and strong hate together as hate, they formed the second dataset. By leveraging morpho-syntactical features, sentiment polarity and word embedding lexicons, the authors designed and implemented two classifier algorithms for the Italian language: one is the traditional machine learning algorithm named SVM and the other is the deep learning recurrent neural network (RNN) named the long short-term memory (LSTM) algorithm. By conducting two different experiments with both datasets, in at least 70% of cases the annotator agreed on the class of the data. SVM and LSTM achieved an F-score of 80% and 79% for binary classification and 64% and 60% for ternary classification, respectively. Another study on Italian tweets-TWITA was reported by Florio et al. [6]. They used SVM and AIBERTo, the Italian BERT language model, and revealed the importance of the time difference between training and test data, because this will impact the performance of both the SVM and AIBERTo models. Another development pertaining to Italian tweets is the creation of a lexicon of hate words, known as Hurtlex, which can be used as a resource to identify hate speech

3 ADVANTAGES OF SOCIAL MEDIA:

1. **Communication:** Social media networks are the fastest means of communication as messages are sent and received almost instantaneously [41].
2. **Connectivity:** This is one of the important advantage of social media i.e. connectivity. Any number of users can connect from any place at any time with the use of internet only. Through social media, the information like images, text or video can be shared across the world, and building relationships with each other also become easy. Social media provides a feeling of closeness between friends and family.
3. **Education:** Social media is not only helpful in making relations only but also has been proved beneficial in the field of education. In this time of pandemic social media makes learning easier by connecting educators and experts all over the world with the learners. Any user can start discussion on it which helps in improving skills by enhancing knowledge and creativity.
4. **Information and Updates**: Social media also remains helpful in keeping yourself up to date, earlier people depends upon newspaper, magazines, TV for information but now these mediums of information too depend upon social media. you could keep yourself updated with the information about any happenings in the world or in someone’s life. Social media helps you to provide correct information by showing the true picture of contents and resources. It helps in showcasing the real-world globally [2].
5. **Awareness:** Social media also helps in creating and spreading awareness in the minds of people. For example, you are organising any social activity and wants to spread among

other people of your specific area only or with the people who are interests in social activates then social media is best for you

6. **Share Anything with Others**: Social media sites are the best platform to post anything you want to like a lyrics of song, a video, a poem, a creation, a recipe, a painting, and much more. It not only provides happiness amount creators but also provide a platform to showcase your hidden talent. In recent times, a lot of people from small areas, poor artists become famous by posting their songs or dance on social media [5].
7. **Noble Cause**: Social media is not only for entertainment but it can be useful for noble deeds also. For example, many NGO's or patients suffering from diseases like cancer or thalassemia and are in need of funds to cure it, they contact people on social media for the same. It is the easiest and quickest to promote a noble cause. Not only for funds but jobs also can be applied through the social media.
8. **Mental Health:** In today's time when everyone has hectic and tight schedules at that point of timesocial media also acts as a great stress buster or mental health reliever by connecting to various people across the world and building positive relationships with them. Some people have created special group/ pages and communities for this purpose only they help you to fight with stress issues, depression, and isolation. It can build healthy relations with people by generating positive vibes and a happier mood.
9. **Companies:** Social media is not only good for people but its also best for companies to share their products, offers with their any number users & links of them. The ability to reach large audiences is huge advantage of social media. It opens the door for any business to find more business [1].
10. **Free:** This is one of the biggest advantage of social media that it is entirely free to join. Most of the large social media platforms are offering free membership with most of the features too are free and for few features they are charging some amount also.

3.1 DIS-ADVANTAGES OF SOCIAL MEDIA SITES:

1. **Affects Social-Emotional Connection**: Social media also has some dis-advantages like it become a hindrance in the way of social-emotional connection like in earlier times at the time of special days like birthday, anniversary, promotion we use to call people and congratulate them but now days everything has been limited to textual content through social media, which results in a lack of personal feelings and connections [1].
2. **Thinking ability:** As we all knows that social media has decreased real-time face-to-face conversations with our friends and family. We now days relying on text messages by simply typing a text or most of the time we copy and paste from Goole or other search engines. We can say that Internet users are not quick-witted; they take time to think and then reply.
3. **Present Physically Not Mentally**: Earlier all family sit together, play, eat and laugh together but now when family sits together but still everyone was busy on their mobile phones only. People recommend to stay connected with their virtual friends beside physical friends. For people it matters most that how much likes are coming on FB post or on Instagram post besides what family saying.It is one of the major reasons behind health issues like depression, stress, and anxiety because we are somewhere missing those real-time friends and interactions with them, which we earlier used to have.
4. **Cyberbullying:** At the beginning pf the 20th century, cyberbullying was not treated seriously when social media was still in it infancy [5]. In 2017, 41% of the US citizens personally encountered online harassment. In current scenario this, most of the social

media users, especially children, have become victims of cyberbullying as it is very easy nowadays we have see thatcreating fake accounts and fake profiles and threaten the other person or asking for money is becoming very normal. Cyberbullying can result into even suicides, depression issues, etc. Social media sites are filled full of fake news and hatred rumours, which has caused an unhealthy environment in society and the country.

5. **Addiction**: This is one the main challenge for parents and individuals that persons specially youth becoming addict to Social media. Most of the social media users spending their whole day and even nightson browsing social media sites. This addiction has ruined their lives leading to serious issues. Use of everything beyond the limit results into wasting their productive time and energy both[1].
6. **Social Media data control Issues**: Social media provides information, shares things with others but sometimes, privacy also got compromised on it. Most people used social media and daily updating their details like phone number, address, work location, and other information of their personal life. Teenagers girls or boys share their photos over their and some mischievous persons use those photos with wrong headings or edit that photos to make them feel insulted. Boys stole passwords of girls from their accounts. One of the most common and security issues found on using social media platforms is stolen the account's password. Anyone can create account on anyone name and post their pictures also [2].
7. **Hate Speech**: On social media sites people posting comments or sharing posts to hurt the other persons or religion. The unwanted trolls, images, comments, videos on another person's life or on religion or on gender makes viewers uncomfortable.Sometime this type of hate speech results into disastrous like riots, suicides, war etc.

4 ABOUT HATE SPEECH

Hate speech is not a new term for the world. However social media have begun playing larger role in hate crimes. A lot of researchers found that the persons involved in hate related terror attacks had an extensive social media history of hate related posts. In some cases, social media can play more direct role in hate crime for instance video footage from the suspect of the 2019 terror attack in Christchurch, New Zealand, was broadcast live on Facebook [6].

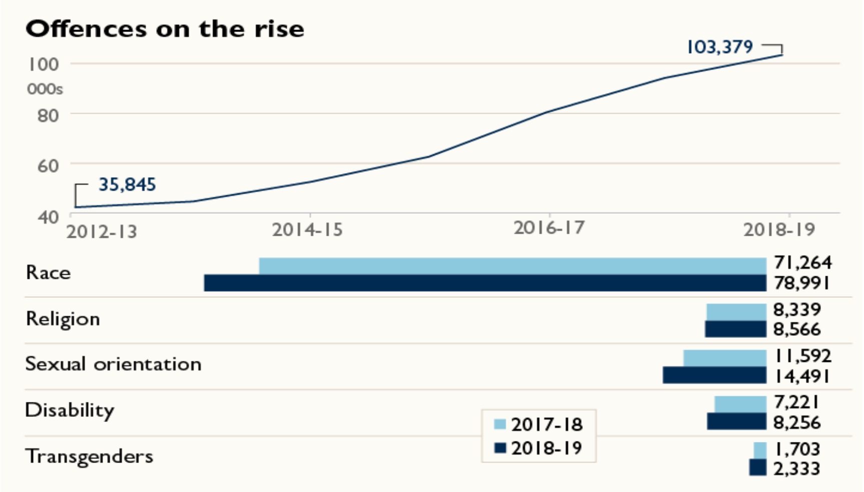

Figure 2: Increase in Hate speech on social media from 2013-19

Hate speech has an adverse effect on the mental health of persons and directly on society. Twitter have seen about 900% increase in hate speech during COVID-19 pandemic. YouTube reported about 500 Million hate comments from June 2019 to September 2019 [7].

According to a survey, 47.3% of students have been a victim of hate speech in Asian countries. Victim report anxiety, depression, fear, self-harming and mental health issues as after-effects. Cyberbullying stats 2020 shows that 42% of online harassment happens on Instagram which has over a billion active users. Facebook and snapchat follows closely, with 39% and 31% respectively. Stats have shown that Cyberbullying victims are 2.9 times more likely to commit suicide [8].

In the European (EU), 80% of people have encountered hate speech online and 40% of felt attacked or threatened via social media sites [9]. These hateful posts and comments not only effect the society at a micro scale but also at a global stage by influencing people's views regarding important events like elections and protests [18].

4.1 HATE SPEECH DEFINITION

Hate speech doesn't have any universal definition or you can say that all are not agreed upon one definition of hate speech as the line between hate speech and appropriate free expression is very thin, making some wary to give hate speech a precise definition [6].A common definition of hate speech is communication towards a specific person or group with aggressive content based on some characteristics such as gender, race, ethnicity, sexual orientation, religion, color or nationality [10]. Each platform like American Bar Association, Facebook, Twitter and others have their own definitions. Few definitions I am going to mentioned below:

A. **Encyclopaedia of the American Constitution**: "Hate speech is speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, dis-ability, sexual orientation or gender identity" [11].

B. **Facebook:** "We define hate speech as direct attack on people based on what we call protected characteristics – race, Ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We also provide some protection for immigration status. We define attach as violent or dehumanizing speech, statements of inferiority or calls for exclusion or segregation" [12].

C. **Twitter:** "Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease" [13].

D. **Davidson et al**.: "Language that is used to express hatred towards a targeted group or is intended to be a derogatory, to humiliate, or to insult the member of the group" [14].

E. **YouTube**: "Hate speech refers to content that promotes violence or hatred against individuals or groups based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status and sexual orientation/ gender identity. There is fine line between what is and what is not considered to be hate speech. For instance, it is generally okay to criticize a nation-state, but not okay to post malicious hateful comments about a group of people solely based on their ethnicity" [17].

4.2 STEPS TO STOP HATE SPEECH

The problem of hate speech getting increased popularity, therefore many initiatives are also taken at Government levels e.g.: The Council of Europe executed the movement of No Hate Speech, Legislation has also been made to eliminate its proliferation, names EU Hate speech code of conduct [33]. The Indian Government has already introduced law that expands the anti-terrorism law to encompass cyberspace in order to prohibit the dissemination of any terrorizing of obscene information [34]. The German federal Government has introduced the Act to Improve Enforcement of the Law in Social Networks (Network enforcement Act) in 2017 [35].

However, policies or compliance rules are difficult to enforce, if hate speech cannot be detected efficiently.

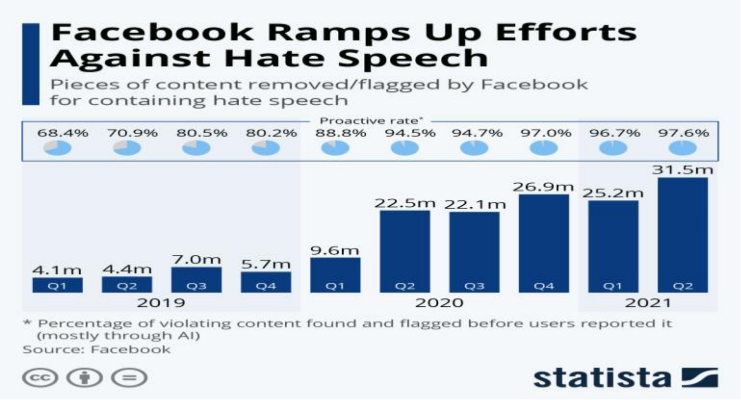

Figure 3: Percentage of Hate speech flagged by users on Facebook

Earlier manual process was adopted to detect hate speech on social media platforms [36]. They hire persons at positions like Community managers are thus in charge of moderating user contributions, by employing various strategies for supervising, controlling, and enabling content submission. When pre-moderation is followed, high security is achieved. However, this method requires a lot of effort, finance and time resources. On the other hand, post moderation policies lead to a simpler and more open approach but lower the quality [37].

To save human resources and time, a lot of automatic hate speech detection algorithms are available in the field using machine language but they have their own issues. As the language use one the web is in a different text style as compared to the day-to-day speech [38]. Another problem with machine learning programs are like some communities tend to use benign words or phrases that have accepted hate speech meaning within their community and specific social context of usage [39]. In simple words, meaning of one word for two communities can be differ.

4.3 DEEP LEARNING

Today, automatic hate speech detection is often based on machine learning approaches. However, while, deep learning models achieve a high performance [35]. Deep learning is the part of machine learning which depends entirely on deep artificial neural networks that learn and identify patterns by mimicking the event in layers of neurons. There are two perspectives in Deep learning. First, the right representation of data. Secondly, the depth of the neural networks is an important factor since the greater depth will give more effective power, because the complicated tasks will be broken into a series of layers [40].

4.3.1 DEEP LEARNING MODELS:

Deep learning techniques can learn automatically from the data following a supervised strategy. Labelled training data need to be provided as an input. These models are competent to understand and analyze text using deep artificial lneural networks with multiple stacked layers [42]. However, two of the most popular examples of deep artificial neural networks are:
CNN: Convolutional Neural Networks.
RNN: Recurrent Neural Networks.
CNNisconsideredasaneffectivenetworkforextract-ingfeaturesfromthedata.Ontheotherhand,RNNis more suitable for modeling orderly sequence tasks [40].In this research, we are going to experiment the effectiveness of using different deep learning settings using RNN alone and using a combination of RNN and CNN. We assume that combining the two architectures will show a better performance as they will be able capture more hate-speech patterns. However, RNN is a family of different architectures with different gating mechanisms ,which includes the following:

- LTSM(Long Short-TermMemorynetwork)—which is capableoflearninglong-termdependenciesbetween wordsbyrememberingwordsforlongperiodoftime.
- GRU(GatedRecurrentUnit)—whichisavariantof LTSMbutGRUissimplerandfasterinthetraining process,wheretheLTSMismorepowerfulandcomplex thanGRU.

Wecanhavedifferentsettingsofourdeepneuralnet-worksarchitecturebyadjustingthelayersofneuralnet-works andfine tuningtheparameters untilthey satisfy our Arabichate-speechproblem.Threemainsettingsofdeep neuralnetworkswillbeexperimented:

- LTSMmodel.
- EnsemblemodelofLTSMandlayerofCNN.
- GRU model.
- EnsemblemodelofGRU andalayerofCNN.

Inordertobuildandexperimentthese4models,we aregoingtouseKeras,whichisadeeplearninglibraryin PythonthatworksontopofTensorFlow.

5. PROPOSED RESEARCH METHODOLOGY

A term “Research Methodology” is something every researcher around the world must have worked upon to get the true understanding of the procedures and protocols through which his / her research will pass. In general terms, Research means search of knowledge, knowing the known in a more scientific and organized manner. Redman and Mory defines research as a “systematized effort to gain new knowledge.” The Advanced Learner’s Dictionary of Current English says that research is “a careful investigation or inquiry especially through search for new facts in any branch of knowledge.” Research is, for the purpose of advancement a novel contribution to the existing reserve of knowledge. It is search for knowledge via goal oriented objective, systematic & scientific methods of finding solution or a new dimension to a subject. The main purpose of research is to explore the unknown dimensions, the unknown facts about the subject.

5.1 NEED AND SIGNIFICANCE OF THE STUDY

When we talk about hate speech detection on social media, then we can see a lot of work has already been done but as per the research all work is not able to provide even 90% successful results whether using machine learning or manual work or using deep learning existing techniques [42]. Hate speech is a complicated & multi-faceted concept that has been difficult to understand, by both human beings and computer systems [3].

5.2 OBJECTIVES OF THE STUDY

- To study the existing deep learning techniques for auto detection of hate speech & record the results
- To propose a new deep learning framework for auto detection of hate speech on social media.
- To categorize detected hate speech using proposed deep learning framework.
- To compare the results of proposed deep learning framework with the results of existing deep learning techniques

5.3 METHODOLOGY

My approach is to use deep neural network to identify and classify hate speech into different subclasses. Methodology divides into five stages.

1. Sample data will be collected from different sources like API's provided by different social media sites and data available on websites.
2. Clean the data from errors and noise, normalized, duplicates were removed
3. Study existing frameworks and their functionality, test available data on them and record their results.
4. Develop framework for auto detection of hate speech.
5. Test framework on available date and compare results with earlier available results.

6. CONCLUSION:

All social media is filled with hate speech and increasing day by day. Hate speech on social media has increases many times in recent years. This increase has encouraged researches to work on automated techniques to detect and categorize hate speech. This paper addresses the hate speech detection problem with three tasks: First classification of content into (Hate or normal), second, categorization of content into (Hate, abusive or normal), and lastly, multi-class categorization of content into (Racism, Religious, color, gender). We have collected data from internet and other social media sites for testing purpose. In the future, we will continue increasing the size of data for testing purpose. We also plan to research more and more on deep learning models. Finally, we will research more on targets and sources of hate speech, for instance, researching on the gender, the color, or the country.

7. REFERENCES

1. Duwairi, R., Hayajneh, A., &Quwaider, M. (2021). A Deep Learning Framework for Automatic Detection of Hate Speech Embedded in Arabic Tweets. *Arabian Journal for Science and Engineering*, *46*(4), 4001-4014.
2. Van der Bank, C. M., & van der Bank, M. (2014). The impact of social media: advantages or disadvantages. *African Journal of Hospitality, Tourism and Leisure*, *4*(2), 1-9.
3. 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*.

4. Beddiar, D. R., Jahan, M. S., &Oussalah, M. (2021). Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, *24*, 100153.
5. Alotaibi, M., Alotaibi, B., &Razaque, A. (2021). A multichannel deep learning framework for cyberbullying detection on social media. *Electronics*, *10*(21), 2664.
6. MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS one*, *14*(8), e0221152.
7. Srivastava, A., Hasan, M., Yagnik, B., Walambe, R., & Kotecha, K. (2021). Role of Artificial Intelligence in Detection of Hateful Speech for Hinglish Data on Social Media. *arXiv preprint arXiv:2105.04913*.
8. Raja, R., Srivastavab, S., &Saumyac, S. (2021). NSIT & IIITDWD@ HASOC 2020: Deep learning model for hate-speech identification in Indo-European languages.
9. Castano-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, 101608.
10. Aljero, M. K. A., &Dimililer, N. (2021). Genetic Programming Approach to Detect Hate Speech in Social Media. *Ieee Access*, *9*, 115115-115125.
11. Nockleby JT. Hate speech. Encyclopaedia of the American constitution 2000; 3: 1277-79
12. Community Standards: Available from: https://m.facebook.com/communitystandards/objectionable_content
13. Hateful conduct policy: Available from: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
14. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, No. 1).
15. de Gibert, O., Perez, N., García-Pablos, A., &Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.
16. Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), 1-30.
17. Hate Speech Policy: Available from: https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436
18. Mishra, S., Prasad, S., & Mishra, S. (2021). Exploring multi-task multi-lingual learning of transformer models for hate speech and offensive speech identification in social media. *SN Computer Science*, *2*(2), 1-19.
19. Albadi, N., Kurdi, M., & Mishra, S. (2018, August). Are they our brothers? analysis and detection of religious hate speech in the arabictwittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 69-76). IEEE.
20. Horsti, K. (2017). Digital Islamophobia: The Swedish woman as a figure of pure and dangerous whiteness. *New Media & Society*, *19*(9), 1440-1457.
21. Froio, C. (2018). Race, religion, or culture? Framing Islam between racism and neo-racism in the online network of the French far right. *Perspectives on Politics*, *16*(3), 696-709.
22. Hanzelka, J., & Schmidt, I. (2017). Dynamics of Cyber Hate in Social Media: A Comparative Analysis of Anti-Muslim Movements in the Czech Republic and Germany. *International Journal of Cyber Criminology*, *11*(1).

23. Faulkner, N., &Bliuc, A. M. (2018). Breaking Down the Language of Online Racism: A Comparison of the Psychological Dimensions of Communication in Racist, Anti-Racist, and Non-Activist Groups. *Analyses of Social Issues and Public Policy*, *18*(1), 307-322.
24. Chen, P. J. (2019). Civic discourse on Facebook during the Australian same-sex marriage postal plebiscite. *Australian Journal of Social Issues*, *54*(3), 285-304.

25. Evolvi, G. (2019). # Islamexit: inter-group antagonism on Twitter. *Information, communication & society*, *22*(3), 386-401.
26. Sainudiin, R., Yogeeswaran, K., Nash, K., &Sahioun, R. (2019). Characterizing the Twitter network of prominent politicians and SPLC-defined hate groups in the 2016 US presidential election. *Social network analysis and mining*, *9*(1), 1-15.
27. Dragiewicz, M., Burgess, J., Matamoros-Fernández, A., Salter, M., Suzor, N. P., Woodlock, D., & Harris, B. (2018). Technology facilitated coercive control: Domestic violence and the competing roles of digital media platforms. *Feminist Media Studies*, *18*(4), 609-625.
28. KhosraviNik, M., & Esposito, E. (2018). Online hate, digital discourse and critique: Exploring digitally-mediated discursive practices of gender-based hostility. *Lodz Papers in Pragmatics*, *14*(1), 45-68.
29. Miro-Llinares, F., & Rodriguez-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.
30. Poole, E. A., Giraud, E., & de Quincey, E. (2019). Contesting# stopIslam: the dynamics of a counter-narrative against right-wing populism. *Open Library of Humanities*, *5*(1).
31. Sundén, J., &Paasonen, S. (2018). Shameless hags and tolerance whores: Feminist resistance and the affective circuits of online hate. *Feminist Media Studies*, *18*(4), 643-656.
32. Trajkova, Z., &Neshkovska, S. (2018). Online hate propaganda during election period: The case of Macedonia. *Lodz Papers in Pragmatics*, *14*(2), 309-334.
33. Qureshi, K. A., &Sabih, M. (2021). Un-compromised credibility: Social media based multi-class hate speech classification for text. *IEEE Access*, *9*, 109465-109477.
34. Mohapatra, S. K., Prasad, S., Bebarta, D. K., Das, T. K., Srinivasan, K., & Hu, Y. C. (2021). Automatic Hate Speech Detection in English-Odia Code Mixed Social Media Data Using Machine Learning Techniques. *Applied Sciences*, *11*(18), 8575.
35. Bunde, E. (2021, January). 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* (p. 1264).
36. Hille, S., & Bakker, P. (2014). Engaging the social news user: Comments on news sites and Facebook. *Journalism Practice*, *8*(5), 563-572.
37. Vrysis, L., Vryzas, N., Kotsakis, R., Saridou, T., Matsiola, M., Veglis, A., ... &Dimoulas, C. (2021). A Web Interface for Analyzing Hate Speech. *Future Internet*, *13*(3), 80.
38. Ghosh Roy, S., Narayan, U., Raha, T., Abid, Z., & Varma, V. (2021). Leveraging multilingual transformers for hate speech detection. *arXiv e-prints*, arXiv-2101.
39. Vijayaraghavan, P., Larochelle, H., & Roy, D. (2021). Interpretable multi-modal hate speech detection. *arXiv preprint arXiv:2103.01616*.

40. Al-Hassan, A., & Al-Dossari, H. (2021). Detection of hate speech in Arabic tweets using deep learning. *Multimedia Systems*, 1-12.
41. Drahošová, M., & Balco, P. (2017). The analysis of advantages and disadvantages of use of social media in European Union. *Procedia Computer Science*, *109*, 1005-1009.
42. Siddiqui, S. (2021). Automatic hate speech detection: A literature review. *International Journal of Engineering and Management Research*, *11*(2), 116-121.
43. Rosa, H., Matos, D., Ribeiro, R., Coheur, L., & Carvalho, J. P. (2018, July). A “deeper” look at detecting cyberbullying in social networks. In *2018 international joint conference on neural networks (IJCNN)* (pp. 1-8). IEEE.
44. Van Huynh, T., Nguyen, V. D., Van Nguyen, K., Nguyen, N. L. T., & Nguyen, A. G. T. (2019). Hate speech detection on vietnamese social media text using the bi-gru-lstm-cnn model. *arXiv preprint arXiv:1911.03644*.
45. Gambäck, B., &Sikdar, U. K. (2017, August). Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online* (pp. 85-90).
46. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion* (pp. 759-760).
47. Figure 1: https://datareportal.com/reports/digital-2021-july-global-statshot
48. Figure 2: https://www.thetimes.co.uk/article/social-media-blamed-for-rise-in-racial-and-transgender-hate-crimes-pd2pdqjdg
49. Figure 3: https://www.statista.com/chart/21704/hate-speech-content-removed-by-facebook/