VOL. 03 NO. 8,
FEBRUARY, 2024
E-ISSN 3026-8893
P-ISSN 3027-1061

African Scholar Publications & Research International
www.africanscholarpub.com

Journal of Education Research and Library Practice (JERLP)

A FRAMEWORK FOR OPTIMIZING AN ENSEMBLE LEARNING FOR OFFENSIVE TWEET IDENTIFICATION

Nanlir Sallau Mullaha

Federal College of Education Pankshin, PMB1027 Pankshin, Plateau State, Nigeria.

mullakns@gmail.com

Abstract

Online derogatory comments are ubiquitous on social media (SM) and are raising serious concerns. SM data is ridden with high dimensional search space due to noise, redundant features, and non-standardized writing style. These problems lead to high computational cost, longer training time with overall low predictive accuracy in machine learning models. To address these problems, we proposed a framework for optimizing stacked generalized ensemble learning to enhance the model performance. The main components of the framework include feature optimizer (FO), ensemble classifiers, and stratified K-foldCV (skfCV) through stacked generalization ensemble architecture. The ensemble classifiers, FO, and skfCV components make our method stable, computationally efficient with the best performance. Our proposed method was validated using three benchmark datasets. The proposed method outperformed the state-of-the-art results in all the evaluation metrics used in those three articles adopted for comparison.

Introduction

There is a pervasive hate speech problem across different social media networks. Over four billion users share their opinions, feelings, current thoughts, and anything of interest in just a few clicks on social media (SM) [1]. This breakthrough has radically transformed the public communication sphere. The advantages of this era are quite huge, as it has made communication faster, cheaper, and has

also removed the distance barrier among distributed communities and many more. Twitter, Facebook, and YouTube were among the early SM that took reprisal action against Russia during the invasion of Ukraine in February 2022.

Nonetheless, the main drawback of SM remains unsolved over decades despite leveraging the power of machine learning [2]. The disadvantage has devastating effects through sharing of hateful and antagonistic posts created by SM users. Many studies have revealed how people with corrupt minds towards others spread their hateful content targeting minorities, political opponents, disadvantaged people of any sort on SM [3]. It has been widely acknowledged that religion, racism, ethnicity, gender, sexual orientation, politics, are among other topics that trigger hate speech [4]. The SM data has four common problems – noisy, highly skewed class distribution, data sparsity, and high dimensionality problem. Non-standardized writing styles and out-of-vocabulary (OOV) tokens also increase the feature space unnecessarily large and have a multiplicative effect on the error rate. These problems account for the poor performance of classifiers in offensive tweets detection models. A multi-dimensional approach is required to solve these problems.

To find a lasting solution, we proposed a framework for optimizing stacked generalization ensemble learning. Through the principle of the "wisdom of the crowd", the ensemble learning techniques use a weighting and aggregating mechanism to add up the advantages of the weaker learners to compensate for the errors for each learner. The integration of suitable feature selection (FS) technique, skfCV on stacked generalization ensemble will solve the problem associated with the SM dataset. The suitable FS technique serves as a preprocessing algorithm that selects the most informative features to be used for ensemble learning [5]. Most FS approaches are inherently unstable. That is, an FS technique always recreates distinct subsets of features for different training datasets, resulting in varied classification accuracy [6]. The Stratification ensures that each fold contains the same proportion of the labels as that of the initial training set.

FS techniques are classified into two categories: wrapper and filter [7]. The combination of the two methods is possible called the hybrid method. The filter methods make use of independent mining algorithms alongside the general properties of the training dataset. The filter techniques employ the ranking technique as the main criterion to score the features and select them after ordering. The filter technique has been reported to be more simple, efficient, and computationally cost-effective than the

wrapper method [7]. The wrapper technique, on the other hand, works based on the principles of relationship that subsist between the optimal feature subset selected and its relevance to the algorithm performance [8]. Different search algorithms can be used to obtain the subset of variables that minimizes the loss function, which is the classifier's predictive accuracy. The wrapper technique is computationally expensive to use.

The contribution of this study is three-fold:

i. To enhance the performance of the prediction model through the ensemble technique.
ii. To optimize the performance of the model using the SelectFromModel (SFM) FS technique
iii. To apply stratified K-foldCV to address the problem of class imbalance in the SM dataset.

The Proposed Framework

Frameworks generally are applied extensively in computing science, especially in software design. A framework is a set of interrelated layers designed to serve as a guide to building a reusable structure. The proposed framework will help researchers to build an optimized stacked generalization ensemble to mine big data. The essence of the new framework is to ensure that high dimensional feature space in SM texts is always reduced to decrease training time and improve model performance. The researchers tested the framework on three benchmark SM datasets. The proposed framework is depicted in Fig.1

The framework has three main components or sections. A section is added to the stacking ensemble. These sections include optimizer, weaker learner models, and the combiner. Each section will perform a specific task toward improving the model performance.

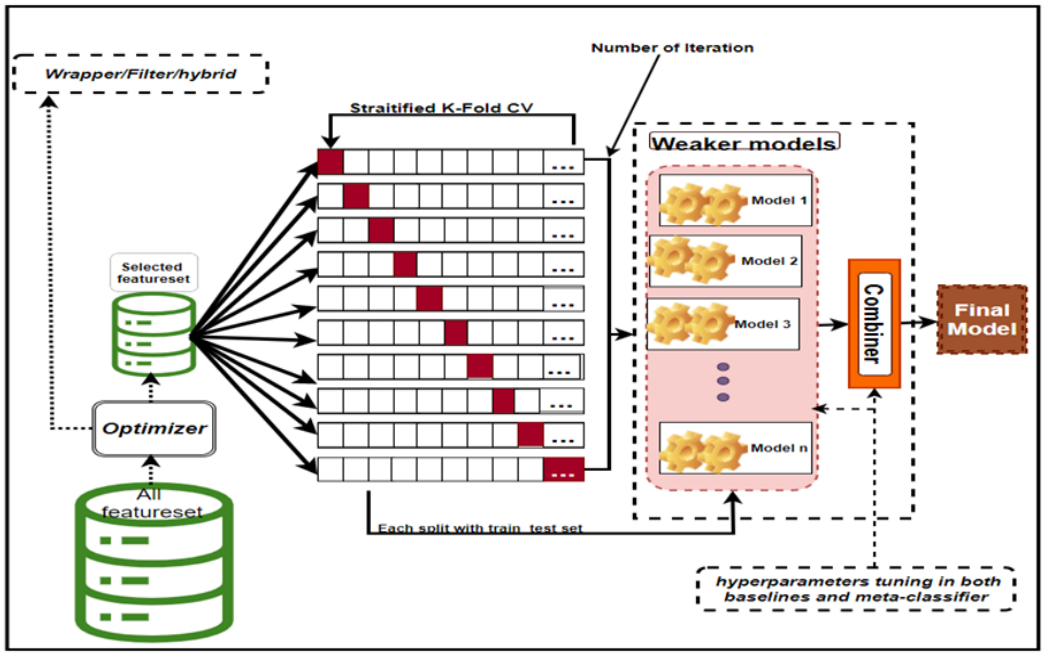


Fig.1 Proposed Framework

Optimizer

This section is to ensure that only informative data is selected in order to solve the problem of high dimensional feature space associated with the SM data. This will help the algorithm to learn the most important aspect of the dataset so that it generalizes well on the unseen data.

Weaker Learner Models

This is where suitable weaker algorithms are arranged in parallel based on the stacked generalization principle. For purpose of this research, we employed ensemble classifiers (boosting ensemble) to serve as the weaker learner models. However, any other learners can be used as the weaker models to build the ensemble following the framework.

Combiner

The output of the weaker learners is used to train the combiner. The combiner must be a linear classifier such as logistic regression (LR), linear support vectors machine (linear SVM), stochastic gradient descent classifiers (SGDC), among others.

Implementation

The proposed framework is implemented through simulation in Python sci-kit-learn framework. Three benchmark datasets were used to test the performance and compare the results for validation. The primary goal of our proposed method is to train a classifier to categorize tweets correctly and efficiently in predefined classes. The essence of integrating FS with stacking ensemble and stratified K-fold technique include – enhancing the performance and robustness of the detection model. Different optimizers were tested and SelectFromModel (SFM) gives the best results as shown in Table 2. The SFM FS creates a subset of the most informative features for training the model. Furthermore, the skfCV technique helps to build a model capable of handling the imbalanced class distribution in SM datasets. A total of three datasets were used for the simulation experiments. The details and descriptions of these datasets are displayed in Table 1.

Table1

Details description of the different datasets used in our experiments.

S/N	Dataset source	Name	Model class	Doc. Size
1	[9]	Mandl	Binary class	7,005
2	[10]	HSE	Binary class	5,876
3	[11]	Davidson	Multi-class	24,783

Ensemble Learning

Ensemble algorithms are a community of learning algorithms with different capabilities and can be combined in parallel or in series to improve classification accuracy [12]. The accuracy of an ensemble classifier must be larger than that of the individual baseline classifier, which is a required and necessary condition and assumption that must be made. This assumption is possible if and only if the errors of the baseline learning algorithms are uncorrelated, that is the errors have no relationship.

The proposed framework is built based on the stacking ensemble. Stacking ensemble is a classifier assembling technique used to aggregate information from various predictive models to generate a more powerful model. Stacked generalization is a 2-stage architecture where the weaker models are arranged in parallel and with combiner models in series. The new framework added another stage called the optimization stage as shown in Fig. 1. The essence of using the ensemble technique

is to accomplish the goal of producing an efficient, reliable, and accurate classification on the training dataset and the best generalization on the testing data. The ensemble techniques improve prediction accuracy by decreasing variance, bias, or obtaining the best trade-off between bias and variance.

Boosting Ensembles

The weaker learners employed are ensemble classifiers or boosted ensemble algorithms such as GBC, RF, XGB, DT, and ABC. Boosting ensemble is a learning strategy that corrects the prediction errors made by previous models to incrementally enhance its accuracy [10]. The boosting ensemble is used as the weaker learner and stacked generalization is employed to build the main classifier.

Model Evaluation Metrics

F1-score and Matthews correlation coefficient (MCC) is the recommended evaluation metrics for the dataset with the imbalanced class distribution [10]. Confusion matrix (CFM) gives in numerical values the performance of the model. It is expressed in terms of true positive (TP), true negative (TN), false positive (FP), and false-negative (FN). TP refers to those data annotated as positive and are predicted as such by the model. FP are those that data that are negative but predicted as positive by the model. TN refers to the dataset coded as negative and predicted as negative by the model. FN as the name implies, the data were predicted as negative, but the prediction was false as they are positive data. The lower the values of FP and FN, the better the model and vice versa. The values TN, FN, FP, and TP can be computed in terms of precision, recall, F1-score, and MCC as shown in Eqs (3) – (6) respectively.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 - score = \frac{TP}{TP + \frac{1}{2}(FN + FN)} \tag{3}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \quad (6)$$

Experimental Results

The effectiveness of the proposed framework was investigated by integrating SFM as an optimizer, boosting ensembles as the weaker models on stacked generalization ensemble to identify hate speech or offensive tweets on SM datasets. The dataset includes both binary and multiclass classifications tasks.

Both the filter and wrapper techniques were employed to identify the suitable algorithm for feature reduction. Particle swarm optimization algorithm (PSOA), SelectFromModel (SFM), cuckoo search algorithm (CKSA), grey wolf algorithm (GWA), and chi-squared technique were among the techniques tested. Our results show that SFM is the best option as displayed in Table 2.

Table 2

FS techniques performances

FS alg.	**Chi-2**	**All features**	**GWOA**	**SFM**	**CKSA**	**PSO**
mcc	0.85	0.84	0.76	**0.86**	0.74	0.73

Table 3 depicted results from the 10-fold CV for the proposed method. The metrics that were used in the previous articles were applied to evaluate the various models in this study for ease of comparison.

Table 3

Evaluation on testing dataset

Fold	HSE		Davidson			Mandl		
Fold	F1	MCC	Pr	Rc	F1	F1-M	F1-W	A
fold0	99	97	95	95	95	85	87	86
fold1	99	97	96	95	95	86	82	81
fold2	99	96	95	95	95	82	85	84
fold3	99	96	95	94	95	84	87	86
fold4	99	95	95	95	95	83	85	84
fold5	100	99	95	95	95	83	86	85
fold6	99	98	96	95	95	77	79	78
fold7	100	98	95	94	95	72	74	73

fold8	99	93	92	92	93	74	75	76
fold9	100	98	95	95	95	65	69	67
Avg	99	97	95	94	95	78	81	80

Comparative Analysis

The datasets used in this research are grouped into multiclass (Davidson) and binary classifications (Mandl & HSE).

Comparison of The Multiclass Classification Task

The versatility of the method was also tested on the multiclass classification model using Davidson's dataset. [13] used the same dataset and applied the BERT-based transfer learning technique. This also allows us to compare the proposed method with the state-of-the-art method. Davidson, Mozafari, and EoEs-SFM are classical ensemble, BERT-based, and optimized ensemble methods respectively. The result of the proposed method is 4% better than Davidson's and 2% better than Mozafari's methods as depicted in Fig. 2.

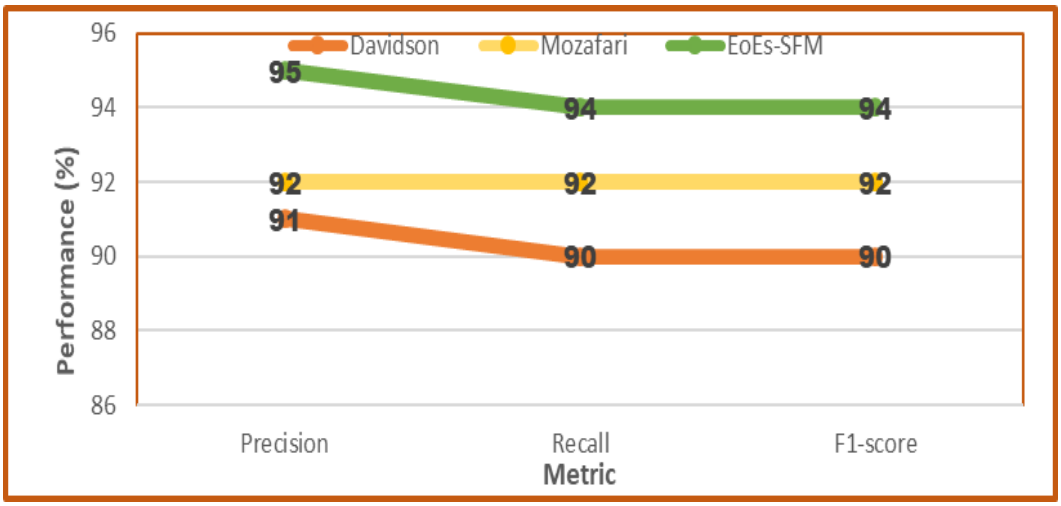


Fig. 2. Multi-Class Results Comparison On Davidson Dataset

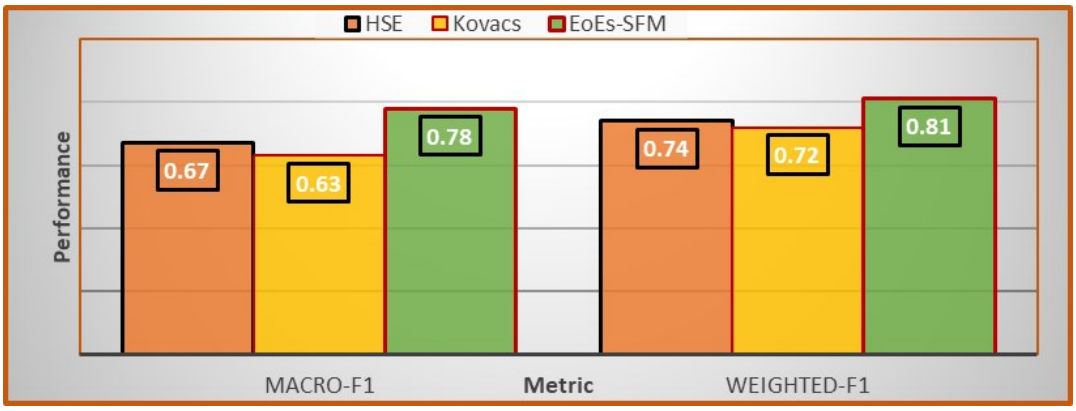


Fig. 3. Binary-class results comparison on Mandl dataset

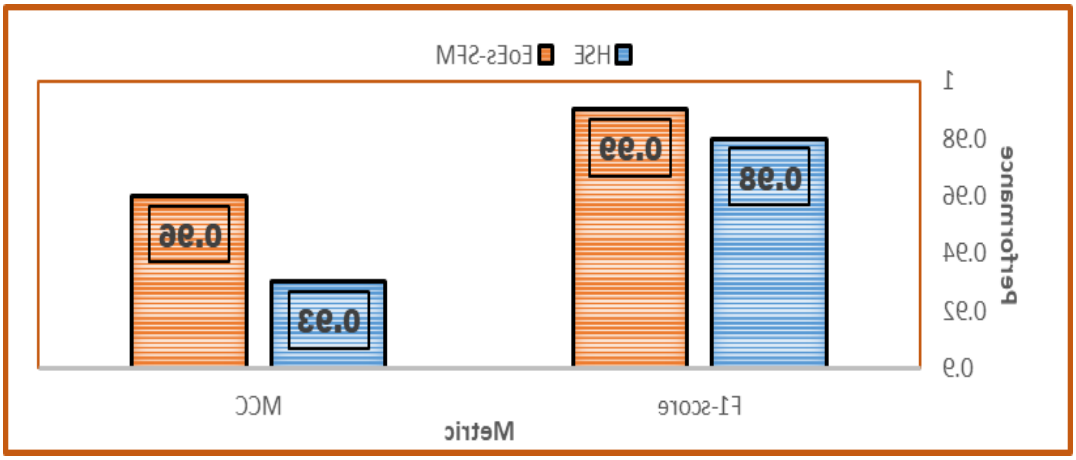


Fig. 4. Binary-class results comparison on HSE dataset

Comparison of The Binary Classification Task

To test for the robustness of the EoEs-SFM method, a binary classification task was used. Two datasets were employed for this comparison: Mandl and HSE datasets and the results are shown in Fig. 3 and Fig. 4.

Kovacs and HSE used F1 macro and weighted averages to evaluate their model. For comparison, EoEo-SFM employed the same in this study. EoEo-SFM is better in both metrics with 0.11 and 0.07 in macro and weighted averages respectively. EoEo-SFM also shows superiority over HSE in MCC with 0.03 better and F1-score with 0.01 higher as shown in Fig. 4.

Conclusion and Future Research Directions

Our proposed method solved the common problems associated with SM datasets: Imbalanced class distribution and improved model prediction generalization on

unseen data. The skfCV helps in reshuffling the dataset to permit the algorithm to learn equally in the class distribution of the training data. The proposed method has proven better on both binary and multi-class classifications tasks.

As seen in Figs. 3 and 4 for the binary classifications, there is a significant improvement in all evaluation metrics. In the case of multiclassification, both classes of interest, hate speech, offensive and neither were improved. This method has proven to be effective and efficient for offensive or hate detection on social media platforms. Further research needs to be conducted in the future to identify the few specific messages that were misclassified and why.

Conflict of interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

P. Singh, R. S. Sawhney, and K. S. Kahlon, "Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government," *ICT Express*, vol. 4, no. 3, pp. 124–129, 2018, doi: 10.1016/j.icte.2017.03.001.

N. S. Mullah and W. M. Na. W. Zainon, "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media : A Review," *IEEE Access*, vol. 9, pp. 88364–88376, 2021, doi: 10.1109/ACCESS.2021.3089515.

A. Tontodimamma, E. Nissi, A. Sarra, and L. Fontanella, "Thirty years of research into hate speech: topics of interest and their evolution," *Scientometrics*, vol. 126, no. 1, pp. 157–179, 2021, doi: 10.1007/s11192-020-03737-6.

L. Vrysis *et al.*, "A web interface for analyzing hate speech," *Futur. Internet*, vol. 13, no. 3, pp. 1–18, 2021, doi: 10.3390/fi13030080.

K. Keykhosravi, A. Hamednia, H. Rastegarfar, and E. Agrell, "Data preprocessing for machine-learning-based adaptive data center transmission," *ICT Express*, 2022, doi: 10.1016/j.icte.2022.02.002.

N. Hoque, M. Singh, and D. K. Bhattacharyya, "EFS-MI: an ensemble feature selection method for classification," *Complex Intell. Syst.*, vol. 4, no. 2, pp. 105–118, 2018, doi: 10.1007/s40747-017-0060-x.

P. H. Prastyo, R. Hidayat, and I. Ardiyanto, "Enhancing sentiment classification performance using hybrid Query Expansion Ranking and Binary Particle Swarm Optimization with Adaptive Inertia Weights," *ICT Express*, 2021, doi: 10.1016/j.icte.2021.04.009.

M. Mafarja and S. Mirjalili, "Whale optimization approaches for wrapper feature selection," *Appl. Soft Comput.*, vol. 62, pp. 441–453, 2018, doi: 10.1016/j.asoc.2017.11.006.

Mandl T, Modha S, Patel D, Majumder P, Dave M, Mandlia C, and Patel A., "Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages.," in *In Proceedings of the 11th forum for information retrieval evaluation*, 2019, pp. 14–17.

N. S. Mullah and W. M. N. W. Zainon, "Improving detection accuracy of politically motivated cyber-hate using heterogeneous stacked ensemble (HSE) approach," *J. Ambient Intell. Humaniz. Comput.*, pp. 1–12, 2022, doi: https://doi.org/10.1007/s12652-022-03763-7.

T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *In Proceedings of the International AAAI Conference on Web and Social Media*, 2017, pp. 512–515.

O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, pp. 1–18, 2018, doi: 10.1002/widm.1249.

M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media," in *International Conference on Complex Networks and Their Applications. Springer*, 2019, vol. 881 SCI, pp. 928–940, doi: 10.1007/978-3-030-36687-2_18.