# International Journal of Mathematical Analysis and Modelling

(Formerly Journal of the Nigerian Society for Mathematical Biology)

Volume 6, Issue 2 (Oct.-Nov.), 2023

ISSN (Print): 2682 - 5694
ISSN (Online): 2682 - 5708

International Journal of Mathematical Analysis and Modelling
Volume 6, Issue 2, October-November 2023, pages 265 – 274

# Yoruba language text-based hate detection model for social media

**O.F. Adegoke*†, F.S. Bakpo‡, B. Tenuche†, E. Ukekwe‡, and J. Raiyetunbi†**

**Abstract**

There is an emerging problem of hate speech sharing and spreading associated with Nigerian social media. This work was targeted to develop classification model for text based hate speech in Yoruba language. Support Vector Machine (SVM) was used on crowd sourced datasets comprising 1,420 Yoruba extracted from social media platforms. Annotated 71% Yoruba text datasets were used to train the SVM classifier. The data sets were preprocessed using the Python libraries. The technologies used for the front end are HTML version 5.0, CSS version 3.0, Java script, jQuery, and Bootstrap. The back end involved the use of Laravel version 9.0, a PHP framework, and MySQL. 1008 and 412 datasets were used to train and test the Yoruba SVM classifier, respectively. The confusion matrix indicated that the trained SVM Classifier for Yoruba was able to identify Yoruba hate speech. Accuracy, Precision, Recall, F1-Score and AUC are 75.49%, 75.21%, 0.0954, 0.8410, and 0.60, respectively for the Yoruba SVM Classifier. The model performance compared with other existing models showed that for Yoruba database; Naïve Bayes gave 69.09%, accuracy, and 72.64% precision. Logistic Regression gave 66.51% accuracy, and 68.89% precision. KNN gave 75.52% accuracy, and 86.91% precision. Random Forest gave 54.88% accuracy, and 50.65% precision.

***Keywords***: classification model; confusion metrics; hate speech; support vector machine; yoruba language

## 1 Introduction

The misuse of social media is on the rise, with hate speeches and abusive language spreading on social media and it is assuming a dangerous and alarming dimension in Nigeria. Hate speech is now ubiquitous as a result of social media, which has encouraged the spreading and sharing of hate speech and is gradually becoming part of our society.

There are many definitions of hate speech from Facebook, Twitter, (Davidson *et al.*, 2017; Fatuna and Nunes, 2018 and Bhardwaj *et al.*, 2020). Common to these and other definitions of hate speech is that it is directed at an individual or group and that it is based on some aspect of the individual or group identity.

To solve this problem, we worked on artificial intelligence that can identify hate speech. The model is made up of supervised learning algorithms trained on capturing different aspects of Yoruba

* Corresponding Author; E-mail: adegoke.fo@ksu.edu.ng
† Department of Computer Science, Prince Abubakar Audu University, Anyigba, Nigeria
‡ Department of Computer Science, University of Nigeria, Nsukka, Nigeria

hate speech. Notably it has a deep learning layer that converts a social media post's language into numerical value and accesses this sequence. In this way, the social media post content can be understood by SVM Classifier which helps it decide whether the post is hateful or not. Some of the past approaches with promising results for detecting hate speech in textual content in indigenous languages (Gaikwad *et al*., 2021; Matamoros-Fernandez and Farkas, 2021 and Saini, 2020) but were not on text based Yoruba hate speech detection using support vector machine (SVM).

## 2 The classier

Support Vector Machine (SVM) is a supervised learning algorithm primarily used for classification problems in machine learning. It creates decision boundaries or hyperplane which segregates dimensional space into classes. The hyperplane classifies the comments into either hate or not hate being a binary classification. Hyperplane is a dot in a one dimensional space (Figure 2.0), a land in two dimensional space (Figure 3.0) and a plane in three dimensional space (Figure 4.0). SVM helps to choose the best separating hyperplane with maximum margin Support Vector determines the hyperplane to choose.

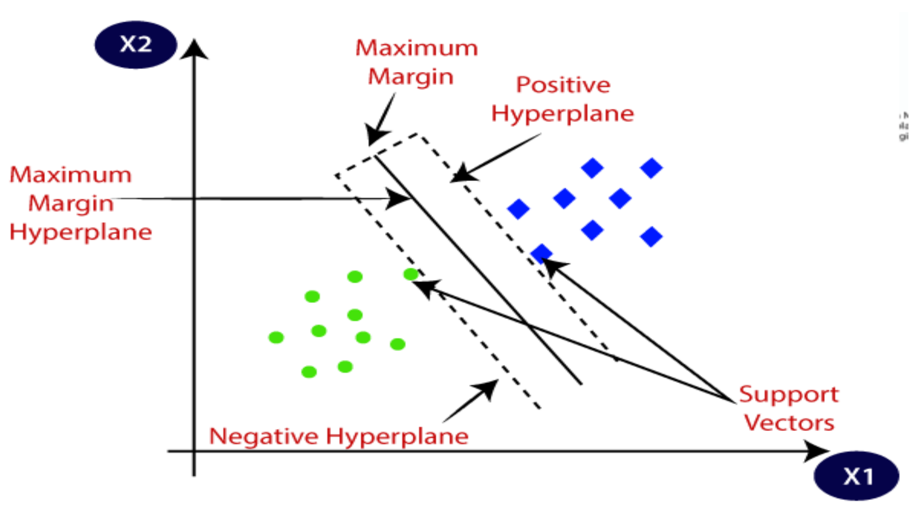


Figure 1: Maximum Margin in SVM (Saini, 2020)

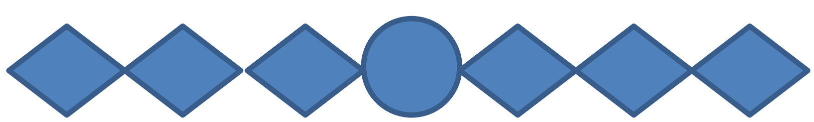

Figure 2: Hyperplane as a dot in one dimensional space

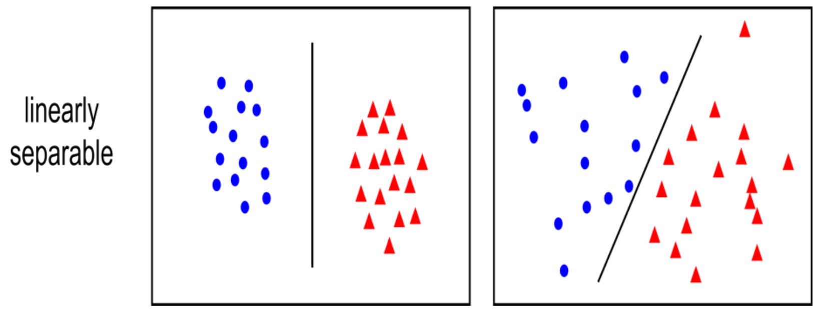


Figure 3: Hyperplane as a land in two dimensional space

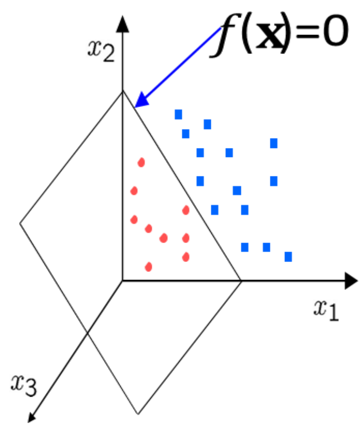


Figure 4: Hyperplane is a plane in three dimensional space

Support Vectors are the data points that lie closest to the hyperplane. They have decision on the linear boundary.

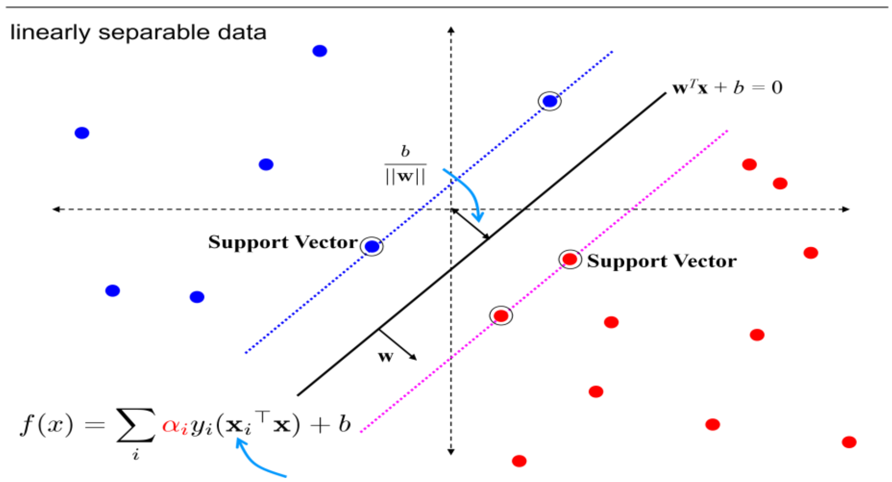


Figure 5: Support Vectors (SVM JavatPoint)

SVM Predicts Outputs

• During inference time, SVM predicts the output Y with input X
• The inputs are set of features x for each sample
• W is the weight vector for each feature

Only support vector determines the weight
Only w is needed to classify new data

Through linear combination of input and weight, output Y is predicted

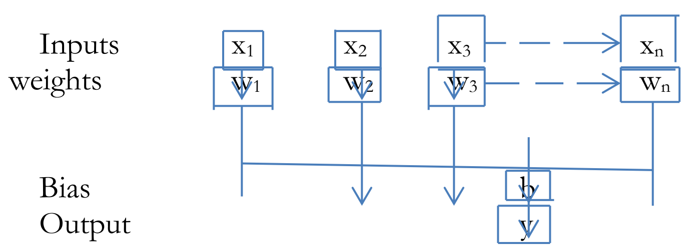


Figure 6: SVM Architecture

Linear Classifier

Given training data $(x_i, y_i)$ for i= 1, 2 …..N, with $x_i \in R^d$ and $y_i \in \{-1, 1\}$,
linear classifier $f$ (x)

such that $f(x_i) \geq 0 \; y_i = +1$, and
$f(x_i) < 0 \; y_i = -1$

• The linear classifier has the form

$$f(x) = w^T x + b$$

• W is known as the weight vector while the b is the bias

• Only w is needed for classifying the unlabeled comments or test data
•The training data is used to learn w and then discarded
- For a new or unlabeled comments, vectorized unlabeled comments would be applied to the classifier equation such that if
    • $w^T x_i + b \geq 1$ it is classified as positive (hate)
    • $w^T x_i + b \leq -1$ it is classified as negative (not hate)

The linear classifier has the form
$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$

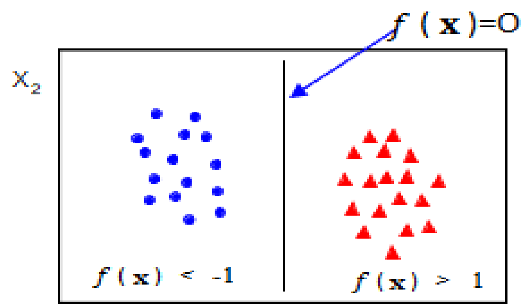


Figure 7.0: SVM Classifier in Binary Classification

$\mathbf{w}^T\mathbf{x}_i + \mathbf{b} \geq 1$ **it is classified as positive (hate)**
$\mathbf{w}^T\mathbf{x}_i + \mathbf{b} \leq -1$ **it is classified as negative (not hate)**

# 3 Methodology

## 3.1 Data sources and collection

Yoruba text comments posted by social media users in response to high conversation on Facebook and twitter social media were the data set source. Yoruba text comments were collected using web scraping in python with twint library and imported into CSV file from Facebook and twitter. The publicly available crowd sourced access was created to be able to copy the content needed by creating secured connection with the source site (my page on face book) where contents were fetched into the document object module (DOM). Value dash pair was used to sort the content of the DOM and data of interest were selected and stored in CSV file. The data set were divided into training and test of data sets using resembling statistical method which is available in scikit-learn.

## 3.2 Data annotation

The Yoruba text data set (71%) were uploaded and distributed via crowd flower into a dedicated web site created from acquired domain for three annotators to manually tag the text comments into either hate or not hate (binary classification). The manually tagged text comments became the training data set used to train the support vector machine (SVM) and optimize the model parameters. The other part that was not tagged was used as test data set 29% for Yoruba language. Every document which contained the text data set was classified into either hate or not-hate.

### 3.3 Data cleaning and preprocessing

Preprocessing of input text can play a key role in text classification by reducing dimensionality and removing unnecessary content (Husain and Uzuner, 2022). Different Preprocessing techniques using Natural language processing (NLP), which is a python library that provides modules for processing text comments in other to filter noisy and non-informative features from the Yoruba text comments. The following were carried out: Removals, replacement, translation, tokenization, lemmatization, stemming, vectorization, data scaling and kernel selection.

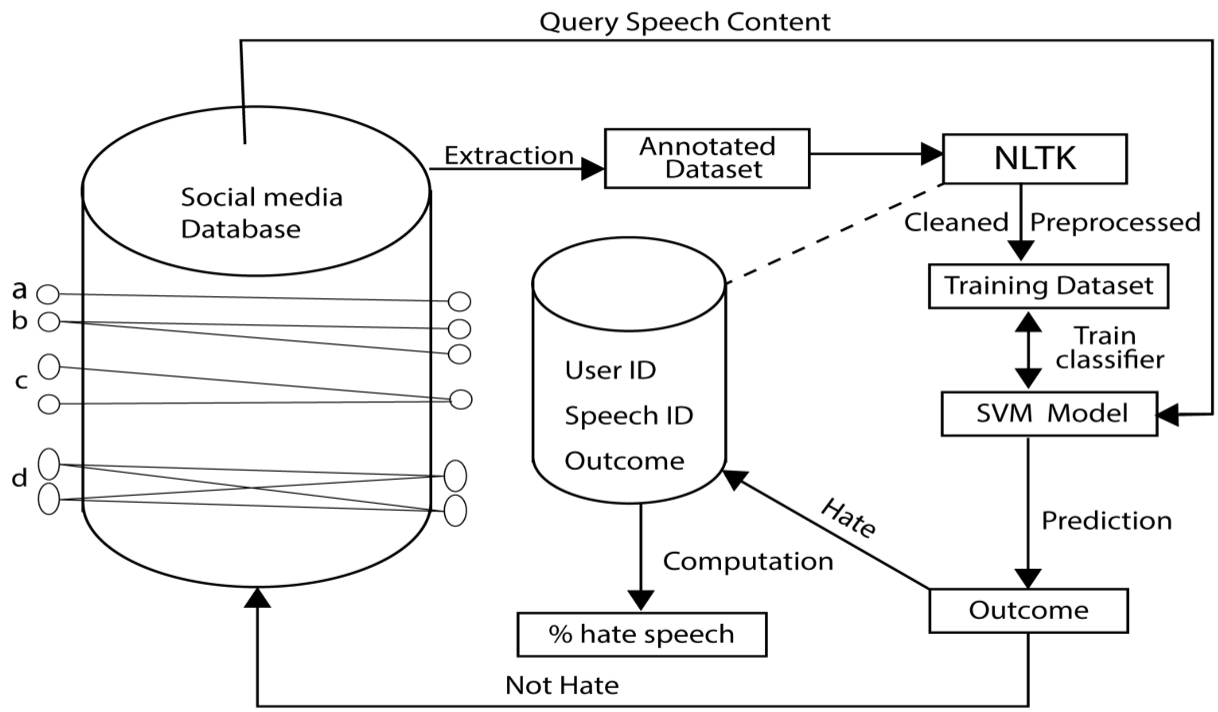


Figure 8.: SVM Classifier for detection of Nigerian indigenous language hate speech in Nigeria Social Media

## 4 Results and discussion

The data set were divided into training (71%) and test (29%) data sets. 732 and 277 data sets were used to train hate and not hate, respectively as shown in Table 1.0 respectively. The data set that was not tagged by annotators (29%) made up of 280 hate dataset and 132 not hate dataset were used as test the SVM Classifier for Yoruba language.

A total of 355 data sets were predicted hate out of which 267 data sets were true hate (true positive) and 88 data sets are not hate but predicted hate (false positive) as shown in Table 2. 13 hate data sets were predicted not hate (false negative) while 44 not hate data sets were predicted not hate (true negative) (Table 2)

Table 1: Yoruba language text Data set Distribution

| Dataset | Hate | Not hate | Row Total |
|---|---|---|---|
| Training | 732 | 277 | 1009 |
| Test | 280 | 132 | 412 |
| Column Total | 1012 | 409 | 1420 |

Table 2: Yoruba Language Data Set Predicted.

| Test Data Labels | SVM Predicted Hate | SVM Predicted Not Hate | Row Total |
|---|---|---|---|
| Hate | 267(tp) | 13(fn) | 280 |
| Not Hate | 88 (fp) | 44(tn) | 132 |
| Colum Total | 355 | 57 | 412 |

Yoruba Language Data Set Confusion Matrix

$$\begin{bmatrix} 267 & 13 \\ 88 & 44 \end{bmatrix}$$

% Trained Yoruba dataset = $^{1008}/_{1420}$ x 100 = 71%
% Test Yoruba dataset = $^{412}/_{1420}$ x 100 = 29%.

Keys:

p = positive (hate)
n = negative (not hate)
tp = true positive (positive and predicted positive)
tn = true negative (negative and predicted negative)
fp = false positive (negative but predicted positive, false alarm, type II error or overestimation)
fn = false negative (positive but predicted negative, missed, type I error or underestimation).

Based on the values in the SVM binary classification confusion matrix the following measurements were carried out. Precision, which calculates total accurate prediction as the percentage of the total instances i.e. % of positives predictions that are correct.

Yoruba SVM model precision = $\frac{\text{tp}}{tp+fr} = \frac{267}{269+88} = 0.7521$,

% of predictions that are correct = accuracy = $\frac{\mathrm{tp}}{tp+tn+fn+fp}$

Yoruba SVM model accuracy = $\frac{267+44}{267+13+14+88} = \frac{311}{412}$ = 0.7549.

Accuracy and precision are not enough to determine if the SVM model is correct. Consequently, F1 score, which is the harmonic mean area of precision, was calculated. In order to calculate the harmonic mean of precision, recall, which is another common metric for classification problem was measured. Recall is also known as sensitivity, it is % of positive cases that were predicted as positive.

Recall = $\frac{\mathrm{tp}}{tp+fn}$.

SVM model recall for Yoruba language = $\frac{267}{267+13} = \frac{267}{280}$ = 0.9536.

$F_1$– score = 2. $\frac{\mathrm{recall.precsion}}{recall+precision}$

SVM model $F_1$ – score for Yoruba = 2. $\frac{0.7172}{1.7057}$ = 0.8410.

The $F_1$– score gives a better and unbiased representation of the performance measurement than accuracy measurement even for unbalanced data.

Misclassification rate of SVM Classifier for Yoruba Language = $\frac{\mathrm{fp+fn}}{tp+tn+fp+fn}$= 0.23

True positive rate (TPR) of SVM Classifier for Yoruba Language = $\frac{\mathrm{tp}}{p}$ = 1- FNR = 0.75

True negative rate (TNR) of SVM Classifier for Yoruba Language = $\frac{\mathrm{tn}}{n}$ = 1- FPR = 0.77

False positive rate (FPR) of SVM Classifier for Yoruba Language = $\frac{\mathrm{fp}}{fp+tn}$ = 1- TNR = 0.67

False negative rate (FNR) of SVM Classifier for Yoruba Language = $\frac{\mathrm{fn}}{fn+tp}$ = 1- TPR= 0.05

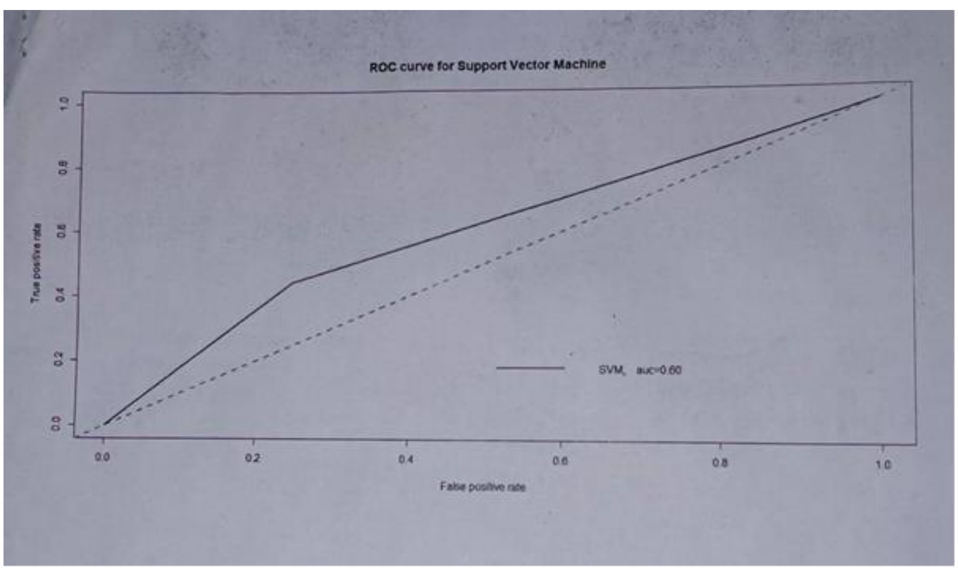


Figure 9: ROC Curve for SVM Classifier for Yoruba

Area under curve (AUC) represents the degree of separability. It tells how much the model is capable of distinguishing between hate and not hate. AUC uses the true positive rates and false positive rates of the models and makes it better in assessment than accuracy. The value of AUC of 0.60 was observed for Yoruba SVM Model (Figure 9.0).

The value of AUC of Yoruba means that the Yoruba model is discriminatory and as such was able to detect Yoruba hate speech. The machine learning classifiers for Yoruba language performed well as shown in Figure 10.0 and other statistical indices.

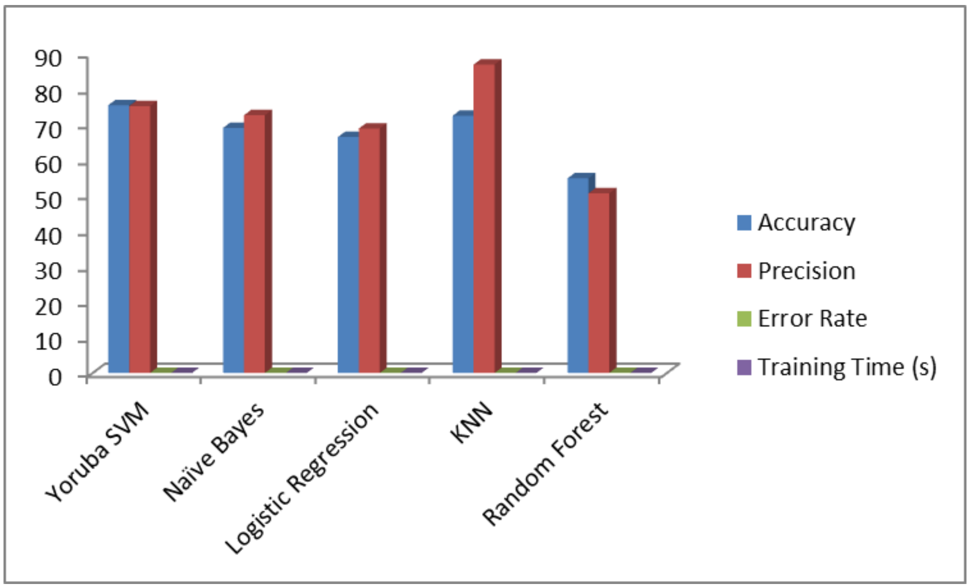


Figure 10: Yoruba SVM Classifier against others

## 5 Conclusion

Support Vector Machine (SVM) classifier developed was able to detect hate speeches in Yoruba text data set. The SVM classifier displayed better performance compared with other classifiers. The deployment of the model as backend could help alleviate or reduce the incidence of hate speech on Nigeria social media in Yoruba languages. Error analysis can be carried out to examine what could have been responsible for wrong prediction of some data sets by the SVM classifier.

## References

[1] Bhardwaj, M., Akhtar, M.S.,Ekbal, A, Das,A. and Chakraborpy, I (2020). Hostility detection dataset in hindi. ACM Transactions on internet Technology (TOIT) 20(2):4-22.

[2] Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Proceedings of the International AAAI Conference on Web and Social Media, 11(1): 512-515.

[3] Fatuna, P. and Nunes, S. (2018). A survey on Automatic Detection of hate speech Text ACM Computer Survey, 51(4):1-30.

[4] Gaikwad, S. Ranasinghe, J., Zampieri, M. and Homan, C.M. (2021). Cross-lingual offensive language identification for low resource languages: The case of Marathi arXiv:2109.03552 https://doi.org/10.48550/arxiv2109.03552 .

[5] Husain, F and Uzuner, O. (2022). Investigating the effect of preprocessing Arabic Text on offensive language and hate speech detection ACM Transactions on Asia and low Resources Language Information Processing, 21(4): 1-20.

[6] Matamoros-Fernandez, A. and Farkas, J. (2021) Racism, hate speech and social media: A systematic review and critique Television and New Media, 20 (2): 205-224.

[7] Saini, A. (2020). Guide to Support Vector Machine Algorithm https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/.

[8] SVM Javat Point https:www.javatpoint.com.>machine-learning-support-vector-machine-algorithm