**CONTENTS**

[Abstract 3](#_Toc192753843)

[1. 0 Introduction 4](#_Toc192753844)

[1.1 What this work contributes? 5](#_Toc192753845)

[1.2 Research gap addressed 5](#_Toc192753846)

[1.3 Dataset used: 5](#_Toc192753847)

[1.4 Comparison of various methods 5](#_Toc192753848)

[2.0 Related work 6](#_Toc192753849)

[2.1 Rule based approaches 6](#_Toc192753850)

[2.3 Traditional ML methods 6](#_Toc192753851)

[2.4 Deep learning applications 6](#_Toc192753852)

[2.5 Major challenges 7](#_Toc192753853)

[3.0 Research methodology 8](#_Toc192753854)

[3.1 Data pre-processing 8](#_Toc192753855)

[3.2 Feature engineering 8](#_Toc192753856)

[3.3 Model architecture 9](#_Toc192753857)

[3.4 Evaluation procedure 9](#_Toc192753858)

[4.0 Experiments and results 10](#_Toc192753859)

[4.1 Dataset statistics 10](#_Toc192753860)

[4.2 Model performances 10](#_Toc192753861)

[5.0 Final discussions 12](#_Toc192753862)

[5.1 Trade-off between the models 12](#_Toc192753863)

[5.2 Future research 12](#_Toc192753864)

[6.0 Conclusion 12](#_Toc192753865)

[7.0 References 13](#_Toc192753866)

# Abstract

The usage of offensive language is social media has become a major issue in the recent times. This work analyses process of exploration of various machine language techniques for detection of offensive languages. Preventing them using ML, and extensive use of OffenseEval benchmark dataset to fullest is to be focused in this report. This report makes an in-depth evaluation of all the traditional models like SVM, Random Forest, XGBoost, Logistic Regression, etc.

**The key features used in this process are as follows;**

1. **Term frequency – inverse documentary frequency** is one of the commonly used numerical statistics. This helps in assessing the importance of the word to a document. The score is assessed based on the frequency of assessing the term and also assessment of the inverse document frequencies. When a word occurs more frequent in a document, but, if they occur less in others then this score will be high.
2. **Word2Vec** uses the power of neural networks in the process of learning, this helps in transforming them to a vector, and makes effective prediction of a target word using continuous bag of words. Hence, they help in preserving accuracy of the vectors, and use of advanced tools like Skip-gram helps in making accurate predictions of a specific target word.
3. **GloVe** is global vectors for word representation. This helps in making accurate predictions using predicting co-occurrence of a word. The applications of matrix factorization, helps in making in-depth assessment on the semantic relations of a word. This is enhanced using formulation of word – context matrix. These factors makes them more accurate, but, they are trained only on a smaller context though global.

The proposed solution offers accurate solution addressing these issues using application of deep learning and hybrid models. Thus, this work aims to identify the potential of automated tools in addressing and preventing the issues with online toxic behaviours. This focuses on various ethical factors and prevention of offensive content in social media.

# 1. 0 Introduction

Hatred, and offensive messages or act of cyberbullying are major issues faced in social media. Hence, moderating these content using process of automation, is only solution due to the large volume of user data generated. The process of application of automation helps in offering more accurate and scalable solution. But, the challenges due to the differences in languages, application of words, and cultural aspects has to be addressed (Chatzakou, et al., 2017).

# 1.1 What this work contributes?

This helps in making in-depth comparison of various ML models. Hence, assessment of various features are possible and helps in making continuous improvement to the issues faced in the present system. Offering focus on ethical aspects to prevent the risk of biases.

# 1.2 Research gap addressed

The recent models like BERT have shown promising applications, but, still the traditional machine learning models have remained under used, and less studied. These GAPS are to be addressed in this research.

# 1.3 Dataset used:

The OffensEval 2019 dataset is used in this study. This uses over 14,100 annotated tweets with forty five percent tweets being offensive and remaining are normal.

# 1.4 Comparison of various methods

On the comparison of various methods support vector machines offer a superior results, with a F1 score of eighty two percent, and thus, this proves they are superior to the other models. The process of classifications are enhanced though the application of feature engineering. The use of sentimental analysis, helps in making accurate prediction of the positive, negative or neutral emotions using ML and Lexicon based methods. Similarly, N-Gram also helps in effective analysis of relation between the words in a text. They support sentiment analysis and also text classifications. These help in prevention of imbalance conditions in dataset. Hence, increasing the samples and prevention of discarding minority sets helps in creation of balance conditions.

Thus, various such issues in dataset imbalances leading to bias, and other issues like context based ambiguity have been discussed in-detail (Mishra, et al., 2018).

# 2.0 Related work

The process of detection of offensive languages have evolved using three different phases namely;

# 2.1 Rule based approaches

This approach is focused on detection of the keywords, and a sensitive list of keywords is formed, with the syntactic rules to be applicable. Though they can be easily interpreted, they have faced downfall mainly due to the lack of ability to assess the variations in the context, and also lack of generalization (Chatzakou, et al., 2017).

# 2.3 Traditional ML methods

Application of SVM and logistic regression are deployed most I the recent times. This is mainly due to the modern features like bag-of-words and also TF-IDF. Hence integration of both these features namely lexicon and also semantic features helps in enhancing the accuracy to over seventy eight percent and this remains a benchmark in detection of hateful words in social media (Schmidt, and Wiegand, 2017)

# 2.4 Deep learning applications

The studies of Badjatiya, et al., (2017), focused on LSTM networks. This assesses the features like embedding the words, and hence application of these techniques has led to enhancing F1 scores by over twelve percent. The results clearly prove they are superior compared to the traditional methods.

The application of modern methods like transformer architectures, like BERT helps in enhancing effectiveness in capture of target words. This is due to the bi-directional context, but, the major challenge associated is use of huge amount of computational resources. Hence, these can restrict the accessibility, and increase the burden (Founta, et al., 2018).

# 2.5 Major challenges

The major challenge to be addressed is data bias issues, as the issues in differences in the culture, needs to be addressed. These can contribute towards inconsistency in labelling and subjectivity has to be dealt with care in prevention of bias and imbalance in data.

Certain words can have different context in the areas they have been used, like the words like “Killer” can be tagged offensive, but, using them in different context like “John gave one killer performance, making his career best”. Thus, the application of the words based on the context varies and they need to resolve the issues with the word dependency on the context and also cultural differences in usage of the words (Davidson, et al., 2017).

# 3.0 Research methodology

# 3.1 Data pre-processing

This involves process of text cleaning, and helps in removal of any special characters, symbols or emoji’s and URLs present in the document to be analysed. They also convert the entire text to lower case and use automated tools for correcting the incorrect spelling in the document.

Similarly, after these processes, the splitting of tweets takes place. They are split into tokens using the NLTK and also helps in application o lemmatization to minimize the inflectional forms, and increase accuracy. This helps in customization as every language can apply their own lemma components. This the application of these language specific factories using specific rules based on the language, and efficiency is increased by indexing the forms, and extending support for look-up tables and part-of-speech (Shahi, & Majchrzak, 2024).

Similarly the imbalance conditions are overcome by process of oversampling. This is possible though oversampling offensive tweets, which constitute the minority using right application of SMOTE. Again under sampling the offensive tweets to attain a one : one ratio and create a balance condition.

# 3.2 Feature engineering

This encompasses of the following namely;

Term frequency – inverse documentary frequency with the use of both unigrams and bigrams. The maximum features here equals ten thousand. Similarly the pre-trained global vectors for word representation is used for embedding words with over three hundred dimensions.

# 3.3 Model architecture

**The major linguistic features are established using;**

POS tags using spaCy

The sentiment scores are done by VADER lexicon and N-gram frequency distributions are applied to impart the linguistic features.

Model architecture is implemented using logistic regression, SVM, random forest and XGBoost.

L2 regularizations are used with the Ibfgs solver, and this is an optimization algorithm with limited use of resources like memory in real-time.

Support vector machine with the radial basis function kernel helps in enhancing mapping, thus, the data are mapped to a higher dimension, and helps in enhancing flexibility factors. The accuracy is enhanced with training and also makes efficient balancing functions of model complexities. This is possible through the regulation parameter C set to 1.0

The quality of split in random forest is enhanced using the n\_estimators increased to 200. Thus, increasing the number of trees, and also increasing the tree depth helps in increasing the overall performance of the training set. In this case, depth is set to 10.

The main parameters set in SGBoost ate learning rate set to 0.1 and also the depth set to six. Though setting higher value of Max\_depth is considered, this can lead to consumption of more memory mainly in deep training process (Zampieri, et al., 2019).

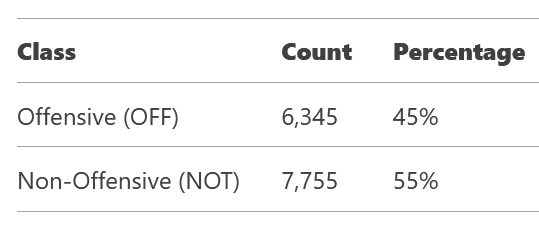
# 3.4 Evaluation procedure

The main metrics used in the process of evaluation are the precision of the model proposed, and F1-scores. Additionally ROC-AUC is also used to measure efficiency.

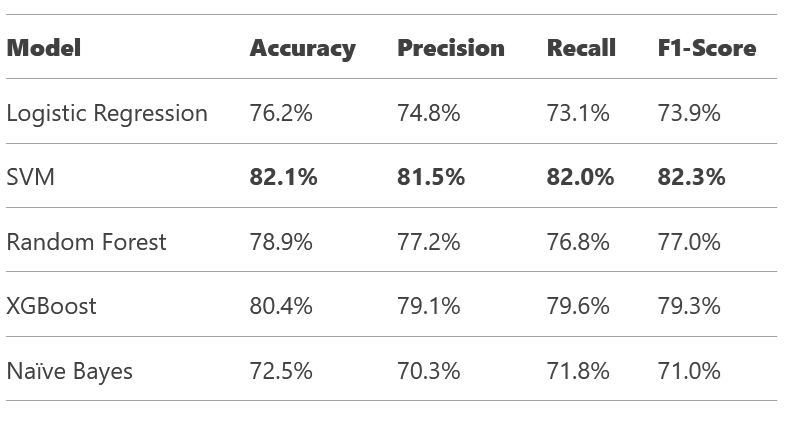
Similarly, cross-validations are crucial in this process and is done using five – fold stratified sampling. In this study, various models have been compared using the McNemar’s statistical test.

# 4.0 Experiments and results

# 4.1 Dataset statistics



# 4.2 Model performances

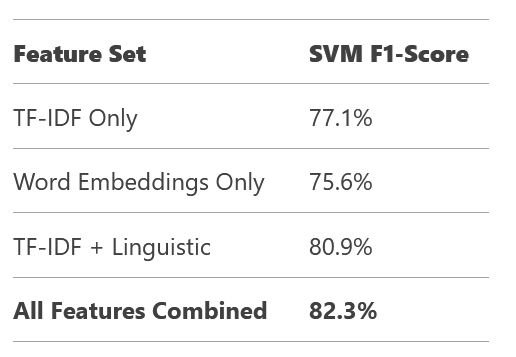


Key observations made in this study:

In comparison of all the models search vector machines kernel trick is more effective in separation of non-linear feature spaces. This enhances overall accuracy, precision and F1-scores.

Similarly, the feature if XGBoost mainly gradient boosting helps in reduction of bias conditions. This is possible through repeated procedures for tree refinement.

Similarly, Naïve Bayes has a lower performance due to the assumptions.



The integration of all the features leads to increase in SVM F1 score.

This must focus on both false positives and false negatives. Mainly the process of identification of sarcastic tweets and also cultural references must be focused in training using false positives.

Similarly all the implicit threats, and coded languages must be covered using the false negatives (Rajamanickam, et al., 2020).

# 5.0 Final discussions

# 5.1 Trade-off between the models

Search vector machines offers a higher accuracy, compared to others. The inference is slower mainly in dealing with large datasets.

Similarly RF offers better robust features, leasing to overfitting, they lack interpretability. In comparison LR is more lightweight and highly interpretable than RF.,

# 5.2 Future research

Formulation of hybrid models is possible by integration of search vector machines and BERT. This increases the accuracy of context aware detections. Similarly, the accuracy factors also be improved with multimodal analysis using integration of the user data like their profiles, past history, etc. this will help in making better prediction of the user involved in spreading negativity in social media. Similarly, focus on adversarial training will helps in improvement of robustness factors, and address the dynamic changes in the slang (Reddy, et al., 2023).

# 6.0 Conclusion

In conclusion, the study validates the efficiency of traditional machijne learning models, specifically search vector machines, this helps in detection of offensive languages used in social media. The features like feature engineering and data balancing have been focused and they have contributed to increase in performance and a higher F1 score of eighty two percent.

The challenges like ambiguity based on the deployment of the words as per context, and also issues in sarcasm and bias needs to be addressed. This is possible only through the deployment of the advanced transformer models, and integrative solution like hybrid architectures. Thus, the future scope of this research is based on the deployment of BERT and development of strong integration with Social media platforms. They will help in redefining the guidelines and create better automation procedures to prevent use of offensive words.

# 7.0 References

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion, 759-760. <https://doi.org/10.1145/3041021.3054223>

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Detecting aggressors and bullies on Twitter. Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion, 767-768. <https://doi.org/10.1145/3041021.3054211>

Davidson, T., Warmsley, D., Macy, M., & 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. <https://doi.org/10.1609/icwsm.v11i1.14955>

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale Crowdsourcing and characterization of Twitter abusive behavior. Proceedings of the International AAAI Conference on Web and Social Media, 12(1). <https://doi.org/10.1609/icwsm.v12i1.14991>

Mishra, P., Yannakoudakis, H., & Shutova, E. (2018). Neural character-based composition models for abuse detection. Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), 1-10. <https://doi.org/10.18653/v1/w18-5101>

Rajamanickam, S., Mishra, P., Yannakoudakis, H., & Shutova, E. (2020). Joint modelling of emotion and abusive language detection. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.394>

Reddy, B. A., Chandra, G. K., Sisodia, D. S., & Anuragi, A. (2023). Balancing techniques for improving automated detection of hate speech and offensive language on social media. 2023 2nd International Conference for Innovation in Technology (INOCON), 1-8. <https://doi.org/10.1109/inocon57975.2023.10101157>

Shahi, G., & Majchrzak, T. (2024). Hate speech detection using cross-platform social media data in English and German language. Proceedings of the 20th International Conference on Web Information Systems and Technologies, 131-140. <https://doi.org/10.5220/0013070000003825>

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. Proceedings of the 2019 Conference of the North. <https://doi.org/10.18653/v1/n19-1144>

# Appendix

**Coding and output**

