Language Resources and Evaluation (2024) 58:695–712
https://doi.org/10.1007/s10579-023-09683-y

ORIGINAL PAPER

Comparative performance of ensemble machine learning for Arabic cyberbullying and offensive language detection

Marwa Khairy[1] · **Tarek M. Mahmoud**[3] · **Ahmed Omar**[2] · **Tarek Abd El-Hafeez**[2,4]

Accepted: 23 July 2023 / Published online: 13 August 2023

Abstract

Since cyberbullying impacts both individual victims and entire society, research on abusive language and its detection has attracted attention in recent years. Because social media sites like Facebook, Instagram, Twitter, and others are so widely accessible, hate speech, bullying, sexism, racism, aggressive material, harassment, poisonous comments, and other types of abuse have all substantially increased. Due to the critical requirement to detect, regulate, and limit the spread of harmful content on social networking sites, we conducted this study to automate the detection of offensive language or cyberbullying. We created a new Arabic balanced data set to be used in the offensive detection process because having a balanced data set for a model would result in improved accuracy models. Recently, the performance of single classifiers has been improved using ensemble machine learning. The purpose of this study is to examine the effectiveness of several single and ensemble machine learning algorithms in identifying Arabic text that contains foul language and cyberbullying. Applying them to three Arabic datasets, we have selected three machine learning classifiers and three ensemble models for this aim. Two of them are offensive datasets that are readily accessible in the public, while the third one was created. The results showed that the single learner machine learning strategy is inferior to the ensemble machine learning methodology. Voting performs is the best performing trained ensemble machine learning classifier, outperforming the best single learner classifier (65.1%, 76.2%, and 98%) for the same datasets with accuracy scores of (71.1%, 76.7%, and 98.5%) for each of the three datasets used. Finally, we improve the voting technique's performance through hyperparameter tuning on the Arabic cyberbullying data set.

Keywords OSN · Arabic offensive language · Cyberbullying · NLP · Ensemble machine learning

Extended author information available on the last page of the article

1 Introduction

For many users, online social networks (OSNs) are becoming the most prevalent and interactive media. The majority of individuals use social media without considering the impact these networks have on our lives, whether beneficial or harmful. However, along with valuable and interesting content, these networks can also broadcast inappropriate or harmful content, such as cyberbullying, hate speech, and insults (Khairy et al., 2021; Mironczuk & Protasiewicz, 2018). The detection of such language is essential because it may cause emotional distress and affect the mental health of social media users (Mengü & Mengü, 2015).

The Arabic language is the fifth most spoken in the world, with more than 420 million speakers (http://istizada.com/complete-list-of-arabic-speaking-countries-2014/). The use of the Arabic language in social media is widespread and continually increasing. As of 2017, the Arabic Social Media Report estimates that Facebook users from the Arab region constitute 8.4% of all Facebook users which is more than 150 million Arab users (Salem, 2017). Because the linguistic format might be sophisticated or slang, classifying Arabic social media texts is a difficult task. The Arabic language has multiple dialects with different lexicons and structures, making high-performance classification difficult.

Ensemble Machine Learning is a machine learning methodology that integrates multiple distinct prediction models into a single model to improve performance. It has to be considered whenever good predictive accuracy is demanded (Džeroski et al., 2009). In addition to Ensemble classifiers have been shown to be more effective than data resampling techniques to enhance the classification performance of imbalanced data (Wei et al., 2018). Having a balanced data set for a model would generate higher accuracy models, higher balanced accuracy, and a balanced detection rate. The results obtained by Wei and Dunbrack (2013) demonstrates that using balanced training data (50% neutral and 50% deleterious) results in the highest balanced accuracy (the average of True Positive Rate and True Negative Rate). Consequently, a balanced data set is essential for a categorization model. As a result, we create a fresh collection of balanced Arabic data for the purpose of offensive detection. Additionally, using both balanced and imbalanced datasets, we will employ ensemble machine learning models to find Arabic-language objectionable posts. In order to examine the effectiveness of employing a Single ML Classifier and an ensemble ML in the identification of abusive language and cyberbullying in Arabic, we used a variety of Machine Learning Models and Feature Extraction approaches.

This study has several major strengths that make it a valuable contribution to the field of NLP such as

- Firstly, the topic of cyberbullying and offensive language detection is highly relevant in today's society, given the widespread use of social media and the impact of harmful content on individuals and communities. Cyberbullying can have severe consequences for victims, including emotional and psychological harm, and can also negatively affect the broader society by creating a toxic

online environment. Therefore, research on the detection and regulation of abusive content is essential for promoting safety and well-being on social networking sites.

- Secondly, the study focuses on the Arabic language, which is relatively underrepresented in NLP research despite its global spread and importance. By developing a model for detecting offensive language in Arabic, the study fills a significant gap in the literature and can help to address the issue of harmful content in Arabic online spaces.
- Thirdly, the study created a new balanced data set specifically for Arabic offensive language detection, which is essential for improving the accuracy of models. Having a balanced data set ensures that the model is trained on an equal number of offensive and non-offensive examples, which can prevent bias and ensure that the model is effective in identifying offensive language.
- Fourthly, the study used both single and ensemble machine learning algorithms to compare their effectiveness in identifying offensive Arabic text, providing valuable insights into the strengths and limitations of each approach. Ensemble machine learning involves combining the outputs of multiple models to improve the accuracy of predictions, and the study shows that this approach is more effective than using a single classifier for offensive language detection.
- Finally, the study demonstrates the effectiveness of hyperparameter tuning in improving the performance of the ensemble machine learning classifier, which has practical implications for optimizing the performance of such models in real-world applications. By fine-tuning the parameters of the model on the Arabic cyberbullying data set, the study shows that it is possible to further improve the accuracy of the model and enhance its ability to detect offensive language in Arabic text.

This study makes a significant contribution to the field of NLP by addressing a relevant topic with potential societal impact, focusing on the underrepresented Arabic language, creating a new balanced data set, comparing single and ensemble machine learning algorithms, and demonstrating the effectiveness of hyperparameter tuning in improving model performance.

The remainder of the paper is structured as follows: In Sect. 2, we go over related research. The backdrop for the ensemble machine learning models is covered in Sect. 3. The suggested strategy to detect cyberbullying and abusive language in Arabic is presented in Sect. 4. Finally, Sect. 5 will give the results and discussion.

2 Related work

Research on Arabic abusive language detection has recently drawn much attention; Mubarak et al. (2017) have many efforts in the Arabic language field, especially in Offensive Language detection. They show how to use popular trends in offensive and rude communications to build and extend a list of offensive words and hashtags using an automated tool. Also, Twitter users were ranked based on whether or not they use any of these offensive terms in their tweets. Using this classification, they

expand the list of bad words and present the results on a newly created dataset of labeled Arabic tweets (obscene, offensive, and clean). Also, they publish a large corpus of classified user comments publicly, which were removed from a famous Arabic news site due to rule violations and guidelines of this site. In Mubarak and Darwish (2019), they present a rapidly creating training dataset for identifying offensive tweets using a seed list of offensive words. They trained a deep learning classifier based on charterer n-grams that can efficiently classify tweets with a 90% F1 score. They recently make a new dialectal Arabic news comment dataset public, which was collected from a variety of social media platforms, including Twitter, Facebook, and YouTube. In abusive comments, they investigate the unique lexical content in connection with the use of Emojis. The results show that the data set model of multiplatform news commentary can capture diversity in various dialects and domains. In addition to evaluating the models' generalization power, they also presented an in-depth analysis of Emojis usage in offensive comments. Findings suggest emojis in the animal category are exploited in offensive comments, like lexical observation (Shammur et al., 2020).

Abozinadah et al. (2015) evaluates various machine learning algorithms for detecting abusive accounts with Arabic tweets, the data set for this analysis was collected based on the top five Arabic swearing words, from the total result we ended up having 255 unique users. Naïve Bayes (N.B.) classifier with 10 tweets and 100 features has the best performance with a 90% accuracy rate. An Arabic word correction method was also suggested in Abozinadah and Jones (2016) to tackle internet censorship systems and content-filtering vulnerabilities. This method achieved an accuracy of 96.5%. A statistical learning approach was used in Abozinadah and Jones (2017) to detect adult accounts with the Arabic language in social media. The uses of obscenity, vulgarity, slang, and swear words in Arabic content on Twitter were examined in order to identify abusive accounts. With this statistical method, a predictive precision of 96% was achieved, and the imitations of the bag-of-word (BOW) approach were overcome.

Alakrot et al. (2018) build an Arabic dataset of YouTube comments to detect offensive language in a machine learning context. The data were collected by the principles of availability, variety, representativeness, and balance, thus ensuring that predictive analytical models for identifying the abusive language in online communication in Arabic can be implemented for training. Lately, Bushr et al. (2020) address the issue of abusive language and hate speech detection. They suggest a method for data pre-processing and rebalancing, and then they used the bidirectional Gated Recurrent Unit (GRU) and Convolutional Neural Network (CNN) models. The bidirectional GRU model augmented with attention layer generated the best results among proposed models on a labeled dataset of Arabic tweets were achieved 85.9% F1 scores for offensive language detection and 75% F1 scores for the purpose of detecting hate speech.

Ensemble machine learning is a powerful machine learning algorithm, the result obtained from an ensemble, a combination of machine learning models can be more accurate than any single member of the group (Dietterich, 2000).

Regarding using Ensemble machine learning methods in Arabic offensive language and cyberbullying detection; Haidar et al. (2019) present

Ensemble-Machine-Learning as another solution for Arabic cyberbullying detection to enhance their previous work. They accomplish an improvement on Precision, Recall, and F1-Score. Also, Husain (2020) examined the impact of applying a single learner machine learning approach (SVM, logistic regression, and decision tree) and ensemble machine learning approach (Bagging, AdaBoost, and random forest) on Arabic offensive language detection. The study shows that applying ensemble machine learning techniques rather than single learner machine learning approaches has a significant effect. With an F1 of 88%, which exceeds the best single learner classifier score by 6%, Bagging performs the best among the qualified ensemble machine learning classifiers in offensive language detection.

3 Ensemble machine learning

Ensemble machine learning, a strong machine learning technology that is used by data science experts in industries as it is considered the state-of-the-art solution for many machine learning problems (https://courses.analyticsvidhya.com/courses/ensemble-learning-and-ensemble-learning-techniques). The result obtained from an ensemble, a combination of machine learning models, can be more accurate than any single member of the group (Dietterich, 2000; Džeroski et al., 2009). See Fig. 1.

The three most popular methods for combining the predictions from different models are (https://courses.analyticsvidhya.com/courses/ensemble-learning-and-ensemble-learning-techniques; Brownlee, 2016):

A. *Bagging* Refers to Bootstrap Aggregation, a technique that involves creating multiple samples from the training dataset and training several weak models on each sample. The predictions from the weak models are then aggregated to produce a final prediction, often resulting in improved performance compared to a single model. Examples of bagging models include Bagged Decision Trees and Random Forest.

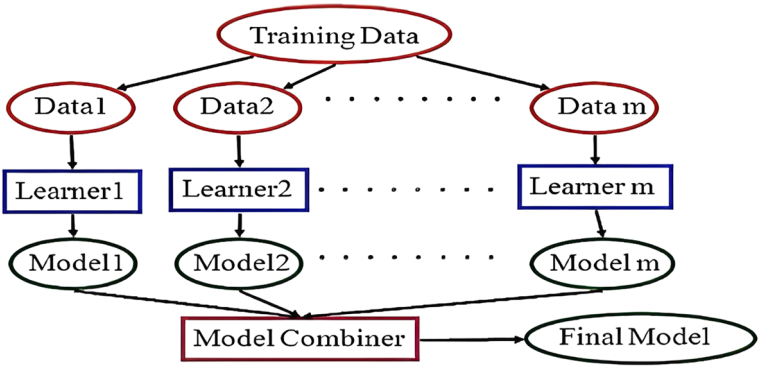

Fig. 1 Ensemble machine learning

B. *Boosting* is a general ensemble technique that combines a number of weak classifiers to produce a strong classifier. This is accomplished by developing a model from the training data, then attempting to fix the faults in the first model with a second model. AdaBoost and Stochastic Gradient Boosting are the two most popular boosting ensemble machine learning techniques.
C. *Voting* is a model that combines the predictions from multiple other and simple statistics are used to combine predictions models to achieve better performance.

4 Proposed method

This section provides a detailed description of the proposed method, including a diagram that illustrates the approach used in this investigation. Additionally, this section outlines the datasets used in the study, the preprocessing techniques employed, the classification strategies implemented, and the performance metrics used to evaluate the models. Regarding the datasets, the study utilized a diverse range of publicly available datasets, including both offensive and non-offensive Arabic text. To ensure that the models were trained on a balanced dataset, a new dataset was also created specifically for this study. Preprocessing techniques were employed to clean and prepare the data for analysis, including tokenization, stop-word removal, and stemming. Multiple classification strategies were implemented for the offensive language detection task, including both single and ensemble machine learning algorithms. The performance of each strategy was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Figure 2 depicts the proposed approach in this study, which involves preprocessing the data, training the models using various classification strategies, and evaluating the performance of the models using standard metrics. The figure provides a visual representation of the steps involved in the proposed method, which can aid in understanding the approach used in this investigation.

4.1 Dataset description

In this paper three Arabic datasets are used, two are offensive datasets that are publicly available at (https://github.com/omammar167/Arabic-Abusive-Datasets) and the third one is a balanced dataset we decide to collect. We created a web crawler to automatically collect the Facebook search results by using a collection of offensive keywords that reflect various forms of offensive language and cyberbullying. We used these keywords to search for tweets on Twitter and posts on Facebook. The tweets produced by the search are collected on Twitter using the Twitter API. Finally, a text file containing all of the gathered tweets and posts is created without duplication. Non-Arabic letters, URLs, and emoticons, for example, have a negative effect on categorization performance. Consequently, it is vital to clean and filter the information gathered. The filtering procedure eliminates all pointless text. We manually awarded a cyberbullying score of 1 and a non-cyberbullying score of 0 to the newly acquired dataset's filtered postings (see the

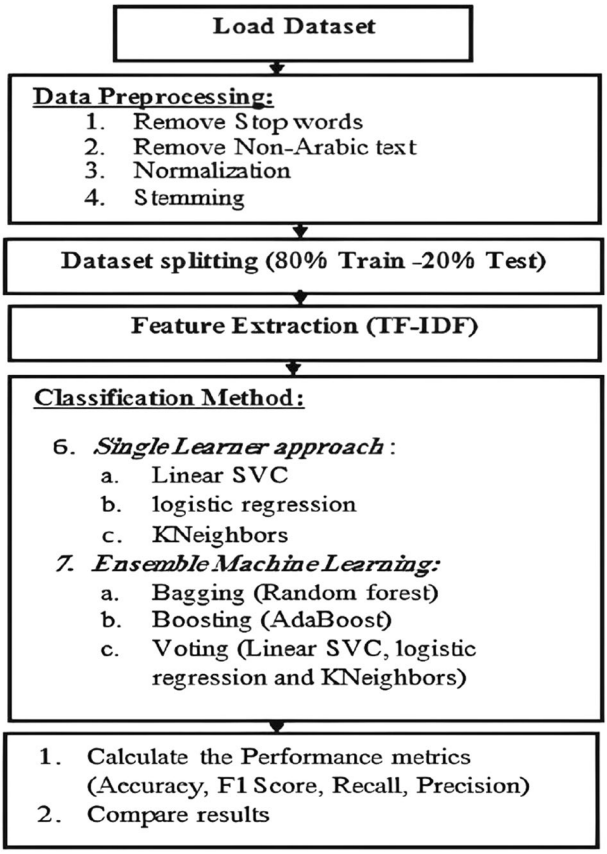

Fig. 2 The proposed method

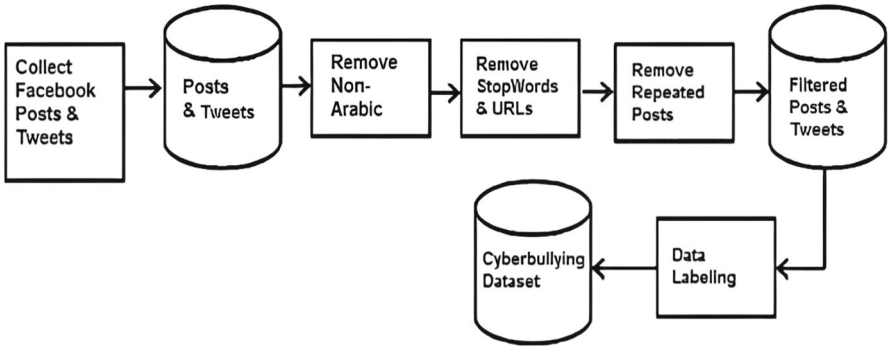

Fig. 3 New Arabic cyberbullying dataset building process

example below). Following preprocessing, the final dataset had 6000 instances of cyberbullying and 6000 instances of non-cyberbullying. Figure 3 shows a new Arabic Cyberbullying dataset building process. It was necessary to tokenize and stem the data when the annotation procedure was finished.

Table 1 lists the original distribution of the datasets that were used, including the size of the dataset, the number of instances of each majority and minority group, and the imbalance ratio for each (IR).

4.2 Classification method

After dividing the dataset into its component parts, we generate the "TF-IDF" and send the features to our classification system, which is based on two strategies (80% for training and 20% for testing). Single Machine Learning Model (SML) is the name of the initial approach. The second model is: Ensemble Machine Learning Model (EML). To evaluate the model's performance, we use a number of indicators (Accuracy, F1, Recall, and Precision). All models are implemented using the Python library Scikit-learn.

4.2.1 Single learner approach (SML)

In this method, we use three single machine learning classifiers: KNeighbors (KNN), Logistic Regression (LR), and Linear Support Vector (Linear SVC).

4.2.2 Ensemble machine learning approach (EML)

For this strategy, we choose three various ensemble machine learning models (bagging, voting, and boosting). The Random Forest model is the one we choose for bagging, and the AdaBoost approach is the one we choose for boosting. The three single individuals from the first method were used for the voting. Three distinct models that each employ a unique ensemble machine learning technique are chosen. A random forest with a maximum of 100 trees was trained. Due to its greater ability to confidently confirm the outcome than separately applied algorithms, hard voting was used to create the ensemble (Alam et al., 2021).

Table 1 The used Arabic datasets

Dataset	Source	Size	Class 0	Class 1	Imbalanced proportion
Arabic1 (Alam et al., 2021)	Twitter	1100	453	647	0.7:1
Arabic2 (Nadali et al., 2013)	YouTube	8577	5332	3245	1.64:1
The proposed dataset	Facebook & Twitter	12,000	6000	6000	1:1

4.3 Performance measures

The metrics measured used to analyze the performance of each classifier are Accuracy, Precision, Recall, and F1_score. The definitions of those metrics are as follows:

A. *Accuracy* The accuracy is the percentage of instances that were correctly classified into their respective classes. It is also called sample accuracy.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

B. *Precision* Precision is used to measure the exactness of the classifier. Moreover, it refers to the fraction of predicted positive which are actually positive. The formula for precision is the number of positive predictions divided by the total number of positive class values predicted.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{2}$$

C. *Recall* Refers to the fraction of those that are actually positive that were predicted as positive. The formula for the recall is the number of positive predictions divided by the number of positive class values in the test data.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{3}$$

D. *The F-measure (or F-score)* Is used to measure the accuracy of the test by considering both precision and recall in computing the score. It conveys a balance between precision and recall wherein it reaches its best value at 1 and its worst value at 0.

$$\text{F1\_Score} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{4}$$

5 Experimental results and discussion

This section presents the results of the experiments, which were carried out to compare the performance of different single and ensemble machine learning algorithms in detecting Arabic messages containing cyberbullying and abusive language.

Table 2 and Fig. 4 have shown the performance metrics for single and ensemble ML models for the first Arabic dataset. From Table 2 we can notice that Voting outperforms all used single and ensemble classifiers in all performance metrics. Logistic Regression (LR) outperforms the other single classifiers with an Accuracy of 65.1%. After that Linear SVC and KNeighbors achieved an Accuracy of 64.2%.

Table 2 Comparative analysis of single and ensemble ML classifiers performance for Arabic1 dataset

Arabic1 dataset						
Performance metrics	Single learner			Ensemble learner		
	Logistic regression (%)	KNeighbors (%)	Linear SVC (%)	Bagging (Random Forest) (%)	Voting (%)	Boosting (Adaboost) (%)
Accuracy	65.1	64.2	64.2	60.5	71.1	56.4
F1_score	62.7	64.1	63.4	58.6	70.9	56
Recall	65.1	64.2	64.2	60.5	71.1	56.4
Precision	64.3	64	63.4	58.4	70.9	64.6

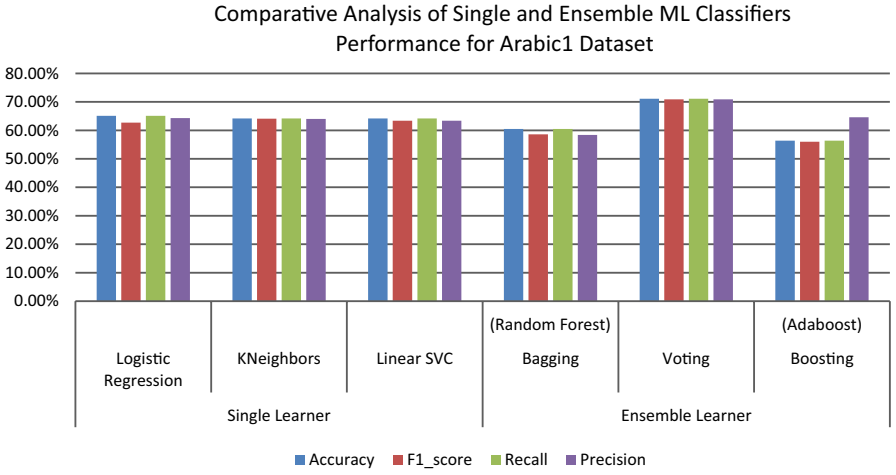

Fig. 4 Comparative analysis of single and ensemble ML classifiers performance for Arabic1 dataset

Table 3 Comparative analysis of single and ensemble ML classifiers performance for Arabic2 dataset

Arabic2 dataset						
Performance metrics	Single learner			Ensemble learner		
	Logistic regression (%)	KNeighbors (%)	Linear SVC (%)	Bagging (Random Forest) (%)	Voting (%)	Boosting (Adaboost) (%)
Accuracy	76.2	75.4	76.1	75.3	76.7	73
F1_score	74.4	74.5	**75.2**	72.8	**75.8**	70.2
Recall	76.2	75.4	76.1	75.3	76.7	73
Precision	76.5	75.3	75.8	77.2	78	74.4

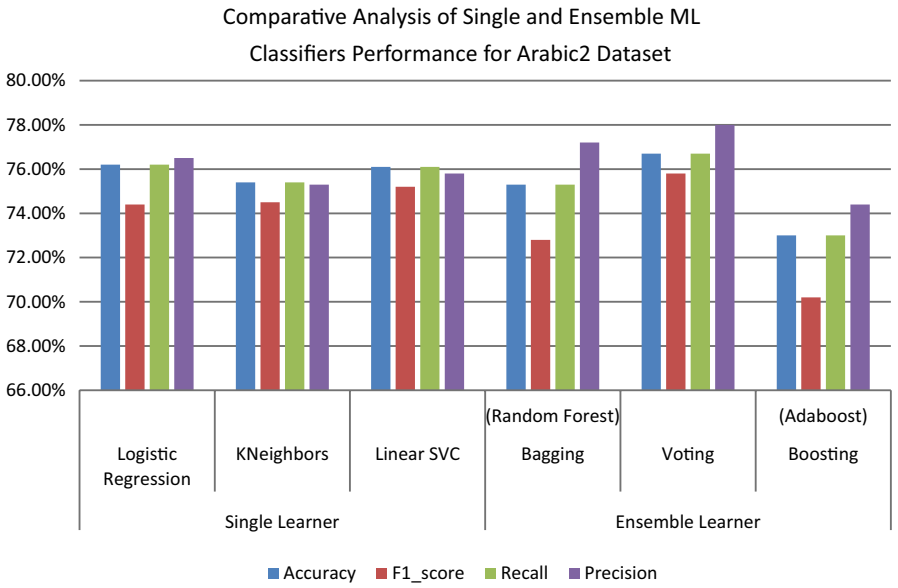

Fig. 5 Comparative analysis of single and ensemble ML classifiers performance for Arabic2 dataset

The Voting EML classifier performs better than the other EML classifiers, with an Accuracy of 71.1%. Next, the Bagging EML and AdaBoost EML's accuracy both hit 60%. The accuracy rises from 65.1% using the best SML classifier to 71.10% using the best EML model, proving the efficiency of ensemble machine learning methods. Our results are inconsistent with Husain (2020).

For the second Arabic dataset, performance metrics for single and ensemble ML models are provided in Table 3 and Fig. 5. Table 3 shows that Voting performs better than all single and ensemble classifiers for all performance metrics. Table 1 shows that this dataset is somewhat unbalanced, therefore rather than utilizing accuracy, which can be deceptive, to measure performance, we will use the F1_score.

Table 4 Comparative analysis of single and ensemble ML classifiers performance for new Arabic dataset

New dataset						
Performance metrics	Single learner			Ensemble learner		
	Logistic regression (%)	KNeighbors (%)	Linear SVC (%)	Bagging (random forest) (%)	Voting (%)	Boosting (Adaboost) (%)
Accuracy	97.4	96.3	98	96.1	98.5	94.8
F1_score	97.2	96.2	97.8	96	98.3	94.5
Recall	97.4	96.3	98	96.1	98.5	94.8
Precision	97.5	96.4	98.2	96.2	98.7	95.1

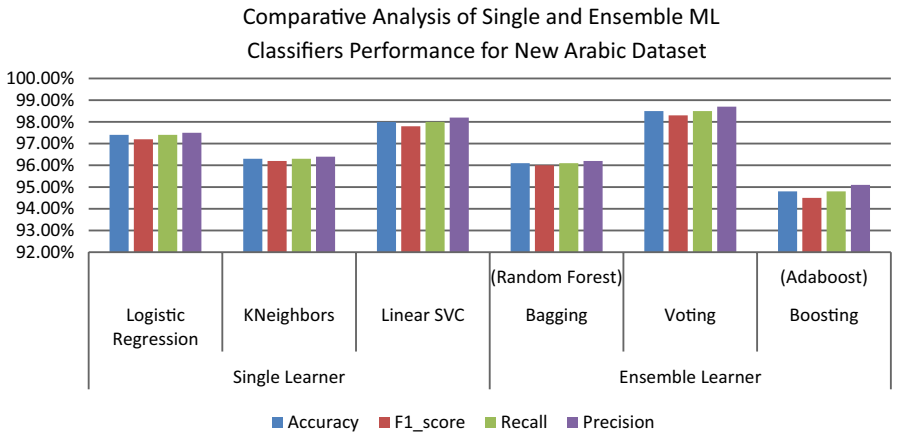

Fig. 6 Comparative analysis of single and ensemble ML classifiers performance for new Arabic dataset

With an F1 score of 75.2%, the SML classifier (Linear SVC) surpasses all other SML classifiers, followed by the KNeighbors (KNN), which achieved a score of 74.5%, and the Logistic Regression (LR), which achieved a score of 74.4%. With an F1 score of 75.8%, the voting classifier outperforms the other EML. Following that, the Bagging EML scored 72.8% on the F1 score, and the Boosting EML scored 70.2%. These results demonstrate the effectiveness of ensemble machine learning techniques, as the F1 score increases from 75.2% using the best SML model to 75.8% using the best EML model. Our results are inconsistent with Husain (2020).

Table 4 and Fig. 6 shows the performance metrics for single and ensemble ML models for the new Arabic dataset. Our new dataset was balanced so we will consider the Accuracy for comparing the results.

From Table 4 we can notice that Voting outperforms all used single and ensemble classifiers in all performance metrics. Linear SVC classifier outperforms the other SML Classifiers with an Accuracy of 98. After that the Logistic Regression (LR) achieved an Accuracy of 97.4% then KNeighbors (KNN) achieved Accuracy of 96.3%. The voting classifier outperforms the other EML Classifiers with an accuracy of 98.5%. After that, the Bagging achieved an accuracy of 96.1%. % then the Boosting achieved an accuracy of 94.8%. These findings prove the efficiency of using ensemble machine learning methods; Accuracy rises from 98% using the best SML model to 98.5% using the best EML model. Our results are inconsistent with Husain (2020).

Table 5 Best parameters obtained by hyperparameter tuning

Classifier	Parameter	Value
KNeighbors	n_neighbors	8
Logistic regression	C	20
Linear SVC	C	20

Our results somewhat confirm the idea that ensemble models outperform single classifiers by default, but not always. In some circumstances, a single classifier called Logistic Regression can produce better results than an ensemble classifier (Bagging and Boosting). It depends on the properties of the studied data set.

5.1 Hyperparameter tuning

We use hyperparameter tuning in the best results produced to improve the prior results, which were obtained by utilizing the default settings for each classifier (Voting). Table 5 shows the best parameters obtained by hyperparameter tuning for three different classifiers: KNeighbors, Logistic Regression, and Linear SVC. The table lists the specific parameter that was tuned and the corresponding value that resulted in the best performance for each classifier and Table 6 presents a comparison of the best performance obtained by hyperparameter tuning and the default values for the selected classifier. The comparison is based on standard performance metrics, including accuracy, F1-score, recall, and precision.

For the KNeighbors classifier, the best value for the n_neighbors parameter was 8. This means that the model performed best when considering the eight nearest neighbors to a given data point when making predictions. For the Logistic Regression classifier and the Linear SVC classifier, the C parameter was tuned, with a value of 20 resulting in the best performance for both models. The C parameter controls the trade-off between maximizing the margin (separating hyperplane) and correctly classifying the training data.

The classifier was trained on an Arabic cyberbullying dataset, and the results show that hyperparameter tuning improved the performance of the model compared to the default values. The best accuracy obtained after hyperparameter tuning was 98.6%, which is a slight improvement over the default value of 98.5%. Similarly, the F1-score, recall, and precision metrics also showed a slight improvement after hyperparameter tuning, with the best F1-score, recall, and precision values being 98.4%, 98.6%, and 98.8%, respectively. These values represent a marginal improvement over the default values of 98.3%, 98.5%, and 98.7%, respectively. Table 6 highlights the effectiveness of hyperparameter tuning in improving the performance of the selected classifier for detecting offensive language in Arabic text. The slight improvement in performance may seem small, but it can have significant practical implications in real-world applications, where even small improvements in accuracy can have a substantial impact on outcomes.

Table 6 Comparison between best performance obtained by hyperparameter tuning and default values

Classifier	Tuning (%)	Default (%)
Accuracy	98.6	98.5
F1_score	98.4	98.3
Recall	98.6	98.5
Precision	98.8	98.7

6 Discussion and future work

The results of this study demonstrate that ensemble machine learning techniques can improve the performance of single learner classifiers in the detection of offensive language and cyberbullying in Arabic text. Specifically, the Voting ensemble machine learning classifier achieved the highest accuracy scores for the three datasets used in this study. These results are consistent with previous studies that have shown that ensemble machine learning techniques can improve the performance of single learner classifiers.

One of the reasons why ensemble machine learning techniques outperform single learner classifiers is that they combine the strengths of multiple models while mitigating their weaknesses. However, it is important to note that the performance of these models is dependent on the quality and diversity of the data used for training. Therefore, it is essential to create larger and more diverse datasets to improve the accuracy and robustness of models in offensive language detection tasks. Furthermore, it would be interesting to investigate the effectiveness of ensemble machine learning techniques for detecting offensive language and cyberbullying in other languages and cultural contexts.

In addition to ensemble machine learning techniques, there are also other machine learning approaches that can be explored for offensive language detection in Arabic text, such as deep learning. Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown promising results in natural language processing tasks, including offensive language detection. Therefore, it would be interesting to compare the performance of deep learning models with ensemble machine learning techniques in detecting offensive language and cyberbullying in Arabic text.

Future work There are several avenues for future work based on the findings of this study. One potential direction is to investigate the effectiveness of ensemble machine learning techniques for detecting offensive language and cyberbullying in other languages and cultural contexts. Offensive language and cyberbullying can manifest differently in different languages and cultures, and models trained on Arabic text may not perform as well on text in other languages. Therefore, it is essential to explore the generalizability of these models to other languages and cultural contexts.

Another potential direction for future work is to investigate the effectiveness of deep learning models for detecting offensive language and cyberbullying in Arabic text. Deep learning models have shown promising results in natural language processing tasks, and it would be interesting to compare their performance with ensemble machine learning techniques in the context of offensive language detection.

Finally, it is important to address the broader societal and ethical implications of automated offensive language detection. While such tools can be useful in regulating harmful content on social media, they can also be used to stifle free speech and silence marginalized voices. Therefore, it is essential to consider the potential unintended consequences of automated detection and to ensure that any

tools developed are used responsibly and ethically. Future work should focus on developing methods to mitigate these risks and ensure that automated detection is used in a responsible and ethical manner.

7 Limitations

While the study has several strengths, there are also some limitations and challenges that should be considered such as

- One major weakness is the reliance on machine learning algorithms to detect offensive language and cyberbullying. While these techniques have shown promising results, they are not infallible and can still produce false positives or false negatives. Additionally, machine learning algorithms are only as good as the data they are trained on, and biases in the data can lead to biased or inaccurate models.
- Another challenge is the limited scope of the study, which focuses solely on Arabic text and may not generalize to other languages or cultural contexts. Offensive language and cyberbullying can manifest differently in different languages and cultures, and models trained on Arabic text may not perform as well on text in other languages.
- Additionally, while the study created a new balanced data set for offensive language detection, the size of the dataset is relatively small, which can limit the generalizability of the results. Creating larger and more diverse datasets can help to improve the accuracy and robustness of models in offensive language detection tasks.
- Finally, the study does not address the broader societal and ethical implications of automated offensive language detection. While such tools can be useful in regulating harmful content on social media, they can also be used to stifle free speech and silence marginalized voices. It is essential to consider the potential unintended consequences of automated detection and to ensure that any tools developed are used responsibly and ethically.

While the study has several strengths, including its focus on a relevant topic, the creation of a new balanced data set, and the comparison of single and ensemble machine learning algorithms, there are also several challenges and limitations to consider. These include the reliance on machine learning algorithms, the limited scope of the study, the small dataset size, and the need to consider broader societal and ethical implications.

8 Conclusion

In order to improve the performance of a single learner classifier, a meta-learning machine learning technique known as ensemble machine learning integrates predictions from several single learner classifiers. In this study, we investigate

how three single learner machine learning approaches (Linear SVC, Logistic Regression, and K Neighbors) and three ensemble machine learning approaches (Bagging-Random Forest, Voting, and Boosting-Adaboost) affect the detection of offensive language and cyberbullying for the Arabic language. The impact of the ensemble machine learning methodology is better than that of the single learner machine learning strategy. Voting, one of the trained ensemble machine learning classifiers, succeeds at detecting offensive language and cyberbullying, with accuracy scores of 71.1%, 76.7%, and 98.5% for the three datasets used, respectively. This score surpasses that of the best single learner classifier, which only achieved accuracy scores of 65.1%, 76.2%, and 98% for the same datasets. To maximize the performance of the voting approach, we applied hyperparameter adjustment, which increased the accuracy to 98.6%.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB). Not applicable.

Data, material and code availability https://github.com/omammar167/Arabic-Abusive-Datasets.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

Abozinadah, E. A., & Jones, J. H., Jr. (2016). Improved micro-blog classification for detecting abusive Arabic Twitter accounts. *International Journal of Data Mining & Knowledge Management Process, 6*(6), 17–28.

Abozinadah, E. A., & Jones, J. H., Jr. (2017). A statistical learning approach to detect abusive Twitter accounts. In *Proceedings of the international conference on computing data analysis—ICCDA '17* (pp. 6–13).

Abozinadah, E. A., Mbaziira, A. V., & Jones, J. H., Jr. (2015). Detection of abusive accounts with Arabic tweets. *International Journal of Knowledge Engineering, 1*(2), 113–119.

Alakrot, A., Murray, L., & Nikolov, N. S. (2018). Dataset construction for the detection of anti-social behavior in online communication in Arabic. *Procedia Computer Science, 142*, 174–181.

Alam, K. S., Bhowmik, S., & Prosun, P. R. K. (2021). Cyberbullying detection: An ensemble based machine learning approach. In *2021 third international conference on intelligent communication technologies and virtual mobile networks (ICICV)* (pp. 710–715).

Brownlee, J. (2016). *Machine learning mastery with Python* (Vol. 527, pp. 100–120). Machine Learning Mastery Pty Ltd

Bushr, H., Zoher, O., Anas, A., & Nada, G. (2020). Arabic offensive language detection with attention-based deep neural networks. In *Proceedings of the 4th workshop on open-source Arabic corpora and processing tools* (pp. 76–81).

Dietterich T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems. MCS 2000. Lecture notes in computer science* (Vol. 1857). Springer. https://doi.org/10.1007/3-540-45014-9_1

Džeroski, S., Panov, P., & Ženko, B. (2009). Machine learning, ensemble methods in. In R. Meyers (Ed.), *Encyclopedia of complexity and systems science.* Springer. https://doi.org/10.1007/978-0-387-30440-3_315

Haidar, B., Chamoun, M., & Serhrouchni, A. (2019). Arabic cyberbullying detection enhancing performance by using ensemble machine learning. In *International conference of Internet of Things* (pp. 323–327).

https://github.com/omammar167/Arabic-Abusive-Datasets

Husain, F. (2020). Arabic offensive language detection using machine learning and ensemble machine learning approaches. *ArXiv Preprint*. https://arxiv.org/abs/2005.08946

Khairy, M., Mahmoud, T. M., Abd-El-Hafeez, T., & Mahfouz, A. (2021). User awareness of privacy, reporting system and cyberbullying on Facebook. In A. E. Hassanien, K. C. Chang, & T. Mincong (Eds.), *Advanced machine learning technologies and applications. AMLTA 2021. Advances in intelligent systems and computing.* (Vol. 1339). Springer. https://doi.org/10.1007/978-3-030-69717-4_58

Mengü, M., & Mengü, S. (2015). Violence and social media. *Athens Journal of Mass Media and Communications, 1*, 211–228.

Mironczuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications, 106*, 36–54. https://doi.org/10.1016/j.eswa.2018.03.058

Mubarak, H., & Darwish, K. (2019). Arabic offensive language classification on twitter. In: *International conference on social informatics* (pp. 269–276). Springer.

Mubarak, H., Darwish, K., & Magdy, W. (2017). Abusive language detection on Arabic social media. In *Proceedings of the first workshop on abusive language online. Vancouver, Canada* (pp. 52–56).

Nadali, S., Murad, M., Sharef, N., Mustapha, A., & Shojaee, S. (2013). A review of cyberbullying detection: An overview. In *13th international conference on intelligent systems design and applications, Bangi* (pp. 325–330).

Retrieved February 16, 2021, from http://istizada.com/complete-list-of-arabic-speaking-countries-2014/

Retrieved June 2, 2021, from https://courses.analyticsvidhya.com/courses/ensemble-learning-and-ensemble-learning-techniques

Salem, F. (2017). *The Arab social media report 2017: Social media and the Internet of Things: Towards data-driven policymaking in the Arab world*. MBR School of Government.

Shammur, A., Hamdy, M., Ahmed, A., Soongyo, J., Beard, J., & Joni, S. (2020). a multi-platform arabic news comment dataset for offensive language detection. In *Proceedings of the 12th conference on language resources and evaluation (LREC 2020)* (pp. 6203–6212) Marseille, 11–16.

Wei, F., Wenjiang, H., & Jinchang, R. (2018). Class imbalance ensemble learning based on the margin theory. *Applied Sciences, 8*, 815. https://doi.org/10.3390/app8050815

Wei, Q., & Dunbrack, R. L., Jr. (2013). The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS ONE, 8*(7), e67863. https://doi.org/10.1371/journal.pone.0067863

Authors and Affiliations

Marwa Khairy[1] · Tarek M. Mahmoud[3] · Ahmed Omar[2] · Tarek Abd El-Hafeez[2,4]

✉ Marwa Khairy
Marwa.kh.mohamed@mu.edu.eg

Tarek M. Mahmoud
tarek@fcai.usc.edu.eg

Ahmed Omar
ahmed.omar@mu.edu.eg

Tarek Abd El-Hafeez
tarek@mu.edu.eg

1 Faculty of Computers and Information, Minia University, El Minya, Egypt

2 Computer Science Department, Faculty of Science, Minia University, El Minya, Egypt

3 Faculty of Computers and Artificial Intelligence, University of Sadat City, Sadat City, Egypt

4 Deraya University, El Minya, Egypt