TOXIC CONTENT CLASSIFICATION USING MACHINE LEARNING

Professor, Department of Information Science and Engineering,
A J Institute of Engineering and Technology, Mangalore

Ashika Ruth Saldana, Athmi B.S, Harshitha H.S, Shreya
A J Institute of Engineering and Technology, Mangalore

ABSTRACT

Online toxic content has become a global issue as the number of internet users is growing and technologies are advancing. The cyber-world is a space for everyone, irrespective of their educational and cultural backgrounds. Identifying and differentiating hate content from other toxic content in the cyber environment is a challenging task for automated systems. Classifying toxic content is a difficult task because it involves text processing and context understanding. Social networking platforms have grown in popularity and are used for a variety of activities such as product promotion, news sharing, and achievement sharing, among others. On the other hand, it is also used to spread rumors, bully people, and target specific groups of people. Hate and offensive posts must be detected and removed from social media platforms as soon as possible because they spread quickly and have a wide range of negative consequences for people. In recent years, offensive content and toxic content detection have become popular research topics. Toxic Content classification is an approach to automatically classify toxic content on Twitter into two classes: hate content and non-hate content. We use different features such as a bag of words, term frequency-inverse document frequency (TFIDF), and N-Grams to train and test machine learning algorithms. We also perform a comparative analysis of the different machine learning models. Classifying toxic content is a difficult task because it involves text processing and context understanding. In our approach, we aim to demonstrate a significant improvement in toxic content classification.

Keywords:
Hate Speech, Machine Learning, ANN, SVM, KNN, Naive Bayes, Twitter

INTRODUCTION

Social Media Platforms has become an open book for all its users to express themselves. It is a powerful resource to learn about different races, traditions, cultures, personalities, and attitude of people. Nowadays due to the availability of low-cost internet which is easily accessible to everyone around the globe the social media giants have seen a rapid increase in the number of users Internet users around the world spent 147 minutes per day on social media on average as of 2022, up from 145 minutes the year before. With an average of more than 2.9

billion active users, Facebook continues to be the most popular social media site. In comparison, the number of daily active Instagram Stories users has climbed from 150 million in January 2017 to 500 million in January 2019. A little over 27.4 million tweets are sent on average every day on Twitter. Because of the site's massive user base, simple functionality, and anonymity, antisocial elements and opponents have been drawn to it to engage in illegal activities such as creating false profiles, trolling, abusing, and spreading rumours. Due to this, the propagation of information and daily news has significantly changed when compared to the traditional media, there is no substitute for a truthful and balanced story as traditional media. It is also a well-known fact that most people get to know about important news from networks like Facebook, Twitter, WhatsApp, and Instagram unlike the olden days when publications would verify and publish factual stuff. Defamation and propagation of fake news has become an easy task. There have been numerous regulations and protocols introduced to keep this in check, despite which it is challenging to restrict specific inappropriate comments and information that contain toxic content. The main objective of developing this project is because there is a strong need to check for toxic content online and keep track of it to prevent it from uncontrolled spread and lower the number of marginalized groups being affected by the content posted against them and protect their rights.

Hate Content and Artificial Neural Networks

Artificial Neural Network models have become the state of art solution in classifying hate speech. Its performance depends on the amount of labeled training data. Higher the data better is the model performance in these methods but then datasets have small quantity of data. The Classification of toxic content using ANN finds its application through two classes i.e., Convolutional neural network (CNN) and recurrent neural network (RNN). A CNN finds its application in voice and image processing, and it is very beneficial in computer vision. It includes one or more convolutional layers that perform a convolutional operation on the input and transfer to the next layer as an output.

[Gambäck, & Sikdar (2017)] conducted experiments is on training 4 CNN models Character 4 grams, Word vectors based on semantic information built using word2vec, Randomly generated word vectors, Word vectors combined with character n-grams. This first baseline model achieved precision, recall and F-score values of 86.68%, 67.26% and 75.63%, respectively the second approach resulted in clearly (7.3%) improved recall, for an F-score of 78.29%, even though the precision was slightly reduced.

[Ribeiro, A., & Silva, N. (2019)] Pre-trained Glove and Fast Text models were also employed against women and immigrants at an individual and group level. In the paper which used the SemEval-2019 dataset, for example, word embedding was utilized to detect hate speech in Spanish and English tweets.

[Al-Hassan, A., & Al-Dossari, H (2021)] A combination of ANN models were experimented ANN models which were experimented namely LTSM model, Ensemble model of LTSM and layer of CNN, GRU model, Ensemble model of GRU and a layer of CNN. This resulted in LTSM model is the slowest in terms of training time and GRU is the fastest and that LTSM models perform better than 2GRU models in terms of recall of all the hate classes.

[Jemima, P. P. et al, (2018) Textual analysis may not be the only way to determine whether or not someone is spewing hate speech. There is a chance that information gained from other modalities (such as pictures sent along with text messages) might be useful as well. Word bags

or word embeddings, provide good classification performance.

[Amrutha B R& Bindu K R (2019)] The ULMFiT model has F1 score of 97 and accuracy of 97.5 which shows that pre-trained models can yield better result. ULMFiT model significantly outperformed two other popular models i.e. GRU and CNN.

[José Antonio Garcia Diaz et al., (2022) Shallow Neural networks with few neurons and few hidden layer behave better than deep neural networks. Results obtained with knowledge Integration are, in general, superior to those achieved with ensemble learning, although there is not great difference.

[Steven Zimmerman et al., [2020]] The best-reported model had three epochs and a batch size of ten for positive and negative classifications, with an F1 average score of 75.98%.
RNNs contain numerous algorithms, including LSTMs and GRUs. The issue of evaporating slants that might be while training conventional RNNs was addressed by LSTMs. The GRU has a forget gate, like long short- term memory (LSTM), but no affair gate. It has smaller parameters as a result. Numerous tasks of hate speech identification, similar as the recognition of hate testament, were also fulfilled using variations of the LSTM.
[Qian, J et al., (2018)]

Two-sub caste RNNs were proposed by [Founta et al. (2019)] The Glove-style tweet characteristics and the metadata on people, networks, and content were used to build the unified model. The cyber bullying dataset, the spiteful dataset, the annoying dataset, the affront dataset, and the vituperative dataset were just a few of the datasets they used to test their approach. Depending on the input features and dataset used, the model gave a range of results; nevertheless, the RNN and metadata interleaved models were the best ones, with an average delicacy of 90.2

Authors	**Purpose**	**Data Set**	**Methodology**	**Results**	**Remarks**
Areej Al-Hassan, & Hmood Al-Dossari.(2021)	Detection of hate speech in arabic tweets using deep learning methods, multiclassification.	Twitter dataset	• LTSM model. • Ensemble model of LTSM and layer of CNN. • GRU model. • Ensemble model of GRU and a layer of CNN.	1)LTSM model is the slowest in terms of training time and GRU is the fastest 2)2 LTSM models perform better than 2GRU models in terms of recall of all the hate classes.	Both GRU and LTSM works good for classifying Arabic hate speech into 5 classes.
Bjorn Gambäck, & Utpal Kumar Sikdar.(2017)	Using Convolutional Neural Networks to Classify Hate-Speech.	Twitter dataset [a dataset of 6655 tweets]	The experiment is based on training 4 CNN models Character 4 grams Word vectors based on semantic	This first baseline model achieved precision, recall and F-score values of 86.68%, 67.26%	Word2vec model without character n-grams still achieved the best results of all the compared

			information built using word2vec, Randomly generated word vectors Word vectors combined with character n-grams.	and 75.63%, respectively The second approach resulted in clearly (7.3%) improved recall, for an F-score of 78.29%, even though the precision actually was slightly reduced.	models.

P. Preethy Jemima, Bishop Raj Majumder, Bibek Kumar Ghosh, Farazul Hoda(2022)	Hate Speech Detection using Machine Learning.	Twitter dataset	Exploratory Data Analysis, Data Cleaning, Pre-processing & Transformation, Data Partition,Modelling,Evaluation.	Textual analysis may not be the only way to determine whether or not someone is spewing hatespeech. There is a chance that information gained from other modalities (such as pictures sent along with text messages) might be useful as well. Word bags or word embeddings, provide good classification performance.	There is a need for a uniform data set that can be used to compare characteristics and approaches.
Amrutha B R, Bindu K R(2019)	Detecting Hate Speech in Tweets Using Different Deep Network Architectures.	WikiText103 dataset	• Gated Recurrent Unit • Convolution Neural Network • ULMFiT • AWD-LSTM	The GRU and CNN model shows an F1 score of 65.4 and 64.16 and accuracy of 96 and 94.5. The ULMFiT model has F1 score of 97 and accuracy of 97.5	ULMFiT model significantly outperformed two other popular models as well as traditional approaches.

				which shows that pretrained models can yield better result.	
1. José Antonio García-Díaz 2.Salud María Jiménez-Zafra · 3.Miguel Angel García-Cumbreras· 4.Rafael Valencia-García1[2022]	Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers.	SpanishMisoCorpus 2020 AMI 2018 EVALITA 2018 HatEval 2019	[1]The DataResolver module acts as input. [2] TextCleaner module cleans andpre processes thetexts. [3]DatasetSplitter module does the training, validation, and testing splits. [4] ModelResolver is the other input and is responsible to select one strategy for evaluate the datasets. [5] HyperParameterSelector module is capable of evaluating different neural network architectures and hyper-parameters	• Shallow Neural networks with few neurons and few hidden layer behave better than deep neural networks. • Results obtained with knowledge integration are, in general, superior to those achieved with ensemble learning, although there is not great difference.	• In This paper we have a study of different datasets • In Order to determine which kind of individual features are most effective for hate-speech detection, how these features can be combined, if linguistic features could provide insights Regarding the identification of hate-speech, and if the methods proposed here outperforms the state-of-the-art results.
Steven Zimmerman, Chris Fox, Udo Kruschwitz [2020]	Improving Hate Speech Detection with Deep Learning Ensembles.		CNN structure to represent 50 tokens based on CNN parameters (epochs, weights, and batch size).	The best-reported model had three epochs and a batch size of ten for positive and negative classifications, with an F1average score of 75.98%.	1] Failure in weight initial-ization of a neural network 2]Different approaches, such as LSTM networks based on character representation should be considered

1]Alison P.Ribeiro 2]N adia F. F. da Silva[2019]	Convolutional Neural Networks for Hate Speech Detection Against Women and Immigrants on Twitter.	The data for the task consists of 9000 tweets in English for training, 4469 tweet in Spanish.	There are two tasks TASK A Two-class classification problem in which participants have to predict whether a tweet, in English or Spanish, with a particular tar-get is hateful or not hateful TASK B (i) classify hate tweets into English and Spanish, where tweets with hate speech, against women orimmigrants, were identified as aggressive or non-aggressive, (ii) identify the harassed target as just one person or group of individuals.	1]Experiments resulted in 0.488 of F1-score for English and0.696 for Spanish with CNN model using word embeddings 2]The proposed model obtained in Task A 0.488 and 0.696 F1-score for English and Spanish.	In this paper, a CNN was implemented based on the architecture proposed by Zhangand Wallace 2015 andafinetuning of hyperpara-meters was not done for the proposed tasks (tasksA andB). In addition, other features were not ex-ploited as sarcasm and irony, inherent in this type of domain.

Hate Content and Support Vector Machine

One of the supervised machine learning techniques used for various categorization issues is the Support Vector Machine (SVM) algorithm. It has uses in information extraction, text categorization, medical diagnosis, and credit risk analysis. SVMs are ideally suited for high dimensional data. There are a lot of arguments in favour of this statement. In particular, the classifiers produce the same hyper plane for repeated training sets, their complexity is determined by the number of support vectors rather than data dimensions, and they are more generalizable

[Oriola , Kotzé (2020)] The official languages of South Africa are English, isiXhosa, Sesotho, Setswana, isiZulu, and Afrikaans. Some tweets may include other native languages to improve expressivity and convenience, and they may also test various hyperparameter combinations of a machine learning classifier. Using the Twitter dataset, multi-tier meta-learning algorithms

consistently captured the combination of words and characters combined with negative sentiment-based features to detect hateful speech. Thereby, we tested it using LogReg, SVM, RF, and GB meta-learners.

[Olusegun Folorunso b et.al. (2018)] The proposed approach is broken down into five modules. metadata extractor, data pre-processing, data representation, detection, and classification. To avoid the fragmentation issue that is present with online clustering, the database of hate speech is used as training data for topic grouping. The Bayes theorem was not used to classify hate speech as a subject cluster based on themes that could not be automatically inferred from the seed database. using the Twitter dataset, the strategy for classifying hate speech on Twitter is presented.

[Asogwa D et.al. (2019)] Hate speech is the outspoken expression of hatred or enmity toward an individual or group based on traits like race, religion, sex, or sexual orientation. It has been tested to classify hate speech using a number of categories using the Twitter dataset, including those that range from offensive to non-offensive. Weka machine learning tools, Java programming language, and NetBeans IDE were used to implement the software.

[Ibrohim, Budi (2018)]This includes identifying the target, category, and intensity of hate speech on Indonesian Twitter. Twitter data annotators annotate whether tweets contain hate speech and abusive language or not. RFDT classifiers, The best accuracy is provided by the RFDT classifier, which uses LP as the transformation technique for quick computing speed. For the annotation process, we built a web-based annotation system to make it easy for the annotators to annotate data.

[Chukwuneke C. et.al. (2017)] Text can be effectively classified using the supervised machine learning algorithm known as the support vector machine (SVM). SVM typically performs well for text classification due to its capacity to generalise into broad dimensions, which text categorization frequently does. Although it achieved an accuracy rate of 83.5%, this system nevertheless faced several difficulties. Addressing issues like managing enormous text corpora, word similarity in text documents, and linking text documents with a subset of class categories are among them.

Authors	purpose	Dataset	Methodology	Results	Remarks
Oluwafemi Oriola , Eduan Kotzé	The official languages of South Africa are English, IsiXosha, Sesotho, Setswana, isiZulu, and Afrikaans. Government and business generally speak English, and the majority of people speak it as a second language. Some tweets may comprise other native tongues for expressivity and convenience.	Twitter dataset	In Python 3.6 [29], every feature category and its hybrid, based on vertical stacking, was examined. For the best performance, we tested various hyper-parameter combinations of machine learning classifiers. We also used a multi-tier meta-learning model and tested it using LogReg, SVM, RF, and GB meta-learners.	Word and character n-gram features' usefulness in identifying hate speech and offensive speech was negatively correlated with their inferior performance in the other areas. , the use of multi-tier meta-learning algorithms consistently captured the combination of words and characters combined with negative sentiment-based features to detect hateful speech.	Lack of Non-Discriminatory Feature (NDSM) - where tweets were misclassified as a result of having specific attributes in other classifications. The inability of the classifier to recognise the contexts of the tweets, which aid to define the class of tweets was the second biggest cause of misclassification (10.48%).
Femi Emmanuel Ayo a,∗ , Olusegun Folorunso b , Friday Thomas Ibharalu b , Idowu Ademola Osinuga c	Support vector machine (SVM) is a supervised machine learning technique developed by Cortes and Vapnik [141] for text classification. SVM is the most used binary classifier for identifying hate speech emotion. When compared to similar algorithms, the implemented SVM performed better.	Twitter dataset	The strategy for classifying hate speech on Twitter is presented. The proposed approach is broken down into five modules: metadata extractor, data pre-processing, data representation, detection, and classification. To avoid the fragmentation issue that is present with online clustering, the database of	The sentiment class of a hate tweet can be classified as mild, moderately severe, or severe. These guidelines were utilised to make informed decisions about the categorisation of hateful tweet sentiment. The Bayes	The sentiment class of a hate tweet can not be classified as mild, moderately severe, or severe. These guidelines were utilised to make informed decisions about the categorisation of hateful tweet sentiment. The Bayes theorem was not used to classify hate speech as a subject cluster based on the themes that cannot be automatically inferred from the seeders database.

			hate speech is used as training data for topic grouping.	theorem was used to classify hate speech as a subject cluster based on the themes that can be automatically inferred from the seeders database.	
Asogwa D., Chukwuneke , Ngene C., Anigbogu G.	Hate speech is outspoken expression of hatred or enmity toward an individual or group based on traits like race, religion, sex, or sexual orientation. It has been tested to classify hate speech using a number of categories, including those that range from offensive to non-offensive.	twitter dataset	Weka machine learning tools, Java programming language, and NetBeans IDE were utilised to implement the hate speech detection and categorization. Supervised learning methods, support vector machines, and Nave Bayes were also used. The sizes of the data that was gathered were decreased from the instances.	The goal of this work was to create an SVM-based system for the detection and classification of hate speech. Weka classifiers and the java programming language were used to create the software. The dataset was obtained from a machine learning repository system for unique client identifiers (UCI).	The main goal of this research was to identify and categorise hate speech using machine learning methods, specifically Naive Bayes and SVM for identifying its attributes. To identify which algorithm will be most effective in detecting hate speech, researchers should compare it to others like Random Forest, k-nearest neighbours' algorithm (k-NN), deep learning, and Artificial Neural Network (ANN).

Hate content and Naïve Bayes

The Nave Bayes algorithm is a supervised learning method for classification issues that is based on the Bayes theorem. It is mostly employed in text categorization with a large training set. The Naive Bayes Classifier is one of the most straightforward and efficient classification algorithms available today. It aids in the development of rapid machine learning models capable of making accurate predictions. Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur.

[Femi Emmanuel Ayo et.al. (2021)] Model is divided into four phases: metadata representation, training, clustering and classification. The clustering task is for clustering real-time tweet. fuzzy logic was used for hate speech classification. The use of an automatic topic spotting measure based on naïve Bayes model to improve features representation was

introduced.Based on the result of the combinatorial algorithm, probability distribution is utilised to differentiate a single hate tweet from all other tweets. In order to categorise tweet items as hate tweets or non-hate tweets based on a certain threshold value, the idea of degree of support in Bayesian networks was modified.

[Tehseen Zia et.al. (2017)] When identifying insults or sentiments, two methods were employed: message classification and human message rating. Message is marked as "flame" if it uses abusive or insulting language. To Extract the tweets popular hash tags, anti-Islamic and anti-Jewish web pages are used. Three extensively supervised learning algorithms—SVM, NB, and KNN—are employed. First, we classified religious viewpoints using these algorithms, and then we determined their sentiment. SVM is the best classifier we could find for categorising sentiment.

[Nabil Badria et.al. (2022)] Our suggested BiGRU-Glove-FT model outputs the likelihood that an input text belongs to the wrong class given an input text (Offensive or Not Offensive). Word embedding matrices from "Glove and FastText" are used to feed the input text into the model.Naive Bayes and SVM models with Word Level TF-IDF as a feature in the second model and Count Vectors in the first model produced the best results.

[Femi Emmanuel Ayo a et.al. (2020)] Naïve Bayes method for opinion classification in Twitterdata. The raw tweets are first pre-processed to remove noise from the dataset. The extract features are fed into the NB classifier for final classification of hate-related opinions. The experimental results showed that the developed NB method outperform the baseline model. When compared to comparable techniques, the created generic metadata architecture for hate speech sentiment categorization outperformed them on the F1 scale.

Authors	**Purpose**	**Data Set**	**Methodology**	**Results**	**Remarks**
Femi Emmanuel Ayo, Olusegun Folorunso b, Friday Thomas Ibharalu b, Idowu Ademola Osinuga c , Adebayo Abayomi-Alli (2021)	To manage negative expressions on Twitter, a probabilistic clustering model for hate speech categorization was created.	Twitter data set	Model is divided into four phases: metadata representation, training, clustering and classification. The clustering task is for clustering real-time tweet. fuzzy logic was used for hate speech classification. The use of an automatic topic spotting measure based on naïve Bayes model to improve features representation was introduced.	Based on the result of the combinatorial algorithm, probability distribution is utilised to differentiate a single hate tweet from all other tweets. In order to categorise tweet items as hate tweets or non-hate tweets based on a certain threshold value, the idea of degree of support in Bayesian networks was modified.	hate speech classification architecture can develop to address the issues of generic metadata architecture, scalability, class imbalance data, threshold settings and fragmentation.

Tehseen Zia, M. Shehbaz Akram, M. Saqib Nawaz, Basit Shahzad, Abdullatie M Abdullatif, Raza Ul Mustafa, Ikramullah Lali (2017)	The purpose of this research is to develop a system that can recognise hate speech in tweets and communications sent over Twitter.	Open source Twitter API and MySQL database to build data collection server.	When identifying insults or sentiments, two methods were employed: message classification and human message rating. Message is marked as "flame" if it uses abusive or insulting language. To Extract the tweets popular hash tags, anti-Islamic and anti-Jewish web pages are used.	Three extensively used supervised learning algorithms—SVM, NB, and KNN—are employed. First, we classified religious viewpoints using these algorithms, and then we determined their sentiment. SVM is the best classifier we could find for categorising sentiment.	Focused primarily on religion and achieved higher accuracy than general techniques. Additionally, a comparison study of classifiers was provided in the results and discussion part, which might be useful in the future when choosing the right classifier.
Nabil Badria,Ferihane Kboubia, Anja Habacha Chaibia (2022)	It offers a technique to detect hate speech on social media platforms based on a mix of Glove and FastText word embedding as input characteristics and a BIGRU model.	OLID dataset	Our suggested BiGRU-Glove-FT model outputs the likelihood that an input text belongs to the wrong class given an input text (Offensive or Not Offensive). Word embedding matrices from "Glove and FastText" are used to feed the input text into the model. Naive Bayes and SVM models with Word Level TF-IDF as a feature in the second model and Count Vectors in the first model produced the best results.	The results collected demonstrate the capability of the suggested model (BiGRU Glove FT) to identify improper content. Using an efficient learning approach that separates the text into offensive and not offensive language, this model detects hate speech using the OLID dataset..	to further investigate the application of deep neural network architectures for the identification of hate speech. We will look at the use of other word embedding methods. We would broaden this effort to incorporate more datasets, such the Arabic dataset.
Femi Emmanuel Ayo a, Olusegun Folorunso b, Friday Thomas Ibharalu b, Idowu Ademola Osinuga(2020)	The work addressed problems with Twitter data streams by presenting a generalised metadata architecture for hate speech categorization in Twitter. On all measures for hate speech identification, generic metadata architecture	Twitter data set	Naïve Bayes method for opinion classification in Twitter data. The raw tweets are first pre-processed to remove noise from the dataset . The extracted features are fed into the NB classifier for final classification of hate-related opinions. The experimental results showed that the developed NB method outperform the baseline model.	When compared to comparable techniques, the created generic metadata architecture for hate speech sentiment categorization outperformed them on the F1 scale. The created approach is excellent for automatically detecting and	These method accessibility can be useful for developing a machine learning model that can recognise and categorise hate speech.

	performed better.			categorising topics, according to the statistical validation of findings.	

Hate content and K-nearest Neighbours

A method of classifying an object based on its closest neighbour is known as “K-Nearest Neighbor (KKN)”. Using previously described data, this method classifies a set of data. The K-Nearest Neighbor algorithm arranges or gathers the closest K neighbours from a document between training documents, using the label of the most similar neighbour K class to predict it. This method is used to categorise unknown documents. KNN is lazy learning and statistical classification algorithm. In terms of memory and time, testing phase of KNN algorithm is costly.

[Janak Sachdeva et al., [2021]] In this the research work, has attempted to categorise tweets and assign them to one of three categories, namely, Racist, sexist, or neither. If hate speech is present, the tweet is further classified as "hate based on racism" or "hate based on sexism". In this study, there are two ensemble-based models: one is based on random forest, KNN, and logistic regression, and the other is based on linear SVC, logistic regression, and random forest. A few deep learning models that categorise Twitter hate speech using self- and pre-trained word embeddings have also been introduced.

[Annisa Briliani et al., [2019]] This study is to develop a system that uses the KNN classification algorithm to determine whether or not comments on Instagram contain. Text classification, POS Tagging, Feature extraction, POS Weighted TF-IDF are the methods used to classify the tweets. Based on the findings of this study, it is possible to draw some conclusions about whether or not the K-Nearest Neighbor algorithm is appropriate for classifying hate speech.

[Mahamat Saleh Adoum Sanoussi et al., [2022]] This article's goal is to identify hate speech in texts written in the "lingua franca," a blend of French and Chadian. The project's objective is to stop the alarming spread of hate speech on Facebook. Text data is first gathered and annotated. Text pre-processing is then applied to the dataset. A feature representation model is created in the end. The machine learning classifier should then be built and trained using labelled data. Each stage's components include data cleaning, text pre-processing, feature extraction, and classification. Main goal is to compare the effectiveness of classifiers and feature strategies for prediction.

[Vijay & Dr. Pushpneel Verma [2021]] In this work, two supervised machine learning algorithms have been used to categorise texts as either having hateful content or not. 10,000 text documents are selected solely for testing after being carefully mixed up. After that, all punctuation, digits, stop words, and URLs were removed, and all text documents were converted to lower case. Next, eliminate all white spaces. Lastly, from the text documents, a document term matrix was created. This experiment shows that SVM, as opposed to KNN, provides the highest accuracy when using a linear kernel function.

[Sindhu Abro et al., [2020]] This study evaluates the performance on a publicly available dataset with three different classes of three feature engineering techniques and eight machine learning algorithms. Proposed system categorised tweets into 3 groups "hate speech," "offensive but not hate speech," and "neither hate speech nor offensive speech". The research technique is comprised of six essential processes processes- data collection, data pre-processing, feature engineering,

Authors	Purpose	Dataset	Methodology	Results	Remarks
Janak Sachdeva, Kushank Kumar Chaudhary, Harshit Madaan, Priyanka Meel (2021)	The purpose of this experiment is to compare ensemble-based and state-of-the-art neural network-based models. Also aims to categorise tweets and assign them to one of three categories, namely, Racist, sexist, or neither.	Twitter Dataset	Deep learning models that use self- and pre-trained word embeddings have been introduced. First ensemble based model-on linear SVC, logistic regression, random forest, KNN and second model- on random forest, KNN, and Logistic regression.	The result shows that compared to the other models, the ensemble classifier model built using RFC, LR, and linear SVC produces a somewhat superior outcome. While the results of linear SVC are marginally better than those of LR and RFC. The KNN model fared the worst.	If hate speech is present, the tweet is classified as "hate based on racism" or "hate based on sexism."
Annisa Briliani, Budhi Irawan, Casi Setianingsih (2019)	The purpose of this paper is to develop a system that uses the K-Nearest Neighbor classification algorithm to determine whether or not the comments on Instagram contain hate speech.	Instagram Dataset	Text classification, Supervised learning, POS Tagging, Feature Extraction, Term Frequency-Inverse Document Frequency (TFIDF), POS Weighted TF-IDF, K-Nearest Neighbor.	Result seen is that the optimal K number is 3, which has 98% precision, 98.13% recall, and 98.13% accuracy. This project's output categorises comments as hate speech or not.	This final project will build a system that can determine whether an Indonesian statement in a comment section on Instagram is hate speech or not using the K- NN technique.

Vijay, Dr. Pushpneel Verma (2021)	The purpose, is to use two supervised machine learning algorithms to categorise texts as either having hateful content or not on social media.	Dynamically Generated hate dataset	Text documents are jumbled. Then, eliminated all punctuation, digits, stop words, and URLs before changing all the text documents to lower case. Then removed all white spaces. Finally, created a document term matrix from the text documents.	KNN algorithm provided accuracy of 60.8%, 62.65%, and 62.75%, respectively. SVM using a linear kernel function gave a maximum accuracy of 65%.	The main understanding from this paper is that 2 main algorithms have been used i.e., Support Vector Machine and KNN algorithm. From this experiment it is noticed that SVM gives maximum accuracy with linear kernel function than KNN.

CONCLUSION

This project aims to compare standard machine learning techniques applied to toxic content classification on social media-specifically to the toxic content classification of the data gathered from the social network twitter and to find the best performing methods for the datasets collected from twitter. As internet material expands, so does the proliferation of poisonous content. We identify and investigate the issues that Twitter data presents for harmful content categorization in text. Furthermore, many current techniques have an interpretability problem, which means it might be difficult to comprehend why the computers make the judgements they do. We present an SVM technique that delivers near-best-practice performance while being simpler and delivering more clearly interpretable choices. We also discuss about artificial neural networks methods, naïve bayes and KNN algorithms to classify the toxic content in the twitter dataset

REFERENCES

[1] Alathur, S., Chetty, N., Pai, R. R., Kumar, V., & Dhelim, S. (2022). Hate and False Metaphors: Implications to Emerging E-Participation Environment. Future Internet, 14(11), 314.

[2] Al-Hassan, A., & Al-Dossari, H. (2021). Detection of hate speech in Arabic tweets using deep learning. *Multimedia Systems*, 1-12.

[3] Amrutha, B. R., & Bindu, K. R. (2019, May). Detecting hate speech in tweets using different deep neural network architectures. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (pp. 923-926). IEEE

[4] Ali, R., Farooq, U., Arshad, U., Shahzad, W., & Beg, M. O. (2022). Hate speech detection on Twitter using transfer learning. Computer Speech & Language, 74, 101365.
[5] Castaño-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, 58, 101608.
[6] Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. Aggression and violent behavior, 40, 108-118.
[7] Corazza, M., Menini, S., Arslan, P., Sprugnoli, R., Cabrio, E., Tonelli, S., & Villata, S (2018) Inriafbk at germeval 2018: Identifying offensive tweets using recurrent neural networks.
[8] Femi Emmanuel Ayo, Olusegun Folorunso b, Friday Thomas Ibharalu b, Idowu Ademola Osinuga c (2021), A probabilistic clustering model for hate speech classification in twitter, 114762.
[9] García-Díaz, J. A., Jiménez-Zafra, S. M., García-Cumbreras, M. A., & Valencia-García, R. (2022). Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers. Complex & Intelligent Systems, 1-22.
[10] 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).
[11] Grolman, E., Binyamini, H., Shabtai, A., Elovici, Y., Morikawa, I., & Shimizu, T. (2022, July). Hateversarial: Adversarial attack against hate speech detection algorithms on twitter. In Proceedings of the 30th ACM Conference on User Modelling, Adaptation and Personalization (pp. 143-152).
[12] J. Serra, I. Leontiadis, D. Spathis, G. Stringhini, J. Blackburn, A. Vakali, Class-based prediction errors to detect hate speech with out-of-vocabulary words, in: Proceedings of the First Workshop on Abusive Language Online, 2017, pp. 36-40
[13] Khan, S., Kamal, A., Fazil, M., Alshara, M. A., Sejwal, V. K., Alotaibi, R. M., ... & Alqahtani, S. (2022). HCovBi-caps: hate speech detection using convolutional and Bidirectional gated recurrent unit with Capsule network. IEEE Access, 10, 7881-789.
[14] Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021, May). Hate explains: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 17, pp. 14867-14875).
[15] M.O. Ibrohim, E. Sazany, I. Budi, identify abusive and offensive language in Indonesian Twitter using deep learning approach, J. Phys. Conf. Ser.
[16] Ò.G. i Orts, Multilingual detection of hate speech against immigrants and women in Twitter at SemEval-2019 task 5: Frequency analysis interpolation for hate in speech detection, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 460–463.
[17] P. P. Jemima, B. R. Majumder, B. K. Ghosh and F. Hoda, "Hate Speech Detection using Machine Learning," 2022 7th International Conference on Communication and Electronics Systems(ICCES), 2022,pp.12741277, doie10.1109/ICCES54183.2022.9835776.
[18] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech intext, ACM Comput Surv. 51 (4) (2018) 85:1–85:30
[19] Roy, P. K., Tripathy, A. K., Das, T. K., & Gao, X. Z. (2020). A framework for hate speech detection using deep convolutional neural networks. IEEE Access, 8, 204951-204962.
[20] Sharma, A, Kabra, A, & Jain, M. (2022). Ceasing hate with MoH: Hate Speech Detection in Hindi–English code-switched language. Information Processing & Management, 59(1), 102760.

[21] Tehseen Zia, M. Shehbaz Akram, M. Saqib Nawaz, Basit Shahzad, Abdullatie M Abdullatif, Raza Ul Mustafa, Ikramullah Lali (2017): Identification of hatred speeches on Twitter.ISSN-2393-2835.
[22] Yin, W., & Zubiaga, A. (2021). Towards generalizable hate speech detection: a review on obstacles and solutions. PeerJ Computer Science, 7, e598.
[23] Zampieri, M., Malmasi, S, Nakov, P, Rosenthal, S, Farra, N, & Kumar, R. (2019) Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983.
[24] Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. 2011. Combining lexicon-based and learning-based methods for Twitter sentiment analysis. Technical Report HPL-2011-89.
[25] Zhao, Y., Tao, X., 2021. Zyj123@ DravidianLangTech-EACL2021: Offensive language identification based on xlm-RoBERTa with DPCNN. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages. pp. 216–22