ZAMBIA INFORMATION COMMUNICATION TECHNOLOGY (ICT) JOURNAL

Volume 7 (Issue 1) (2023) Pages 79-84

Detecting Hate Speech and Offensive Language using Machine Learning in Published Online Content

Sinyangwe Clement Mulenga
Chalimbana University
Lusaka, Zambia.
clementsinyangwe1@gmail.com

Douglas Kunda
ZCAS University
Lusaka, Zambia
douglas.kunda@zcasu.edu.zm

William Abwino Phiri
Chalimbana University
Lusaka, Zambia
williamabwino@gmail.com

Abstract*—Businesses are more concerned than ever about hate speech content as most brand communication and advertising move online. Different organisations may be incharge of their products and services but they do not have complete control over their content posted online via their website and social media channels, they have no control over what online users post or comment about their brand. As a result, it became imperative in our study to develop a model that will identify hate speechand, offensive language and detect cyber offence in online published content using machine learning. This study employed an experimental design to develop a detection model for determining which agile methodologies were preferred as a suitable development methodology. Deep learning and HateSonar was used to detect hate speech and offensive language in posted content. This study used data from Twitter and Facebook to detect hate speech. The text was classified as either hate speech, offensive language, or both. During the reconnaissance phase, the combined data (structured and unstructured) was obtained from kaggle.com. The combined data was stored in the database as raw data. This revealed that hate speech and offensive language exist everywhere in the world, and the trend of the vices is on the rise. Using machine learning, the researchers successfully developed a model for detecting offensive language and hate speech on online social media platforms. The labelling in the model makes it simple to categorise data in a meaningful and readable manner. The study establishes that in fore model to detect hate speech and offensive language on online social media platforms, the data set must be categorised and presented in statistical form after running the model; the count indicates the total number of data sets imported. The mean for each category, as well as the standard deviation and the minimum and maximum number of tweets in each category, are also displayed. The study established that preventing online platform abuse in Zambia requires a comprehensive approach that involves government law, responsible platform policies and practices, as well as individual responsibility and accountability. In accordance with this goal, the research was effective in developing the detection model. To guarantee that the model was completely functional, it was trained on the English dataset before being applied to the local language dataset. This was because of the fact that training deep learning models with local datasets can present a number of challenges, such as limited, biased data, data privacy, resource requirements, and model maintenance. However, the efficacy of these systems varies, and there have been concerns raised about the inherent biases and limitations of automatic moderation techniques. The study recommends that future studies consider other sources of information such as Facebook, WhatsApp, Instagram, and other social media platforms, as well as consider harvesting local data sets for training machines rather than relying on foreign data sets; the local data set can then be used to detect offences targeting Zambian citizens on local platforms.

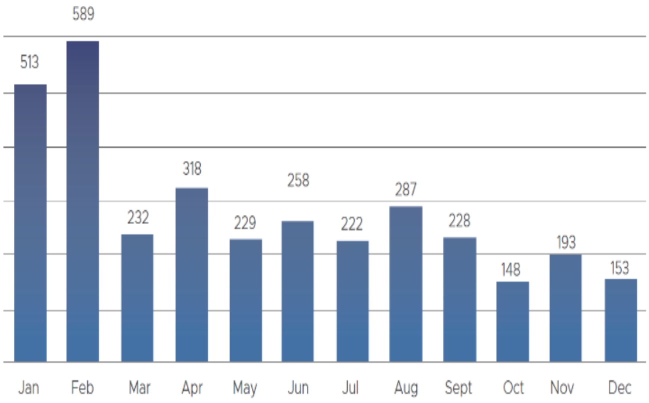
Keywords—**Hate Speech, Offensive Language, Online Content, Machine Learning.**

INTRODUCTION

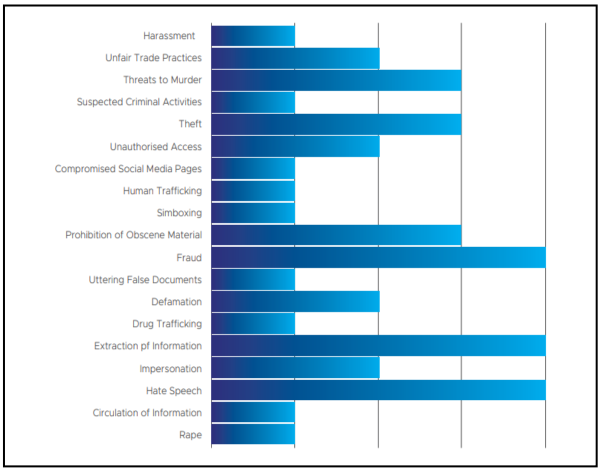
Hate speech refer to such messages of hatred that expresses an incitement to harm against specific social or demographic groups [14]. These includes expressions that build a climate of bias and intolerance, which is supposed to encourage discrimination, antagonism, and attacks. [7] Cybercrime is defined as a crime done by utilizing computers or other forms of communication to induce fear and worry in others or to damage, harm, or destroy property. Cybercrime is a broad word that encompasses computer-assisted crime in which computers and technology play a supporting role [7]. Bob Thomas designed the creeper virus in 1971 in order to infect the systems of the Advanced Research Project Agency Network. At the Massachusetts Institute of Technology in 1988, Robert T. Morris invented the first computer worm in 1988 at the Massachusetts Institute of Technology. In 1998, protesters used a technology called FloodNet to launch a denial-of-service attack on Mexico's president's website [9]. DDoS attacks, Man-in-the-middle attacks, information escape, PROBE, User-To-Root, Remote-To Local, hate speech, and cyberbullying are examples of attacks. To reduce or avoid the amount of cyber-attacks, enhanced security measures should be implemented [6].

PROBLEM STATEMENT

The digital space has expanded dramatically over the last few decades, with the most notable phenomenon being the expansion of the social media space. Although the researcher managed to obtain data relating to this study, it should be noted that due non existance of dataset based on local scenarios, we obtained dataset from kaggle based on american english. And according to available data, social media is used by many people accross the world, making it an essential component of any brand's primary marketing platform. Social media has created a space for people to communicate, express their opinions, and share content, in addition to marketing and brand building. Content posted on social media draws criticism from a variety of user demographics, some of which constitute hate speech. Hateful comments on social media are a growing problem in the online world and a top priority for social media developers, marketers, and law enforcement. Hateful comments and discourse on social media can cause public tension and even violence [15]. Many concerns have been raised in Zambia about hate speech, particularly hate speech on online and social media platforms. Evidence from the Zambia Information and Communications Technology Authority supports the speedy growth in the people engaging in cybercrime and hate speech. According to data obtained from ZICTA. The number of complaints coming as a result of social media and online platforms has skyrocketed [11]. Below are the statistics showing various cyber offences reported to the authority.



Cyber Cases Received in 2020



complaint volumes by category

According to the statistics presented above, this issue has grown in importance over the past decade, and manually detecting such content on the web is a time-consuming task. Detecting hate speech or offensive language is a time-consuming and resource-intensive process when done manually and without the use of a tool [2]. A governmental or private organization would lack the resources to track all issues related with brand content in a bid to detect hate-speech. As a result, there may be need for people to expand their personnel to fulfill the increased need for this form of control to digital content. It's also worth mentioning that humans' capacity to detect hate speech published content is influenced by a number of things, these may include; their energy levels, abilities to reading, as well as personal preconceptions regarding what qualifies as unpleasant content. As a result, there is a need to develop an automated model capable of detecting such toxic content on the internet [2]. The objectives of this research was to create a model that will detect cyber offense in online published content using machine learning, develop a model for detecting hate speech and offensive language on online social media platforms using machine learning and evaluate the effectiveness of the machine learning model for detecting offensive language and hate speech published online, especially on social media platforms [3].

The research focused on detecting cybercrimes that are often committed on the cyberspace specifically hate speech and offensive language in online published content. The study took the case of twitter as a source of the content based on the fact that twitter is one of the widely used platform as indicated in the ZICTA report for 2020. Social learning and general strain theories were used to guide this study [11].

LITERETURE REVIEW

[19] A study on automated hate speech identification and span extraction in underground hacking and extremist forums. They employed a hate speech dataset made up of posts from Hack Forums, an online hacking forum, as well as Storm Front and Incels.co, two extremist communities. The researchers merged their dataset with a Twitter hate speech dataset to train a multi-platform classifier, demonstrating that training a classifier on various data sources does not always boost performance [21].

[18] a new investigation to assess the detection of fake news and hate speech in Ethiopian languages. To improve the performance of all evaluation indicators from various social media sites, deep learning (DL) and machine learning (ML) methodologies were advocated. With the rise of social media platforms that enable anonymity, simple access, online community formation, and online debate, it was established that identifying and tracking hate speech has become a serious problem for society, people, politicians, and researchers. Many research communities have recently expressed interest in automatic fact or claim verification, but the findings are still unsatisfactory [25].

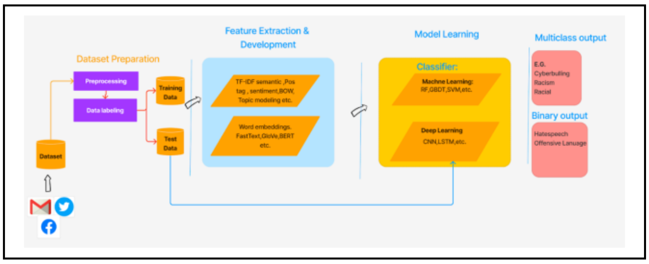
[16] conducted another an investigation on the detection of hate speech in videos using machine learning. The video dataset was created by utilizing a crawler to look for and

download videos that contained inflammatory terms. Training four models with three different feature sets taken from the dataset was used in the experiments. [22] In terms of video classification, the Random Forest Classifier model produced the best results [20]. Researchers' curiosity has been spurred by advancements in the field of machine learning and deep learning, driving them to explore and develop solutions to the problem of hate speech. Machine learning approaches are currently being applied to textual data to detect hate speech [23].

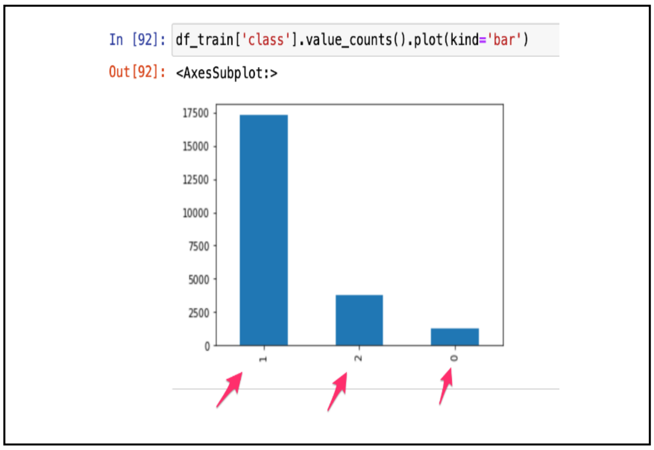
[17] conducted a study on detection of hateful comments on social media. Naive Bayes classifier was found to be 62.75% accurate, but neural algorithm gained an improved accuracy of 87%. The researcher noted that social media usage has grown, but some users abuse the channels by spreading hatred, leading to a lack of an empirical method for detecting, quantifying, and categorizing hateful comments.

METHODOLOGY

This study employed quantitative (epidemiological) and empirical research approaches. The experiment was carried out in order to provide tangible results from the dataset, which were then used to answer the hypothesis. The performance of the algorithms will be assessed by comparing the performance accuracy of each algorithm. CRISP-DM methodology was used and the following major steps were undertaken in the analysis, understanding and preparation of the data [4].


Dataset Preparation
Preprocessing
Data labeling
Training Data
Test Data
Dataset
Feature Extraction & Development
TF-IDF semantic ,Pos tag , sentiment,BOW, Topic modeling etc.
Word embeddings. FastText,GloVe,BERT etc.
Model Learning
Classifier:
Machne Learning: RF,GBDT,SVM,etc.
Deep Learning CNN,LSTM,etc.
Multiclass output
E.G. Cyberbulling Racism Racial
Binary output
Hatespeech Offensive Lanuage

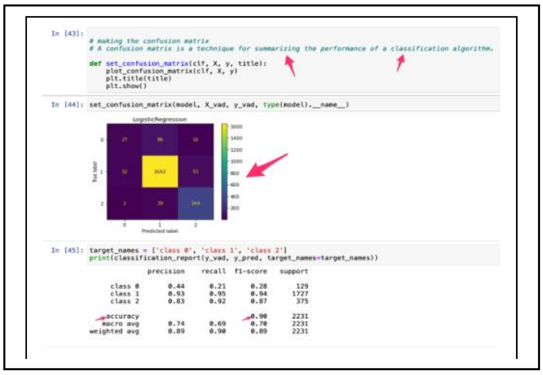
The logic of the detection model is that, once the system is trained, it classifies, label the data and categorise it. The moment the system is run; it is able to produce a report inform of a graph showing different categories [12]. The graph below, shows the classification of data into classes and the number of occurrences per class. From the graph, it can be learnt that the highest was a class of offensive language, while hate speech had the least committed offenses. The other class indicates that though there were some tweets recorded, they did not fall in either of the class.



```
In [92]: df_train['class'].value_counts().plot(kind='bar')
Out[92]: <AxesSubplot:>
```

Training and classifying data sets into classes

To get more information on how well this machine classifier worked. A confusion matrix which is a strong predictive analytic tool in machine learning was used. It is a Predictive analysis software or a comparison table of expected and actual values [10]. A confusion matrix is a statistic used in machine learning to analyze the performance of machine learning classifiers. The confusion matrix was used to depict critical predictive metrics such as remember, specificity, correctness, and precision. the confusion matrix was considered valuable because it was capable of providing straightforward comparisons of variables such as Positive Instances, False Positives, True Negatives, and False Negatives [10]. Other machine learning classification measures, on the other hand, such as "Accuracy," provide less meaningful information because accuracy is a measure of the difference between right and wrong predictions divided by the total number of predictions. The confusion matrix below obtained after running the logistic regression shows the accuracy level of 90% .



```
In [43]:
         # making the confusion matrix
         # A confusion matrix is a technique for summarizing the performance of a classification algorithm.

         def set_confusion_matrix(clf, X, y, title):
             plot_confusion_matrix(clf, X, y)
             plt.title(title)
             plt.show()

In [44]: set_confusion_matrix(model, X_vad, y_vad, type(model).__name__)
```

```
In [45]: target_names = ['class 0', 'class 1', 'class 2']
         print(classification_report(y_vad, y_pred, target_names=target_names))

              precision    recall  f1-score   support

     class 0       0.44      0.21      0.28       129
     class 1       0.93      0.95      0.94      1727
     class 2       0.83      0.92      0.87       375

    accuracy                           0.90      2231
   macro avg       0.74      0.69      0.70      2231
weighted avg       0.89      0.90      0.89      2231
```

Confusion Matrix using Logistic Regression

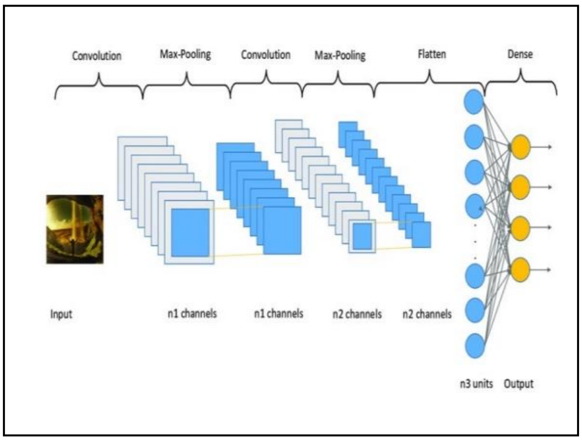
Deep learning is an artificial intelligence (AI) and machine learning (ML) technique that mimics how people learn specific types of information. Deep learning is a key component of data science, along with statistics and predictive modeling [4]. Deep learning is particularly valuable for data scientists who must collect, analyze, and interpret massive amounts of data; deep learning speeds up and simplifies this process [5].

Deep learning can be conceived of as a straightforward method of automating data modeling. In contrast to typical machine learning algorithms, deep learning algorithms are stacked in a structure that makes advantage of variety and abstraction.

Deep Learning has emerged as a feasible method for analyzing vast amounts of data by teaching computers to learn from experience, classify, and recognize data/images in the same way that the human brain does. Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) are two common deep learning algorithms, while other algorithms were also evaluated [8].

Images and objects are identified and classified using Convolutional Neural Networks (CNN). Deep Learning uses a CNN to recognize objects in photos. CNNs can be used for image processing, machine vision applications like as localization and classification, analysis, detecting obstacles in self-driving cars, and speech recognition in natural language processing. CNNs are widely used in Deep Learning because they play such an essential role in these ever-expanding and novel areas.

It is a neural network with many layers that processes data in a grid-like fashion to extract important features. The use of CNNs has the substantial advantage of requiring no image pre-processing [8].



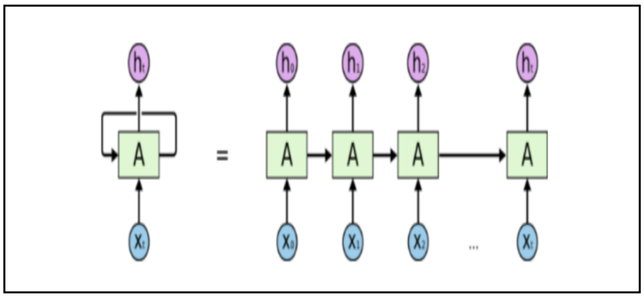
A Convolutional Neural Network (CNN) representation

LSTM - LSTM Recurrent Neural Networks

LSTM Recurrent Neural Networks is a type of neural network that has numerous layers that processes data in a grid-like format to extract significant properties. One significant advantage of employing CNNs is that it does not require prior processing of images .

A feedforward neural network with internal memory is referred to as a recurrent neural network. RNNs are recurrent in nature since they perform the same function for each data input, and the outcome of the input sequence is reliant on the prior computation [1]. The data is then reproduced and returned to the recurrent neural network. When making a decision, it considers both the current input and the outputs learned from the previous input.

recurrent neural networks, unlike machine learning techniques, may handle input sequences using their internal state (memory). As a result, they can be used for tasks like Character recognition that is not segmented, linked, or speech recognition. Many other neural networks' inputs are fully independent of one another. However, with RNN, all of the inputs are connected to one another.



An Unrolled Recurrent Neural Network

It takes X (0) from the input sequence and then outputs h (0), which when combined with X (1) is the entry for the next step. As a result, h (0) and X (1) are the next step's inputs. As a result, h (1) from the previous stage is the input for the following phase, and so on. This way, it keeps the context during training.

The current state's formula is:

$$h_t = f(h_{t-1}, x_t)$$

Using Activation Function:

$$h_t = \tanh\,(W_{hh}h_{t-1} + W_{xh}x_t)$$

W is the weight, h refer to convolutional vector, whh is really the weight from the preceding hidden, whx is the load from the current input, and tanh is the Perceptron, which gives Non-linearity by entirely destroying the range's inputs. [-1.1]

Output:

$$y_t = W_{hy}h_t$$

The output state is Yt. What is the significance of a weight at the output.

Import the text file you prepared. You can now style your work by using the scroll down window on the left side of the MS Word Formatting toolbar.

FINDINGS AND DISCUSSIONS

The outcomes and conclusions of this indicates that all the objectives were achieved. However, while the study accomplished all of its objectives, it encountered some problems that could not be addressed in this study but would be addressed in future related investigations.

According to the first aim, the study indicated that Zambia, like many other countries, has internet venues that are vulnerable to misuse. There have been reports of cyberbullying, harassment, hate speech, and the propagation of false information on online platforms.

Sending threatening or abusive texts, disclosing personal information without consent, or disseminating embarrassing photographs or videos are all examples of cyberbullying and harassment.Such behavior can create emotional distress, worry, and even physical harm. Racist, sexist, homophobic, or xenophobic statements, for example, can promote unfavorable perceptions and contribute to social division and violence. Fake information, or "fake news," can sometimes have serious consequences, such as undermining public trust in institutions and endangering public health and safety.

The government of the republic of Zambia has taken attempts to address internet abuse, such as passing the Cyber Security and Cyber Crimes Act in 2021. This law criminalizes a wide range of internet behaviors, including cyberbullying, harassment, and spreading false information.

Concerns have been expressed about the potential impact of such laws on free expression, as well as the need for clear definitions and procedures to ensure that they are not used to muzzle dissenting voices or limit access to information. Internet platforms are also accountable for resolving abuse on their platforms. Numerous platforms, including machine learning algorithms and human moderators, have policies and processes in place to detect and delete abusive content.

However, the success of these methods varies, and some people have expressed concerns about them.

Therefore, preventing online platform abuse in Zambia requires a comprehensive approach that involves government law, responsible platform policies and practices, as well as individual responsibility and accountability.

In accordance with this goal, the research was effective in developing the detection model. To guarantee that the model was completely functional, it was trained on the English dataset before being applied to the local language dataset. This was due to the fact that training deep learning models with local datasets can present a number of challenges, such as limited, biased data, data privacy, resource requirements, and model maintenance.

However, the efficacy of these systems varies, and there have been concerns raised about the inherent biases and limitations of automatic moderation techniques. Further findings suggested that having numerous language domains can bring several challenges that can impair the accuracy of a deep learning model. A language domain is a specific area or field of usage of language, such as medical terminology, legal terminology, or technical jargon.

CONCLUSION

The study establishes that in fore model of detecting hate speech as well as offensive language on online published content on various platforms, the data set must be categorised and presented in statistical form after running the model [3]; the count indicates the total number of data sets imported. The mean for each category, as well as the standard deviation and the minimum and maximum number of tweets in each category, are also displayed.

For future studies, the study recommends that other sources of information such as Facebook, WhatsApp, Instagram, and other social media platforms, as well as consider harvesting local data sets for training machines rather than relying on foreign data sets; the local data set can then be used to detect offences targeting Zambian citizens on local platforms.

REFRENCES

[1]. K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink &, J. Schmidhuber. LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10), 2222-2232, 2017.

[2]. C. S. McKeever, and S. J. Delany. Abusive text detection using neural networks," in CEUR Workshop Proceedings. vol. 2086, pp. 258–260. 2017.

[3]. D. Sessink. Using Machine Learning to Detect ICT in Criminal Court Cases. University of Twente, The Netherlands, 2018

[4]. T. M. Connolly &, C. E. Begg. Database Systems: A Practical Approach to Design, Implementation, and Management 6th edition. Pearson Education Limited, 2014

[5]. A Karpathy. Convolutional neural networks for visual recognition. Stanford University course notes. This provides a comprehensive introduction to CNNs, including their architecture, training, and applications, 2016

[6]. H. Sameer and K. Brandon. Curtailing cyber and information security vulnerabilities through situational crime prevention. Security Journal, 26(4), 383-402, 2013

[7]. T. J. Holt and A. M. Bossler. Cybercrime in Progress: Theory and Prevention of Technology-Enabled Offenses. Crime Sciences Series. New York: Routledge, 2016

[8]. Mohan, White and Barnes. Countering Hate Speech in Elections: Strategies for Electoral Management Bodies Praeger: Westport, Connecticut, p. 79. 2018. Elections and Electoral Crises in Africa at https://www.ituc-africa.org/

[9]. S. Nahar. Al-Maskari, X. Li, and C. Pang, "Semi-supervised Learning for Cyberbullying Detection in Social Networks," in Databases Theory and Applications, 2014, pp. 160–171.

[10]. C. Goutte &, E. Gaussier. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In European Conference on Information Retrieval (pp. 345-359). Springer, Berlin, Heidelberg, 2005.

[11]. ZICTA Annual Report Annual Report, Advancing the Nation to a Digital Society. December, 2–2, 2020

[12]. V. Behzadan, et al. Corpus and Deep Learning Classifier for Collection of Cyber Threat Indicators in Twitter Stream. IEEE International Conference on Big Data (Big Data), 2018.

[13]. Z. Waseem and D. Hovy. "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," Proc. NAACL Student Res. Work., pp. 88–93, 2016.

[14]. B. Warner and J. Hirschberg. "Detecting Hate Speech on the World Wide Web," no. Lsm, pp. 19–26, 2012.

[15]. C. Wilson, A. Richard and L. Molly. "Hate Speech on Social Media: Content Moderation in Context", 2021. Faculty Articles.

[16]. Unnathi (2019). Detection of Hate Speech in Videos Using Machine Learning. San Jose State University. USA.

[17]. Essa (2022). Detection of Hateful Comments on Social Media. Rochester Institute of Technology.

[18]. Wubetu and Ayodeji (2022). Detection of fake news and hate speech for Ethiopian languages: a systematic review of the approaches. Demilie and Salau Journal of Big Data (2022) 9:66 https://doi.org/10.1186/s40537-022-00619-x

[19]. Zhou, Pete and Hutchings (2022). Automated hate speech detection and span extraction in underground hacking and extremist forums University of Cambridge, Cambridge CB2 1TN, UK

[20]. Aditya, G, Vikrant, D, Shrikant, K, & Laxmi, B. (2018). Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach. Maharashtra Institute of Technology, Pune Pune, India.

[21]. Calvin, S, & Derwin, S. (2020). A Systematic Literature Review of Different Machine Learning Methods on Hate Speech Detection. Bina Nusantara University, Jakarta, Indonesia.

[22]. Duwairi, Hayajneh and Quwaider (2021). A Deep Learning Framework for Automatic Detection of Hate Speech Embedded in Arabic Tweets. King Fahd University of Petroleum & Minerals.

[23]. Aneri and Sonali (2022). Emotion Based Hate Speech Detection using Multimodal Learning.

[24]. Burruss, George W., Bossler, Adam M. And Holt, Thomas J. (2012). Assessing the mediation of a fuller social learning model on low self-control's influence on software piracy. Crime and Delinquency, 59(5), 1157-1184.

[25]. Chen, S. McKeever, and S. J. Delany, "Abusive text detection using neural networks," in CEUR Workshop Proceedings, 2017, vol. 2086, pp. 258–260.