International Journal of Engineering and Management Research
www.ijemr.net
e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-2 (April 2021)
https://doi.org/10.31033/ijemr.11.2.17

Automatic Hate Speech Detection: A Literature Review

Mohiyaddeen[1] and Dr. Sifatullah Siddiqi[2]
[1]Student, Department of Computer Science, Integral University, INDIA
[2]Professor, Department of Computer Science, Integral University, INDIA

[1]Corresponding Author: moinsheik@student.iul.ac.in

ABSTRACT

Hate speech has been an ongoing problem on the Internet for many years. Besides, social media, especially Facebook, and Twitter have given it a global stage where those hate speeches can spread far more rapidly. Every social media platform needs to implement an effective hate speech detection system to remove offensive content in real-time. There are various approaches to identify hate speech, such as Rule-Based, Machine Learning based, deep learning based and Hybrid approach. Since this is a review paper, we explained the valuable works of various authors who have invested their valuable time in studying to identifying hate speech using various approaches.

Keywords— Classification Algorithm, Machine Learning, Hate Speech, Deep Learning, Supervised Learning

I. INTRODUCTION

Social networking sites are the most efficient way to meet new people. However, as social networking sites have grown in popularity, people have discovered an illegal and immoral way to use them. The most commonly encountered and most dangerous misuses of online social media are the expression of hate and harassment. Hate speech may be characterized as violence, hate, intimidation, racism, threats, harassment, insults, provocation, or sexism. These are some of the biggest threats to a social media site online. Several studies have already been worked into the identification of hateful messages in social media platforms[1], along with the dissemination of hateful messages on the dark web[2]. Certain studies have implemented the domain of detection of hate speech but are primarily focused on supervised learning approaches[3]–[5].instruction set. The electronic file of your paper will be formatted further at IJEMR. Define all symbols used in the abstract.

1.1 Hate Speech on Social Media

Hate speech is a form of writing that disparages and is likely to cause damage or danger to the victim on social media. It is a partial, aggressive and malicious speech that targets an individual or a group of people because of their conscious or unconscious intrinsic characteristics[6]. It is a type of speech that shows a strong intent to cause harm, provoke violence, or encourage hate. The social media environment and collaborative worldwide web offer a conducive environment for hate messages against an alleged enemy group to be created, shared, and exchanged.

In 2013, N. Sambuli et al. worked on a project called "Umati: Monitoring Online Dangerous Speech." The project was based on monitoring Hatebase and dangerous speech[7]. According to them, dangerous expressions can be observed in the following ways:

a) It is targeted to a group of people and not a single person. Dangerous speech is an offensive speech that encourages the audience to participate in acts of violence against a particular group of people, therefore In the internet domain, the most prevalent forms of hate speech are related to religion, race, sexual orientation, nationality, class, and gender.

b) Hate Speech may contain one of the pillars of dangerous speech, for instance, statements that classify people as vermin, which claims that a group of people is like rodents or insects.

c) Dangerous speech often incites the listener to support or commit acts of violence against the specific group. The six most common calls to action in dangerous speech are: kill, riot, beat, loot, forcefully evict, and discrimination.

The Internet is inherently open and dynamic, but various communities have their own rules to define the limits of speech. These boundaries differ from one culture to the next and are shaped by historical events and cultural norms[6].

The manual method of detecting and eliminating hate speech posts or comments is time-consuming and computationally expensive. Because of these issues and the prevalence of hateful content on social media, there is a strong case for automated hate speech identification.

Since hate speech, abusive language, and offensive language have recently become subjects of general concern, detecting hate speech has grown to be a major topic by the community of natural language processing (NLP), as demonstrated by the creation of datasets in a variety of languages[8]–[11].

The implementation of systems for automatically detecting abusive and offensive language has followed a general pattern in NLP. Feature-based linear classifiers[8],

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-2 (April 2021)
https://doi.org/10.31033/ijemr.11.2.17

[12], fine-tuning pre-trained language models[13], [14] , and neural network architectures [15]–[17].

There are many approaches by which hate speech detection can be carried out, such as Machine learning, Deep learning, and the Rule-based approach.

II. APPROACHES FOR HATE SPEECH DETECTION

1. Rule-Based Linguistic Approaches

In the Linguistic rule-based approach, Hate speech detection uses a linguistic engine that understands the grammar, morphology, and semantics of a specific language. Furthermore, the program adds rules that check for unique core semantic terms in the sentence in order to determine their potential meanings. For instance, if we input the keyword "bad." The linguistic engine will automatically search for the terms "terrible/awful/unsatisfactory" as well.

2. Machine Learning Approaches

Machine learning creates a mathematical model based on training data to make predictions or decisions without being explicitly programmed. The aim of Machine learning is to make a classifier or regression model through learning the training data set and then use test data set to evaluate the performance of the classifier or regression model. Machine learning can be classified into the following categories based on the nature of the training data. e.g. Supervised learning, Unsupervised learning, Semi-supervised learning.

3. Deep Learning Approaches

The deep learning approach uses neural networks to solve complex problems in an innovative way. When you feed a neural network a series of examples, such as pictures of humans, It can recognize the features that are shared by those pictures. When we use layers of neural network side by side, these layers recognize every detail of the picture to create an effective model. After sufficient training, a neural network becomes refined and capable of classifying unlabeled pictures.

4. Hybrid Approaches

Each solution has its own collection of limitations. And it seems a good solution to merge either two or more approaches into the hybrid approach where one complements another. In the Hybrid approach, we generally combined machine learning, rule-based and deep learning approaches to make an effective model.

III. RELATED WORK

A. Linguistic Rule-Based Approach

In 2014, C. J. Hutto et al. proposed an approach to classify sentiment using VADER, which is a rule-based approach [18]. At first, they created a list of lexical features that are highly sensitive to the sentiment of social media posts. After then they combined that list of lexical features with five general rules that encapsulate syntactical and grammatical rules for presenting sentiment intensity. At last, they have found that VADER performed 96% accuracy using the rule-based model on Twitter sentiments.

Dennis Gitariet al. in 2015 proposed a method to identify the Sentiment Analysis of the Social Media Text using the Rule-based method [19]. In this work, They categorized the hate speech problem into three fields religion, nationality, and race. The main objective of this paper is to develop a classification model that employs sentiment analysis. The developed model not only detects subjective sentences but also classifies and ranks the polarity of sentiment phrases. After then they relate the semantic and subjective features with hate speech. Finally, they achieved 71.55 % precision using the lexicon-based approach.

B. Supervised Learning Approach

Fatahillah et al. (2017) used Naive Bayes Classifier Algorithm to detect hate speech on Instagram using the k-nearest neighbor classifier [20]. They collected the data set using Twitter API from Twitter and annotated those data set manually. After preprocessing and feature engineering phase, they applied the Naive Bayes Classifier algorithm and found 93% of accuracy.

M. Ali Fauzi et al. (2018) proposed an approach to identify hate speech using a set of supervised learning algorithms [21]. They ensembled five different classification algorithms, including K-Nearest Neighbours, Random Forest, Naive Bayes, Support Vector Machine, and Maximum Entropy. They collected the data set using Twitter API and annotated those data set manually. In preprocessing phase, They employed tokenization, filtering, stemming, and term weighting methods. They utilized the bag of words features with TFIDF techniques. The naive Bayes algorithm performed best with 78.3 % of accuracy among all the other five stand-alone classifiers.

In 2019, P. Sari et al. proposed an approach to detect hate speech using logistic regression on Twitter. [22] They collected the data from Twitter and employed Case Folding, Tokenizing, Filtering, and Stemming methods in preprocessing phase. After Pre-processing, the TF-IDF technique is used for vectorization. After Feature engineering, the Logistic regression algorithm has been applied, and they have found 84% of accuracy.

In 2020, Oluwafemi Oriola et al. proposed an approach to detect offensive speech on tweeter [5]. The author collected the data set using Twitter API and annotated those data set into two sections, free speech 'FS' and hate speech 'HT.' In preprocessing phase, they removed special characters, emojis, punctuations, symbols, hashtags, stopwords to clean the data. In the feature

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-2 (April 2021)
https://doi.org/10.31033/ijemr.11.2.17

engineering phase, they employed the TF-IDF technique to transform the text into feature vectors. After applying an optimized support vector machine with n-gram, they have found 89.4% of accuracy.

In 2020, Annisa Briliani et al. proposed an approach to identify hate speech on Instagram using the k-nearest neighbor classifier [23]. They collected the data set using Instagram API from Instagram and annotated those data set manually. They divided the dataset into 2 labels, namely zero and one. In preprocessing phase, they cleaned the data and employed the TF-IDF technique in the feature engineering phase. After then, they applied the k-nearest neighbor algorithm and found 98.13% of accuracy.

C. Unsupervised Learning Approach

Rui Zhao et al. (2015) proposed an approach to detect cyberbullying using Semantic-Enhanced Marginalized Denoising Auto-Encoder [24]. They used two sources of data set. The first source is Twitter, and the second source is Myspace. Twitter data was collected through Twitter stream API, and Myspace data was collected using the web crawling technique. They have achieved 84.9 % accuracy using smSDA for the Twitter dataset, and they have got 89.7% of accuracy with smSDA with the MySpace dataset.

Axel Rodríguez et al. (2019) proposed an approach to detect hate speech content using sentiment analysis on Facebook [25]. They used Graph API to extract the post and comments from Facebook. To remove the unrelated texts VADER and JAMMIN were used. In preprocessing phase, they filtered out all unnecessary stopwords or symbols. Preprocessed documents converted into the vector using TFIDF. The resulting matrix is passed to the k-means clustering algorithm as an input matrix. The most negative articles and responses were collected using sentiment and emotion analysis.

Sylvia Jaki et al. (2019) demonstrated an approach to detect hate speech content using unsupervised learning on Twitter [26]. They collected over 50,00 data set using Twitter API. They used NLP techniques to group the words into similar clusters. They computed three clusters of the top 250 most biassed terms using spherical k-means clustering and skip-grams. As a result, they have got an 84.21% F1 score.

Michele Di Capua et al. (2019) proposed an approach to detect cyberbullying using unsupervised learning [27]. They collected over 54,000 data set from YouTube and Annotated all data sets manually. The GHSOM network algorithm was implemented using the SOM-Toolbox-2 platform. They trained and tested GHSOM using a K-fold method with K = 10. As a result, they have got 64% of accuracy.

D. Deep Learning Approaches

Hugo Rosa et al. (2018) proposed an approach to detect cyberbullying using deep learning [28]. In this paper, the training and testing data set was collected from Kaggle. At first, they initiated CNN, which holds a certain similarity to the issue of cyberbullying. It starts with a single-layer CNN and continues with a completely linked layer with a dropout of 0.5 and softmax performance. Then they combined CNN-DNN-LSTM to achieve maximum accuracy. They employed TFIDF for vector representation. They achieved 64.9% precision with google embeddings.

Tin Van Huynh et al. (2019) proposed an approach to detect hate speech using Bi-GRU-CNN-LSTM Model [29]. In this paper, they collected data from Twitter and categorized their data into three labels (OFFENSIVE, HATE, and CLEAN). After cleaning the data, they implemented three neural network models such as Bi-GRU-LSTM-CNN, Bi-GRU-CNN, and TextCNN to identify hate speech. They achieved a 70.57% of F1 score as a result.

Gambäck et al. (2019) utilized a deep learning algorithm to detect hate speech on Twitter [30]. In this paper, they collected data from Twitter and divided the data set into four categories(sexism, racism, combined(sexism and racism), and non-hate-speech). They employed four CNN models that were trained with character n-gram, word2vec, random vectors combined(word2vec and character n-gram). The author utilized a 10-fold technique to improve the accuracy of the model. Among all four models, word2vec based CNN model performed well with a 78.3% of F-score.

E. Hybrid based Approach

Viviana Patti et al. (2019) proposed a Hybrid based approach to detect hate speech [31]. In this paper, they employed two models. In their first model, they implemented a linear support vector classifier (LSVC), and in the second model, they employed a long short-term memory (LSTM) neural model with word embedding. They concatenated 17 categories, such as HurtLex, with two types, namely LSVC and LSTM. Joint learning with a multilingual word embedding model, including HurtLex, performed best with 68.7% of F1-score.

Safa Alsafari et al. (2020) proposed a Hate speech detection model for Arabic social media [32]. In this paper, they collected the data set using Twitter search API, and the data set is categorized into four classes (Religious, Nationality, Gender, and Ethnicity). They cleaned the data set in preprocessing phase by removing unnecessary words such as URLs, punctuations, symbols, tags, and stopwords. They implemented a three-class classification with CNN and Bert to achieve 75.51% of the F1-score.frequent validation or on demand validation - both can generate considerable, often unnecessary, network traffic and the latter reduces much of the latency gains offered by caching. The viable alternative in such circumstances is resource-driven invalidation where the server invokes a callback on the cache to inform it whenever an update has

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-2 (April 2021)
https://doi.org/10.31033/ijemr.11.2.17

occurred [7][8]. Although this solution involves the server maintaining knowledge of its caches there will be applications which are willing to accept these memory costs in preference to the communication costs of polling-based invalidation.

Various works have already been done in this field. We have categorized all previous works into 5 sections such as Linguistic Rule-Based, unsupervised learning, supervised learning, deep learning, and hybrid approaches. We have also pointed out algorithms and features used in respective research works (Table 1-5).

IV. COMPARATIVE ANALYSIS

Table 1: Supervised Learning Approach (Comparison Analysis)

Paper	Year	Platform	Features and Algorithm	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)
[20]	2017	Twitter	TF-IDF, Naive Bayes	-	-	93.0	-
[21]	2018	Twitter	TF-IDF, Essembled method	-	-	83.4	79.8
[20]	2019	Twitter	TF-IDF, Multinomial Logistic Regression	80.02	82.0	87.68	-
[5]	2020	Twitter	n-gram, Optimized Gradient Boosting	-	-	80.3	-
[23]	2020	Instagram	TF-IDF , K-Nearest Neighbor	94.0	93.0	97.19	93.0

Table 2: Unsupervised Learning Approach (Comparison Analysis)

Author	Year	Platform	Features and Algorithm	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)
[27]	2015	MySpace, Twitter	Bag-of-words (BoW), Latent Semantic Analysis (LSA), smSDA	-	-	87.70	77.60
[24]	2019	Facebook	VADER and JAMMIN, TF-IDF, k-means	-	-	74.42	-
[25]	2019	Twitter	n-gram and k-means	84.21	83.97	-	84.21
[26]	2019	Twitter, YouTube, Formspring	GHSOM network algorithm, SOM-Toolbox-2	60.0	94.0	69.0	74.0

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-2 (April 2021)
https://doi.org/10.31033/ijemr.11.2.17

Table 3: Linguistic Rule-Based Approach (Comparison Analysis)

Author	Year	Platform	Features and Algorithm	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)
[19]	2014	Micro blogging sites	SentiWordNet, VADER,	-	-	96.0	-
[18]	2015	Twitter, Amazon	LIWC, GI, ANEW, SCN,WSD,	81.0	75.0	75.0	-

Table 4: Deep Learning Approach (Comparison Analysis)

Author	Year	Platform	Features and Algorithm	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)
[28]	2018	Kaggle dataset, Formspring, Google,Twitter	CNN-LSTM, Twitter Embedding	84.5	84.2	-	84.2
[29]	2019	Twitter	Bi-GRU-CNN, Bi-GRU-LSTM-CNN, TextCNN,	-	-	-	70.57
[30]	2019	Twitter	CNN, word2vec, character n-grams,	86.61	70.42	-	77.38

Table 5: Hybrid Approach (Comparison Analysis)

Author	Year	Platform	Features and Algorithm	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)
[31]	2019	Benchmark corpora	Word embedding, LSVC, LSTM and HurtLex	60.4	79.8	-	68.7
[32]	2020	Twitter	CNN and mBert	76.95	81.52	--	78.99

V. CONCLUSION

In this paper, we carried out a comprehensive review of various approaches to detect hate speech on social media platforms that have been employed in recent years, along with a brief description of comparative analysis.

The survey work is divided into five major categories: the Linguistic Rule-Based approach, Supervised Learning, Unsupervised Learning, Deep Learning, and Hybrid approaches for hate speech identification, including significant activities in those fields

Taking limited and public datasets for training hate speech detection model is one of the limitations found, and the model can be improved by using real-time big data sets. We have also found that the hate speech is not limited with texts only, but other modes of interactions, such as image and video detection, can also focus on the future.

REFERENCES

[1] E. Spertus. (1997). Smokey: automatic recognition of hostile messages. In: *Innov. Appl. Artif. Intell. - Conf. Proc.*, pp. 1058–1065.
[2] A. Abbasi & H. Chen. (2007). Affect intensity analysis of dark web forums. In: *IEEE Intell. Secur. Informatics*, pp. 282–288, 2007. DOI: 10.1109/isi.2007.379486.
[3] H. Watanabe, M. Bouazizi, & T. Ohtsuki. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, *6*, 13825–13835.

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-2 (April 2021)
https://doi.org/10.31033/ijemr.11.2.17

[4] F. Rodriguez-Sanchez, J. Carrillo-de-Albornoz, & L. Plaza. (2020). Automatic classification of sexism in social networks: An empirical study on Twitter data. *IEEE Access*, 219563–219576.
DOI: 10.1109/ACCESS.2020.3042604.
[5] O. Oriola & E. Kotze. (2020). Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. *IEEE Access*, *8*, 21496–21509. DOI: 10.1109/ACCESS.2020.2968173.
[6] R. Cohen-Almagor. (2011). Fighting hate and bigotry on the internet. *Policy & Internet*, *3*(3), 89–114.
[7] N. Sambuli, F. Morara, & C. Mahihu. (2013). *Umati: Monitoring online dangerous speech*. Available at: http://www.ihub.co.ke/blog/wp-content/uploads/2014/06/2013-report-1.pdf.
[8] Z. Waseem & D. Hovy. (2016). *Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter*.
[9] A. Founta *et al.* (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. *No. Icwsm*, 491–500.
[10] M. O. Ibrohim. (2019). *Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter*.
[11] Ç. Çöltekin. (2020). *A corpus of Turkish offensive language on social media*, pp. 6174–6184.
[12] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. F. Almeida, & W. Meira. (2018). Characterizing and detecting hateful users on Twitter. *arXiv*, Icwsm, 676–679.
[13] P. Liu, W. Li, & L. Zou. (2019). *NULI at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers*.
DOI: 10.18653/v1/s19-2011.
[14] S. D. Swamy, A. Jamatia, & B. Gambäck. (2019). Studying generalisability across abusive language detection datasets. In: *CoNLL 2019 - 23rd Conf. Comput. Nat. Lang. Learn. Proc. Conf.*, pp. 940–950.
DOI: 10.18653/v1/k19-1088.
[15] R. Kshirsagar, T. Cukuvac, K. McKeown, & S. McGregor. (2018). Predictive embeddings for hate speech detection on twitter. *arXiv*. DOI: 10.18653/v1/w18-5104.
[16] P. Mishra, H. Yannakoudakis, & E. Shutova. (2018). Neural character-based composition models for abuse detection. In: *arXiv*.
[17] J. Mitrović, B. Birkeneder, & M. Granitzer. (2015). *nlpUP at SemEval-2019 task 6: A deep neural language model for offensive language detection*, pp. 722–726.
DOI: 10.18653/v1/s19-2127.
[18] C. J. Hutto & E. Gilbert. (2014). VADER : A parsimonious rule-based model for sentiment analysis of social media text In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, pp. 216–225.
[19] N. D. Gitari, Z. Zuping, H. Damien, & J. Long. (2015). A lexicon-based approach for hate speech detection. *IJMUE*, *10*(4), 215–230.
[20] N. R. Fatahillah, P. Suryati, & C. Haryawan. (2018). Implementation of naive bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech. In: *Proc. - 2017 Int. Conf. Sustain. Inf. Eng. Technol. SIET 2017*, pp. 128–131.
DOI: 10.1109/SIET.2017.8304122.
[21] M. A. Fauzi & A. Yuniarti. (2018). Ensemble method for indonesian twitter hate speech detection. *IJEECS*, 294–299. DOI: 10.11591/ijeecs.v11.i1.pp294-299.
[22] P. Sari & B. Ginting. (2019). *Hate speech detection on twitter using multinomial logistic regression classification method*, pp. 105–111.
[23] A. Briliani, B. Irawan, & C. Setianingsih. (2019). Hate speech detection in indonesian language on instagram comment section using K-nearest neighbor classification method. In: *Proc. - 2019 IEEE Int. Conf. Internet Things Intell. Syst. IoTaIS 2019*, pp. 98–104.
DOI: 10.1109/IoTaIS47347.2019.8980398.
[24] R. Zhao & K. Mao. (2016). *Cyberbullying detection based on semantic-enhanced marginalized denoising*.
DOI: 10.1109/TAFFC.2016.2531682.
[25] A. Rodriguez, C. Argueta, & Y. L. Chen. (2019). Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis. In: *1st Int. Conf. Artif. Intell. Inf. Commun. ICAIIC 2019*, pp. 169–174.
DOI: 10.1109/ICAIIC.2019.8669073.
[26] S. Jaki & T. De Smedt. (2018). *Right-wing German hate speech on twitter : Analysis and automatic detection*, pp. 1–31.
[27] M. Di Capua, E. Di Nardo, & A. Petrosino. (2016). *Unsupervised cyber bullying detection in social networks*, pp. 432–437.
[28] H. Rosa, D. Matos, L. Coheur, & P. Carvalho. (2018). A 'Deeper' look at detecting cyberbullying in social networks. In: *2018 Int. Jt. Conf. Neural Networks*, pp. 1–8.
DOI: 10.1109/IJCNN.2018.8489211.
[29] T. Van Huynh, D. Nguyen, K. Van Nguyen, N. L. Nguyen, & A. G. Nguyen. (2019). *Hate speech detection on vietnamese social media text using the Bi-GRU-LSTM-CNN Model*.
[30] B. Gambäck & U. K. Sikdar. (2017). Using convolutional neural networks to classify hate-speech. *No. 7491*, 85–90.
[31] E. W. Pamungkas, V. Patti, & D. Informatica. (2019). *Cross-domain and cross-lingual abusive language detection : A hybrid approach with deep learning and a multilingual lexicon*, pp. 363–370.
[32] S. Alsafari, S. Sadaoui, & M. Mouhoub. (2020). Hate and offensive speech detection on Arabic social media. *Online Soc. Networks Media*, *19*, 100096, 2020.
DOI: 10.1016/j.osnem.2020.100096.