Journal on
Artificial Intelligence

Tech Science Press

https://doi.org/10.32604/jai.2026.086037

ARTICLE

Detecting and Classifying Abusive Content on Social Media Data Using Multi-Classifier Models Based on Ensemble Strategies

Aline Z. Tsague[1,*], Benoît Martin A. Quimatio[2], Brice Armel D. Namou[1] and Elie Tagne Fute[1,2]

[1]Department of Computer Engineering, University of Buea, Buea, Cameroon
[2]Department of Computer Science, University of Dschang, Dschang, Cameroon

*Corresponding Author: Aline Z. Tsague. Email: tsague.aline@ubuea.cm

Received: 22 May 2026; Accepted: 30 July 2026; Published: 12 October 2026

ABSTRACT: The massive rise of user-generated content (UGC), in particular on social media networks, has led to a steadily increase amount of abusive content, hence the need for moderation. This is due to the proliferation of social media platforms, which have empowered their users to freely create, share, and exchange content of various natures for interacting and communicating with each other. However, this accessibility has inadvertently provided a forum for offensive content, posing significant risks to user safety and business continuity. While users' sensitivity is being hurt and the feeling of stigma, insecurity, distrust, and skepticism is created, businesses' image and reputation on the other hand, are tarnished. This paper addresses this issue by seeking to build a strong multi-classifier model that can efficiently detect and classify such content. We propose two ensemble models built upon three base classifiers, namely Multinomial Naïve Bayes (MNB), Linear Support Vector Machines (LSVM), and Logistic Regression (LR). These models employ stacking and a hybrid combination of bagging and stacking strategies, resulting in the Ensemble Stacking-Voting (ESV) model and the Ensemble Bagging-Stacking (EBS) model. ESV and EBS were trained and tested on a constructed dataset to detect and classify five different classes of abuse that are considered in our classification task. The dataset construction involved collecting and labeling data from four different datasets to address the unavailability of a well-annotated dataset. Experimental results demonstrate a clear performance progression from lightweight individual models to ensemble-based methods. While MNB achieves high computational efficiency with an accuracy of 86.4%, margin-based classifiers such as LSVM and LR significantly improve performance, reaching accuracies of 89.6% and 89.2%, respectively, and achieving strong class-level F1-scores, particularly for hate speech (91%) and abusive content (93%). Ensemble approaches further enhance classification robustness, with the proposed EBS attaining the best overall performance, achieving 92.6% accuracy and consistently high F1-scores (≥90%) across all classes compared with the individual classifiers and the ESV model. Notably, substantial gains are observed for semantically overlapping classes such as *Others* and *Sexism*, both reaching F1-scores of 94%.

KEYWORDS: Abusive content; ensemble learning; machine learning; multi-classifier; stacking strategy; bagging strategy; social network

1 Introduction

During the past decade, the use of social media platforms such as X (formerly Twitter), Facebook, YouTube, and Instagram has experienced a sharp increase because of many users joining these platforms daily and sharing their updates [1]. Several users spend a considerable amount of time on these social media platforms carrying out various activities and expressing their views and beliefs, yielding a vast source of text data [2]. For example, the current monthly active Facebook users are estimated at 2.9 billion [3,4],

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

while the daily active users of X are reported to hundreds of millions [4,5]. Almost every second, there is an uncontrolled number of tweets, posts, and comments posted on a social media platform, making it difficult to track or govern the content of such sites [1]. Social media today is a hotbed for curbing hate speech, as suggested by several studies. The sheer ease with which hateful utterances can be made, typically from the comfort of one's home, and the lack of any immediate negative repercussions, using today's digital communication technologies, are responsible for their significant increase and global ubiquity [6].

Based on statistics, there is evidence of the occurrence of abusive content on social media platforms [7]. Abusive language refers to information that contains mean/dirty/spam words or phrases and commonly comes from displeasing circumstances such as mental illness, sexual divergence, physical disability, or the absence of modernization [8,9]. Detecting abusive content in online data is not just a moderation task; it is a core security function. It protects users from emotional, physical, and digital harm. It also protects systems from abuse, manipulation, and loss of trust, while ensuring platforms remain constructive spaces for communication and interaction within society. In recent years, there has been a growing interest in this topic from various groups, including social scientists who aim to understand the reasons behind abusive online behavior, as well as politicians who have recognized that a substantial portion of public debates and social discourse takes place online. Likewise, the perception of some concepts, politicians, elections, and civil rights movements can be influenced through targeted social media marketing campaigns. As a matter of fact, our current online environment, including news and communication platforms, holds an unprecedented level of social, political, and economic importance.

This situation creates a plethora of challenges with regard to the key question of how to best address the sensitivity and relevance of online media and online content by technological means while at the same time avoiding the establishment of a centralized infrastructure that can be misused for the purpose of censorship or surveillance. One challenge is to separate high-quality content from offensive, hateful, abusive, or massively biased content. While journalists have traditionally been responsible for these responsibilities, they are increasingly being transferred to the end user of online content. This means that the analysis, processing, and evaluation of information are no longer solely the domain of professional news editors or journalists, the burden of fact-checking is now increasingly placed on the reader. Therefore, the automatic detection of abusive content on social media needs serious attention, and it is important to highlight what happens. This work aims at the systematic design, adaptation, and empirical evaluation of stacking and hybrid bagging-stacking ensemble models tailored to a multi-class abusive content classification task. Considering that most publicly available abusive language datasets, predominantly sourced from X, are limited to binary or three-class annotations, we aim to construct a unified tweet-based dataset from some publicly available sources to support five-class abusive content classification, and to address class imbalance via random sampling techniques to obtain a balanced dataset suitable for the task. This data preparation effort enables a more reliable evaluation of ensemble strategies under realistic and challenging conditions. Within this challenging setting, we demonstrate consistent performance gains using stacking and hybrid bagging-stacking ensembles, providing practical and reproducible insights into ensemble design for fine-grained abusive language detection.

Overall, this work addresses a critical gap in abusive language detection research by bridging the gap between coarse-grained abusive language datasets and the practical need for fine-grained classification, while providing actionable insights into ensemble model design under realistic and challenging conditions.

The rest of this paper is organized as follows: Section 2 provides a review of previous research works, exploring various terminologies that are utilized in abusive content classification, as well as techniques employed by state-of-the-art implementations. Section 3 presents our approach and describes the methodology used for detecting and classifying abusive content. Section 4 showcases the results obtained through this

methodology. Section 5 is dedicated to the conclusion of the work by summarizing the findings, addressing any challenges encountered, and suggesting avenues for future research.

2 Related Review

Due to the massive rise of user-generated web content, in particular on social media networks, the amount of abusive content is also steadily increasing. The automatic detection of abusive content is a challenging task due to disagreements on different abusive content definitions. Moreover, some content might be hateful to some individuals and not to others, based on their concerned definitions. In this section, a review of common terminologies in this field of study and the exploration of existing works is provided.

2.1 Abusive Content Detection

In the literature of abusive content detection several research works have contributed to this field with the aim to lay the foundation for the effective detection of abusive content. Consequently, a significant number of foundational requirements for abusive content detection have been established by these pioneering studies. In [10], Schmidt and Wiegand conducted a survey on hate speech detection in which they discuss that the amount of online hate speech increases given the steady rise in UGC. A description of key areas that have been explored to automatically recognize such utterances using natural language processing is provided for clear understanding of application areas. X and Instagram datasets were used in their research and SVM was found to be the most used model. Moreover, the authors emphasized on the importance of well labeled corpora and discussed how datasets are sampled and annotated. In [11], a general typology for detecting abusive contents was proposed. The authors also suggested that high-quality datasets are important, particularly for more nuanced expressions of abuse [11]. Based on a survey on Automatic Detection of Hate Speech in Text [12], in which authors explored existing methods, it was found that most research papers are focused on hate speech and that X was the most widely-studied platform. Yet, they had a poorly annotated dataset. Based on a study carried out in [13], several works done on the field of hate speech and abusive language detection are racial bias in all datasets, as classifiers trained on them tend to predict that tweets written in African-American English are abusive at substantially higher rates. If these abusive language detection systems are used in the field, they will therefore have a disproportionate negative impact on African-American social media users. The authors concluded that to avoid underfitting, datasets need to be well trained [13,14].

On the other hand, some researchers attempted to detect abusive content by training some models on some data and achieved some results. In [15], an ensemble deep learning technique, which lies beyond the scope of this work, for detecting abusive content is proposed. The authors combined Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), then trained the combined model on Kaggle competition dataset and succeeded to obtain an accuracy of 98.46%. Using the Automated Hate Speech Detection and the Problem of Offensive Language (HASOC) 2021 dataset in three different languages, English, Hindi and Marathi, an Ensemble-based Machine Learning Model for Hate Speech and Offensive Content Identification was proposed in [16]. The authors used an ensemble of Random Forest (RF), Multi-Layer Perceptron (MLP) and Gradient Boosting (GB) to detect abusive content and were able to obtain an F-measure of 82.5%.

Recent research on abusive content detection has advanced in multiple directions, including enhanced feature representation and context-aware classification. Mnassri et al. [17] propose an emotion-aware shared encoder that integrates external emotional features into transformer-based classifiers for hate speech and offensive language detection. By jointly learning emotional and abusive content representations, their multi-task approach improves detection performance and reduces false positives compared to single-task baselines,

particularly across multilingual settings using models such as BERT and mBERT. In a different line of work, Mednini et al. [18] focus on the detection of brand-related hate speech using natural language processing techniques. Their study examines both traditional machine learning and deep learning models, demonstrating the utility of combining lexical and contextual features for identifying hate speech in brand-related social media commentary. This work underscores the need for domain-specific modeling strategies when addressing nuanced or context-dependent forms of abusive content.

These studies highlight ongoing efforts to move beyond basic binary classification by leveraging richer feature representations and specialized target domains. However, they are typically constrained to limited class schemes or specific content areas, motivating the present work's focus on comprehensive multi-class abusive content classification with systematic evaluation of ensemble learning strategies.

2.2 Abusive Content Classification

For an automatic classification of abusive language and personal attacks in various forms of online communication, using NLP technologies such as Bayes, Bayes Expectation Maximization, Multivariate Logistic Regression, Maximum Entropy and Winnow2, the work in [6] used datasets from X and Wikipedia in English and German and could classify abusive contents into racism, sexism, hate speech, personal attacks and aggression. The authors used BoW for feature extraction and obtained F-score of 81.58%. The accuracy obtained using by Bayes, Bayes Expectation and Logistic Regression were 84.61%, 84.01% and 85.67% respectively for the English Twitter dataset.

For classifying abusive comments using various machine learning algorithms, the work in [15] focused on data from X to classify abusive content as toxic, severely toxic, obscene, insult, threat and identity hate. The authors carried out feature extraction and could obtain a weak prediction model by applying Logistic Regression, Multinomial Naïve Bayes, and Random Forest techniques. Gradient Boosting was applied to this model to obtain the final prediction model that had an accuracy of 69.7%. A neural network was also applied using bi-directional Long Short-Term Memory to obtain an accuracy of 97.8%. Then using the MNB and LR techniques respectively, an accuracy of 60.4% and 67.9% was obtained.

In [2], a classification approach based on DistilBERT is proposed in which hate speech and offensive content is classified abusive content as HATE, OFFENSIVE, PROFANE and NONE. The authors employed the HASOC dataset and used ML techniques such as LR, Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), RF, LSTM and DistilBERT. The best result was obtained using DistilBERT, followed by LSTM with an accuracy of 61% and 60% respectively. Overall, ML techniques did not perform well and the authors justified this situation by the fact that the model could not properly differentiate between profane, hateful and offensive posts.

Maximilian Wich et al., in [19] developed an abusive language detection model leveraging user and network data to improve the classification performance. The work considered three datasets in two languages (i.e., English and German) and fine-tuned DistilBERT model to which the explainable AI framework SHAP (SHapley Additive exPlanations) was integrated to alleviate the general issue of missing transparency associated with deep learning models. This enabled the assessment of the model's vulnerability toward bias and systematic discrimination reliably. An average F1 score of 90% was obtained from the implementation.

For identifying hate speech and offensive content, Ref. [16] proposed ensemble-based machine learning models built on HASOC 2021 dataset in three different languages namely English, Hindi and Marathi. The authors used an ensemble model by stacking RF, MLP and GB models to classify abusive contents as Profane, Hate or Offensive. Using the English corpora, they obtained an F-measure of 57.7%.

Overall, there exist a variety of techniques for classifying abusive contents, ranging from classical Machine Learning Techniques to Transformers, passing through Deep Learning Techniques. For each of these categories, there are several different ways in which a classification model could be built provided that its performance should be assessed. This is often done using criteria such as accuracy, precision, recall, F1-score, and the model's ability to identify different levels of abuse. Accuracy measures the overall correctness, while precision and recall focus on identifying true positives and avoiding false positives and false negatives, respectively. The F1-score provides a balanced measure of both precision and recall. Additionally, the model should be evaluated on its ability to classify abusive content into different levels of severity (e.g., mild, moderate, severe).

Recent advances in abusive content detection have focused not only on improving classification performance, but also on addressing annotation challenges and expanding the class granularity of abusive language. Walsh and Greaney [20] highlight the importance of multi-class hate speech classification and propose an aggregated dataset re-annotated into fine-grained categories (e.g., ethnicity, gender, religion, sexuality, non-hate), demonstrating that deep learning models such as LSTM and contextual models (e.g., BERT) can benefit from richer label schemes over traditional binary classifiers.

Beyond dataset construction, researchers have investigated the impact of contextualized and content-based features on abuse detection. The research study [21] examines how incorporating contextual information (e.g., preceding tweets) alongside content features can improve abuse detection on Twitter, revealing that contextual modeling addresses limitations of isolated content classifiers in identifying nuanced abusive expressions.

Another emerging area involves multi-task learning and user information incorporation. The authors in [22] propose a unified multi-task learning architecture that leverages intra- and inter-user information with deep neural network models (CNN, GRU, BERT, ALBERT), showing significant gains in macro-F1 and weighted-F1 across several benchmark hate speech datasets. These works demonstrate the field's shift toward *multi-class and context-aware models*, motivating the present study's focus on systematic dataset construction and ensemble model evaluation for fine-grained abusive content classification.

Despite the advancements in the detection and classification of abusive content, several limitations exist within the current body of literature. While earlier researches have made strides in addressing some aspects of abusive content, the unresolved issues of multi-class classification and the need for comprehensive, up-to-date datasets underscore the necessity of further inquiry. The current literature still requires efforts to adequately address how to build sophisticated models capable of capturing the complexity of abusive language across different contexts, thus highlighting the research gap that this study seeks to fill. In the next section, we will be developing two models based on Ensemble Learning (EL).

3 Methodology

3.1 Preliminaries

EL is a machine learning paradigm that consists of building a strong learning model by combining multiple models known as base models with the aim of achieving more accurate results. Ensemble models have been used in many fields for detection using ML. By combining multiple base models, ensemble methods can leverage their diverse perspectives to make more accurate and reliable predictions. The authors in [23] emphasized that EL significantly contributed to the improved performance of their model which consist of combining four supervised ML algorithms, and employing two ensemble techniques. In our experiment, we combined three popular supervised ML algorithms tailored for multi-class target labels using Bagging and Stacking strategies. The development of a multi-classifier model for the detection and

classification of abusive content, as proposed in this study, is grounded in the critical insights gleaned from the reviewed literature. The major gaps and limitations identified in the literature require a methodological approach that embodies enhancement of classification capability, utilization of ensemble learning, addressing dataset limitations, focus on cultural sensitivity, and rigorous model evaluation. Our proposed methodology is a direct response to the major gaps and limitations identified in existing studies by employing two multi-classifier ensemble models trained on a well-curated and diverse dataset, enabling more nuanced and accurate fine-grained classification of abusive content in social media environments. While advanced deep learning approaches have the potential to outperform classical models in many NLP tasks, the practical considerations around resource limitations, dataset characteristics, interpretability, and the focus on ensemble strategies justify the decision to use classical machine learning models in this study. This approach aligns with the research objectives and ensures that the methodology remains accessible and effective for the task at hand.

The Bagging strategy as illustrated by Fig. 1, consists of training multiple models (known as weak learners) implementing the same algorithm (such as the Decision Tree) on different sampled data drawn randomly with replacement from the dataset. This technique is used in EL to create a diverse group of ensemble members by varying the training data. Typically, a single machine learning algorithm, such as an unpruned decision tree, is used to train each model on a different sample from the same training dataset. The predictions made by the ensemble members are then combined using simple statistics like voting or averaging. The key aspect of this method is how each sample is prepared for training the ensemble members. Each model receives its own unique sample from the dataset. The samples are drawn randomly from the dataset, with replacement. Replacement refers to the fact that if a sample is selected, it is returned to the training dataset and may be selected again. This allows for a sample to be chosen zero, one, or multiple times for a particular training dataset. This process is known as a bootstrap sample and is commonly used in statistics to estimate the statistical value of a data sample. By creating various bootstrap samples and evaluating a statistical measure, along with computing the average of these evaluations, it becomes possible to obtain a more accurate estimation of the desired measure compared to directly estimating it from the original dataset. Fig. 1 shows the architecture of a Bagging Ensemble with sampled data labeled Sample1, Sample2, Sample3, etc.

Similarly, multiple training datasets can be generated and employed to develop a predictive model and generate predictions. Taking the average of these predictions across the models generally leads to improved accuracy compared to using a single model based solely on the original training dataset. The key benefits of bagging can be summarized as: Bootstrap samples of the training dataset; Unpruned decision trees fit on each sample; and Simple voting or averaging of predictions.

The Stacking Strategy consists of training multiple weak learners implementing different algorithms on the same dataset, in this case a balanced subset sampled from the original dataset as shown on Fig. 2. This technique aims to create a diverse group of members by employing various types of models on the training data and utilizing a model to merge their predictions.

The members of the ensemble are generally referred to as level-0 models, while the model responsible for combining the predictions is called a level-1 model. Typically, a two-level hierarchy of models is used, although additional layers can be incorporated. For instance, instead of a single level-1 model, we might have 3 or 5 level-1 models, along with a solitary level-2 model that integrates the predictions from the level-1 models to make a final prediction. In Fig. 2, Model1, Model2 and Model3 are level-0 models while Model is level-1. Any machine learning model can be employed to aggregate the predictions, although it is common to use a linear model such as linear regression for regression tasks or logistic regression for binary classification. This approach ensures that the complexity of the model resides in the lower-level ensemble

member models, while simpler models learn how to harness the diversity of predictions. The diversity arises from utilizing different machine learning models as ensemble members [17]. The key benefits of stacking can be summarized as: Unchanged training dataset; Different machine learning algorithms for each ensemble member; and Machine learning model to learn how to best combine predictions.

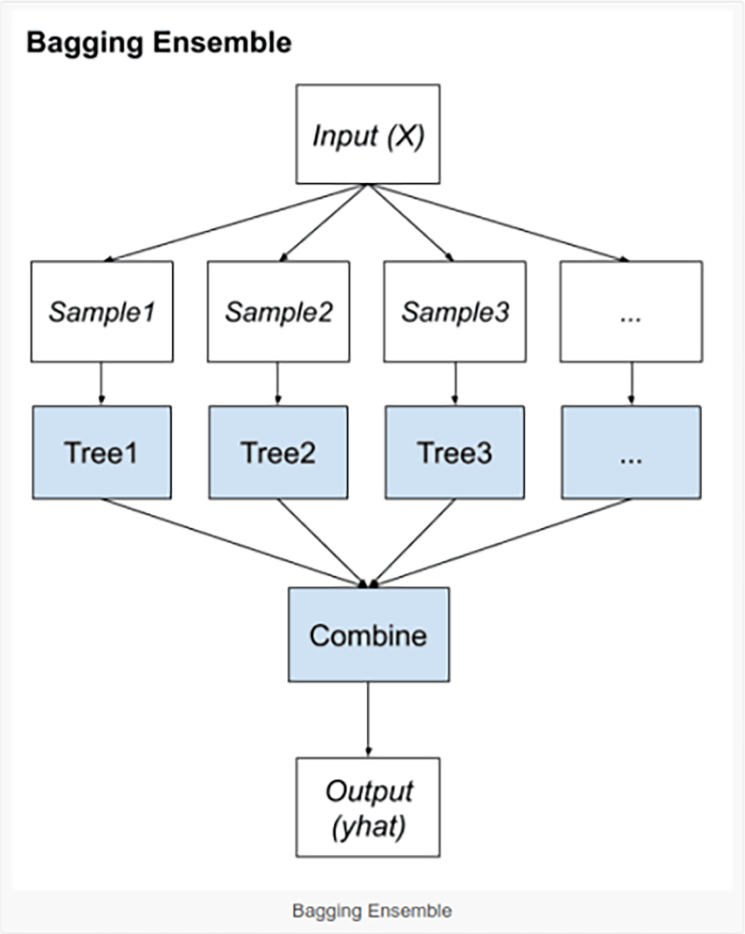


Figure 1: Bagging ensemble [24,25].

There exist different types of abusive contents as they refer to language, images, or actions that are intended to insult, demean, harass, threaten, or emotionally harm an individual or a group. On online platforms, abusive content degrades the quality of interaction, can cause psychological harm to victims, and contributes to hostile digital environments, making its detection and moderation an important challenge in cybersecurity and content moderation systems. This work considers five classes of online content namely Hate Speech, Racism, Sexism, Abusive, and Others. Hate Speech, Racism, and Sexism represent specific forms of abusive content targeting protected characteristics such as ethnicity, nationality, religion, race, skin color, gender or sex. Hate Speech denotes content containing hateful words or expressions, while Racism and Sexism refer to content exhibiting discriminatory or hostile attitudes based on race and gender, respectively. Content that is offensive but does not fall into these specific categories is labeled as Abusive. The Others class encompasses content that cannot be identified as hate speech, racism, or sexism and may include both abusive and non-abusive content. This taxonomy enables the distinction between targeted forms of abuse and more general offensive language, thereby supporting fine-grained abusive content classification.

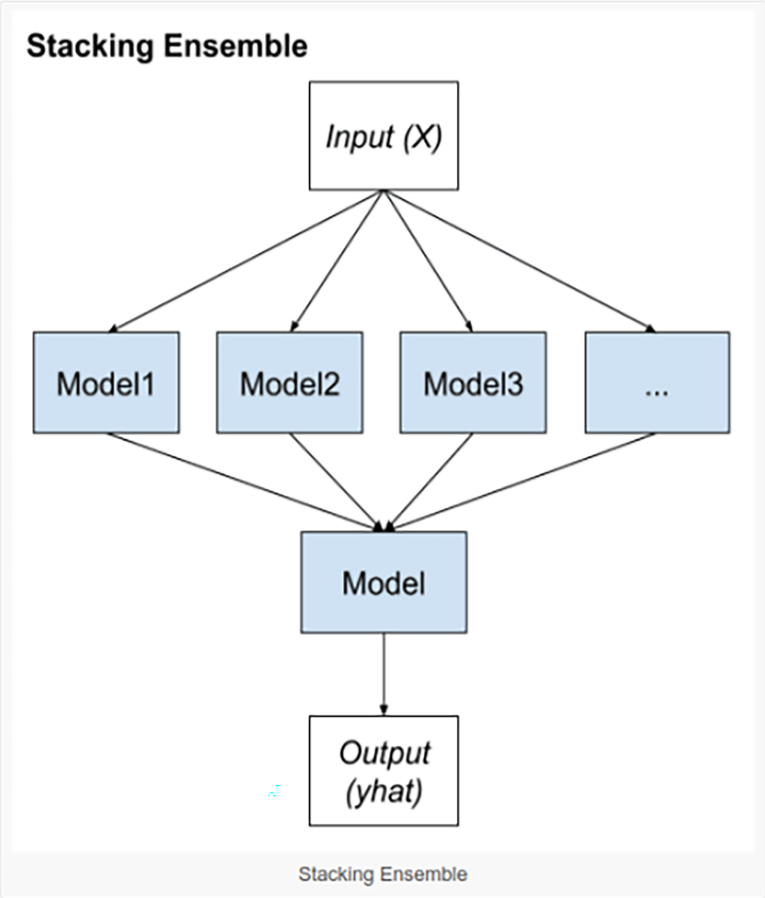


Figure 2: Stacking ensemble [24,25].

3.2 Modelling

This paper proposes two models based on EL techniques to classify abusive contents into Racism, Sexism, Hate speech, Abusive and Others. The first model applies the ensemble stacking strategy while the second model applies a combination of ensemble bagging and stacking strategies. The dataset used for training these models was constructed as elaborated in Section 3.3. Table 1 presents a description of each of the five different classes of abusive contents considered in this paper.

Three traditional machine learning classifiers namely MNB, LSVM, and LR were employed as our base models. Each model performs the multi-classification by implementing the One-vs-Rest (OvR) multi-classification strategy as illustrated in Fig. 3. The OvR approach decomposes the multi-class problem into a set of binary classification tasks, allowing each classifier to focus on distinguishing one target class from all remaining classes. This strategy offers flexibility, improved handling of class imbalance, and enhanced interpretability, making it well suited to the objectives of abusive content detection.

Specifically, each model performs five binary classification tasks corresponding to the following class combinations: (i) Sexism vs. {Racism, Hate Speech, Abusive, Others}, (ii) Racism vs. {Sexism, Hate Speech, Abusive, Others}, (iii) Hate Speech vs. {Sexism, Racism, Abusive, Others}, (iv) Abusive vs. {Sexism, Racism, Hate Speech, Others}, and (v) Others vs. {Sexism, Racism, Hate Speech, Abusive}. The final class prediction for both ensemble models is obtained through a Voting Classifier (VC), which aggregates the outputs of the base learners to produce the final decision.

Table 1: Classes of abusive content considered.

Class	Description
Hate Speech	Content containing hateful words or phrases directed toward individuals or groups based on protected characteristics.
Sexism	Content expressing prejudice, hostility, or discrimination based on gender or sex.
Racism	Content expressing prejudice, hostility, or discrimination based on race, ethnicity, nationality, or related attributes.
Abusive	Content that is offensive or abusive but does not qualify as hate speech, sexism, or racism.
Others	Content that cannot be identified as hate speech, racism, or sexism and may be abusive or non-abusive.

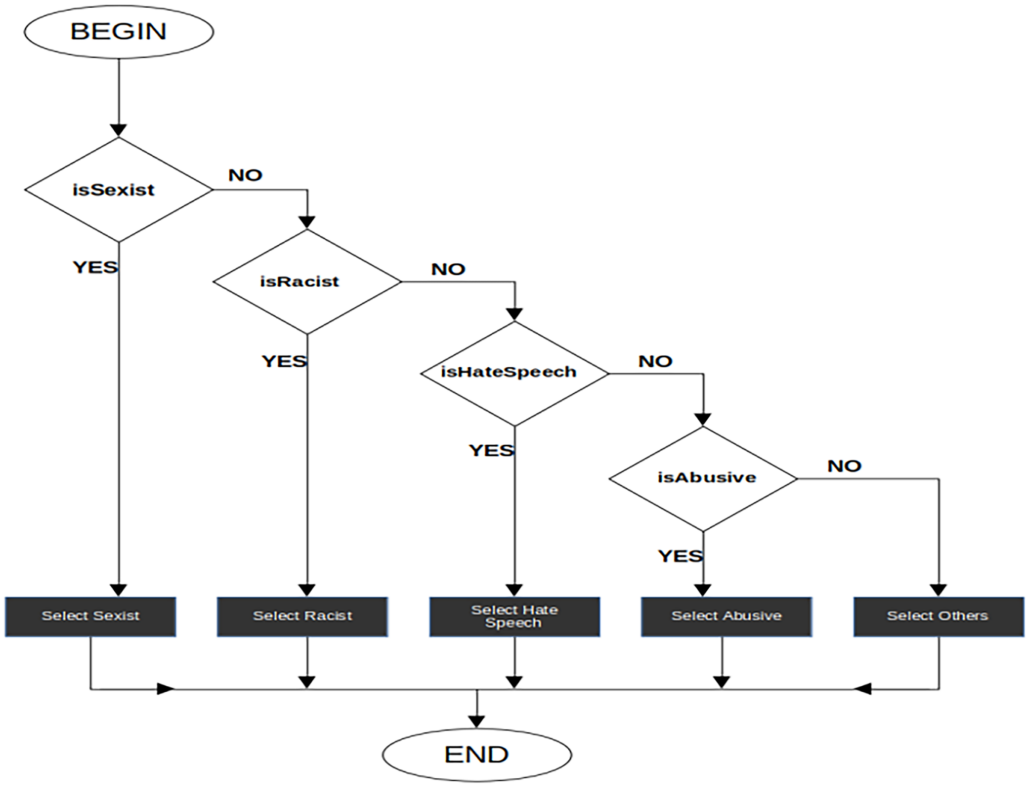


Figure 3: Base model classification algorithm.

3.3 Dataset Construction

In this section, we explain how datasets from different sources were combined, present the original label definitions of each source dataset and explain the mapping strategy used to create the unified taxonomy. For the development of a multi-class classification model, we required a dataset annotated with the five classes defined earlier in Table 1. However, most publicly available abusive language datasets, primarily derived from X, are limited to two or three classes. To address this limitation, we constructed a unified dataset by integrating four existing tweet-based datasets, enabling consistent labeling across the five target classes. Tweets were selected given that X is a widely used social media platform whose linguistic characteristics and

abusive expressions are comparable to those observed on other platforms such as Facebook. To construct a unified dataset for abusive content classification, we integrated four publicly available social media datasets, as presented in Table 2, namely: HASOC-19, Davidson et al. dataset, CMSB, and RacismDetectionDataSet. Since these datasets were originally developed for different abusive language detection tasks and employed distinct annotation schemes, a label harmonization process was performed to ensure consistency across the target classes. The final dataset was organized into five mutually exclusive categories: Hate Speech, Abusive, Racism, Sexism, and Others. Furthermore, since the resulting dataset exhibited class imbalance, we applied random sampling techniques to obtain a balanced dataset suitable for multi-class classification. Fig. 4 shows the individual hate speech labels and their distribution.

Table 2: Datasets used and their sources.

Dataset	Source
HASOC-19	https://hasocfire.github.io/hasoc/2019/dataset.html
Davidson et al.	https://github.com/t-davidson/hate-speech-and-offensive-language
CMSB	https://www.kaggle.com/datasets/ccymforhpl/cmsb-sexist?resource= download&select= CMSB_all.csv
RacismDetectionDataSet	https://www.kaggle.com/datasets/rishabhjohri/racismdetectiondataset

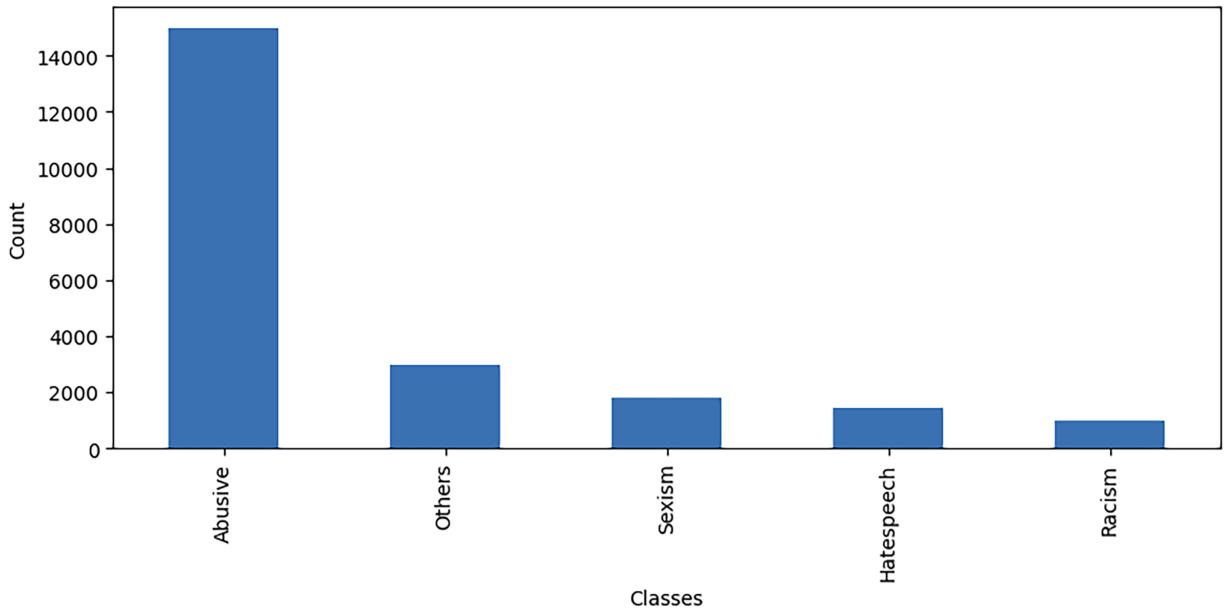


Figure 4: Constructed dataset label distribution.

The 2019 Hate Speech and Offensive Content Identification (HASOC-19) [26] dataset consists primarily of tweets collected from Twitter (now X) and is annotated to capture different forms of abusive content. The dataset labels texts as Hate and Offensive (HOF) or Non-Hate and Offensive (NOT), further categorizing Hateful and Offensive instances into *Hate Speech (HATE)*, *Offensive (OFFN)*, and *Profanity (PRFN)*. For the purpose of this study, all instances labeled as HATE (1143 samples) were extracted and mapped to the Hate Speech class. In the unified dataset, label 0 (Hate) was mapped to the Hate Speech class.

The Automated Hate Speech Detection and the Problem of Offensive Language by Davidson et al. (2017) [27,28] dataset provides annotations for both binary and fine-grained classification tasks, and which has been widely adopted as a benchmark for evaluating ML and deep learning models in hate speech and offensive language detection. The dataset is annotated using three labels: 0 (Hate), 1 (Offensive/Abusive), and 2 (Neither). To improve annotation reliability and reduce label noise, only instances with a minimum of three annotator votes were retained. Following this filtering process, the dataset comprised 308 Hate instances mapped to the Hate Speech class, 14,988 Offensive/Abusive instances mapped to the Abusive class, and 2994 Neither instances mapped to the Others class. In the unified dataset, label 1 (Offensive/Abusive) was mapped to the Abusive class, while label 2 (Neither) was mapped to the Others class.

The Call Me Sexist But (CMSB) dataset is a curated resource for sexism detection in social media text, designed to support nuanced classification of sexist language. The dataset comprised 13,337 comments annotated as either Sexist or Not-Sexist. All instances labeled as Sexist (1809 samples) were selected and mapped to the Sexism class. In the unified dataset, label 4 (Sexist) was mapped to the Sexism class.

The RacismDetectionDataSet dataset has been widely used in machine learning and deep learning research to train and assess models that identify racist language in short-form social media content. It comprises social media comments labeled for the presence of racist content, facilitating classification of racist vs. non-racist text. The dataset comprises 1987 English-language comments annotated as either Racist or Non-Racist. All instances labeled as Racist (1000 samples) were retained and mapped to the Racism class. In the unified dataset, label 3 (Racist) was mapped to the Racism class.

To ensure label consistency, a unified mapping strategy was applied across all datasets. The resulting class definitions were: Hate Speech (content expressing hatred toward individuals or groups), Abusive (offensive, insulting, or profane language), Racism (content targeting individuals or groups based on race or ethnicity), Sexism (content targeting individuals based on gender), and Others (non-abusive content). This harmonization process reduced inconsistencies arising from heterogeneous annotation schemes and enabled the development of a coherent multi-class classification framework. Table 3 summarizes the final class distribution and the corresponding label mappings used in the constructed dataset.

Table 3: Constructed dataset with labels.

Label	Class	Total Count	Source Dataset	Source Count
0	Hatespeech	1451	HASOC-19 Davidson et al.	1143 308
1	Abusive	14,988	Davidson et al.	14,988
2	Others	2994	Davidson et al.	2994
3	Racism	1000	RacismDetectionDataSet	1000
4	Sexism	1809	CMSB Dataset	1809

The constructed dataset exhibited a significant class imbalance problem. As illustrated in Fig. 4, the Abusive class was highly dominant, containing more than 14,000 instances. To address this imbalance, we generated four sampled data (balanced subsets) by randomly sampling 1000 instances per class with replacement, thereby ensuring an even class distribution, as shown in Fig. 5. For model training, a sufficient number of samples per class was required. Consequently, the Racism class (class 3), which contained the fewest instances (1000), was used as the reference class. Each balanced subset therefore comprised

1000 instances per class, resulting in a total of 5000 instances per sample/subset. With this configuration, each subset needed to have 1000 tweets of each class, totaling 5000 tweets per subset, which would not be possible when sampling without replacement. Sampling without replacement would have required at least 4000 instances per class to construct four balanced subsets. Accordingly, with the exception of the Racism class, which remained unchanged, all other classes, including the Abusive class, were sampled with replacement to ensure consistency and fairness across the subsets.

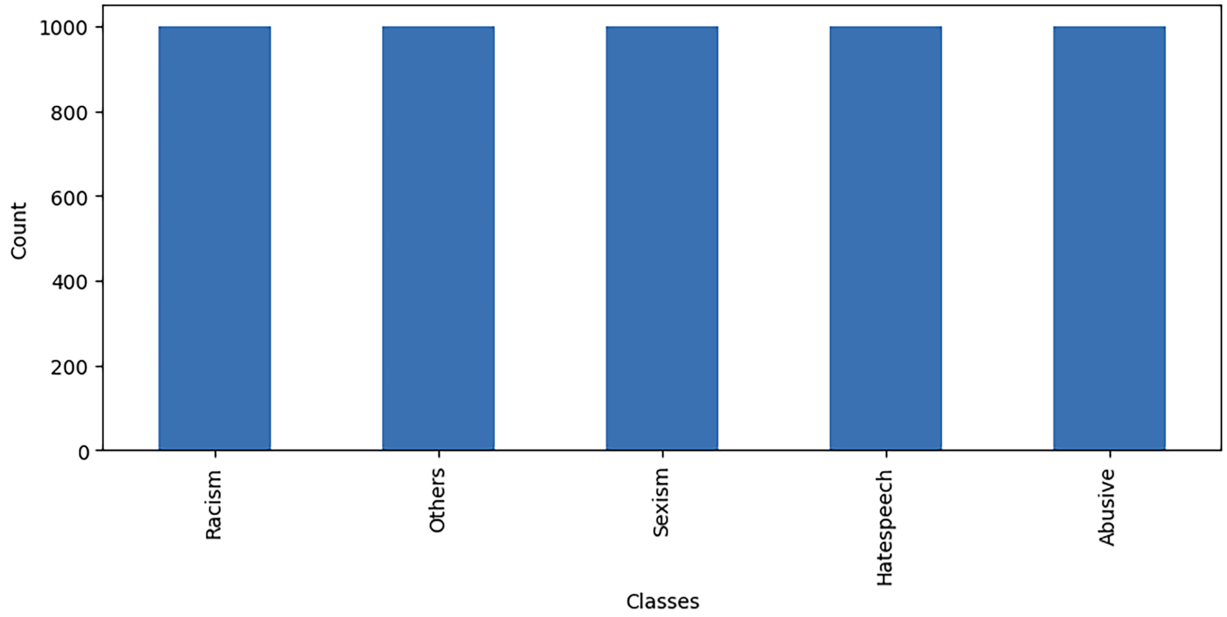


Figure 5: A sampled dataset label distribution.

To prevent potential data leakage and ensure the validity of the experimental results, all datasets were merged prior to partitioning. A deduplication process was then performed to identify and remove duplicate or repeated text instances originating from multiple datasets. Following deduplication, the unified dataset was randomly partitioned into training, validation, and testing subsets. Performing dataset splitting only after the deduplication stage ensured that identical or near-identical samples could not appear simultaneously in different partitions, thereby eliminating a major source of data leakage and preventing artificially inflated performance estimates.

The constructed dataset is intended as a task-specific benchmark for conversational content classification across five classes in low-resource settings. It supports baseline model development and comparative evaluation by capturing key characteristics of conversational data, such as informal language and class overlap. Although its size limits direct generalization to large-scale deployments, the dataset is suitable for controlled experimentation and preliminary validation, and can be extended or combined with larger corpora in future work.

3.4 Models' Construction

This paper proposes two ensemble-based models that employ stacking and bagging strategies for the classification of abusive content into five categories: Racism, Sexism, Hate Speech, Abusive, and Others. The first model, referred to as **Ensemble Stacking-Voting (ESV)**, utilizes a stacking approach, while the second model, termed **Ensemble Bagging-Stacking (EBS)**, integrates both bagging and stacking techniques.

Both ESV and EBS models are built using three different base models namely MNB, LSVM and LR. During the training of the ESV model, the base models are trained and tested on the same dataset and then their predictions are sent into a VC for final prediction. Fig. 6 depicts the architecture of the ESV model.

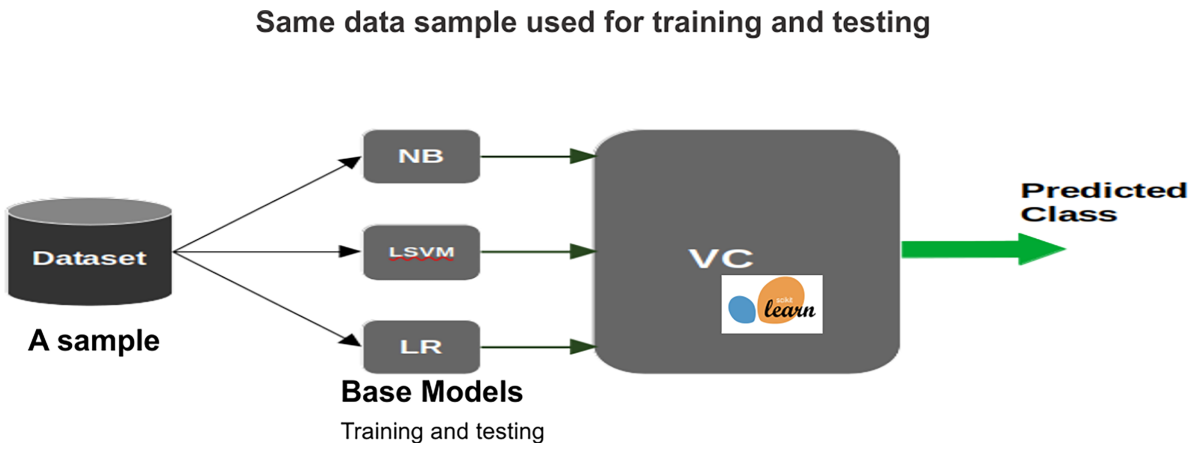


Figure 6: Architecture of ESV-based model.

The architecture of the EBS model is shown on Fig. 7. Each base model is trained using a different dataset sample. The pre-trained models are then tested using another dataset sample and a voting classifier is used to determine the final prediction. Like Bagging, different samples are used and like Stacking, different algorithms or models are used as base models.

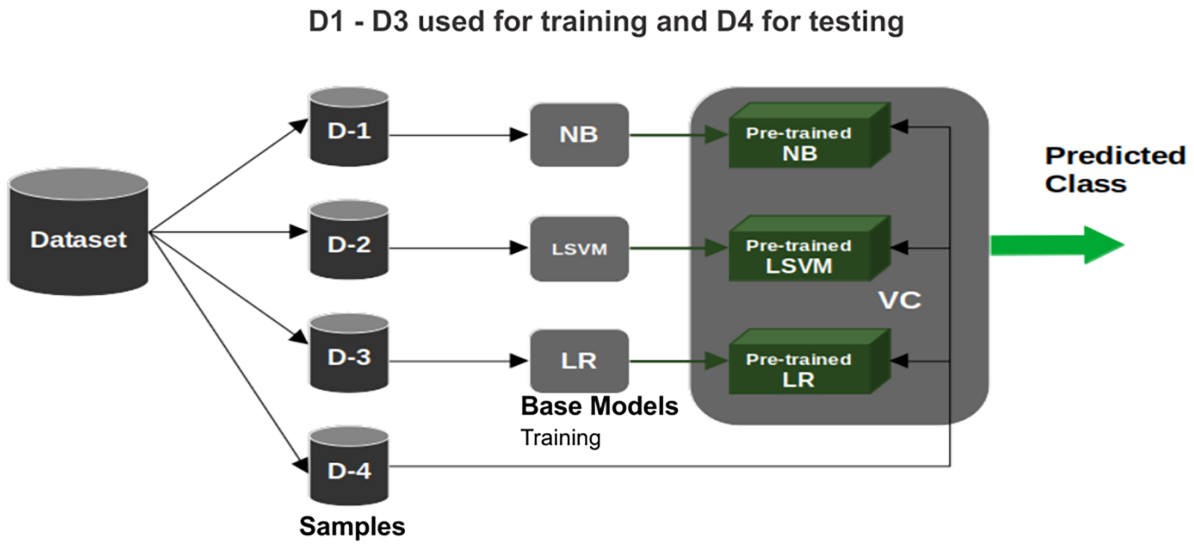


Figure 7: Architecture of EBS-based model.

3.5 Implementation Process

Following the construction of the dataset described in Section 3.3, the implementation process commenced with dataset balancing to address class distribution disparities. Four balanced subsets, denoted as D-1, D-2, D-3, and D-4, were generated through sampling, each containing an equal number of instances per class. These balanced datasets were subsequently used for training and evaluating the proposed models.

Five classification models were implemented and evaluated, namely MNB, LSVM, LR, the ESV model based on Stacking Strategy and the EBS model based on Stacking and Bagging Strategies. The ensemble models employ MNB, LSVM, LR as base learners. All models were trained following the standardized workflow illustrated on Fig. 8.

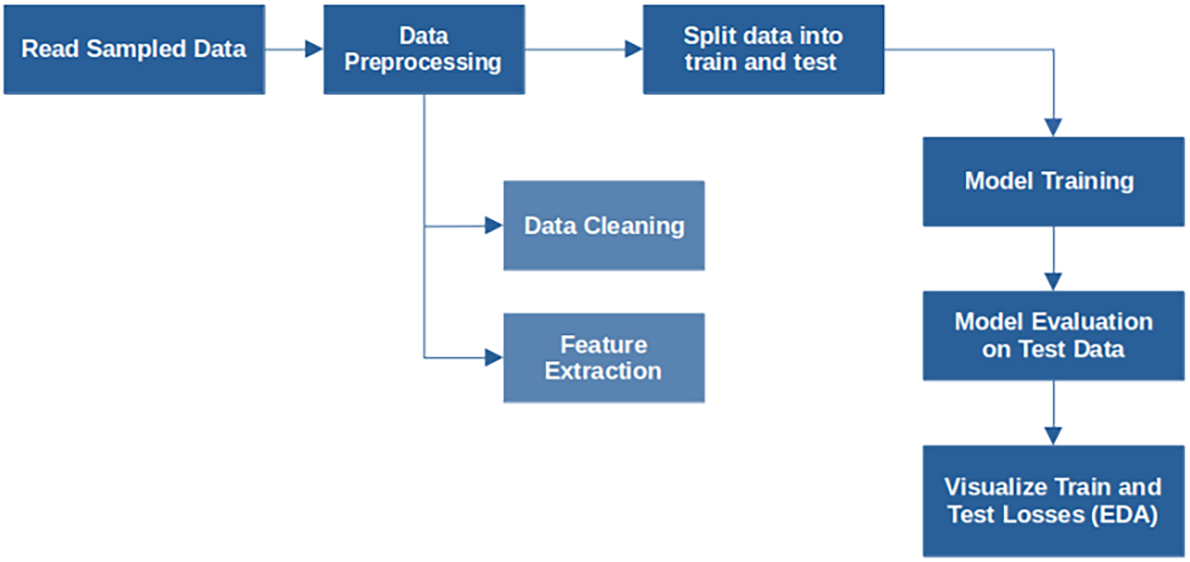


Figure 8: Training process of each base model.

The implementation pipeline begins with data loading and preprocessing. Preprocessing operations included decoding the text, converting all characters to lowercase, replacing unwanted characters with whitespace, and removing stop words. This is a critical step in preparing the dataset for effective training and validation of the multi-classifier model. This stage ensures that the input data is free from noise and irrelevant information, thereby enhancing the quality and reliability of the training process. Specifically, the following tasks were also carried out during this cleaning stage:

- Data collection review: The initial datasets, obtained from various sources (e.g., Kaggle, and existing annotated datasets), were examined to identify common patterns, structures, and any existing issues.
- Removing duplicate entries: Duplicate records can skew the training results. The dataset was thoroughly checked for identical entries, and any duplicates were removed to maintain uniqueness and prevent bias.
- Tokenization: The cleaned text was then split into individual words or tokens. This process facilitates easier analysis and manipulation of the text data for subsequent steps like feature extraction.
- Stemming and Lemmatization: Various forms of words were reduced to their base (stem) or root (lemma) form. For example, “running”, “runner”, and “ran” may all be reduced to “run”. This process aids in consolidating similar words, enhancing the model’s understanding of context.
- Filtering: An initial lexical filtering stage was applied during dataset construction to identify candidate abusive content and eliminate clearly irrelevant or corrupted records (e.g., spam, duplicate entries, and non-informative posts). The lexical resources were used solely to facilitate dataset curation and were not employed to remove abusive words from the textual content used for model training. Consequently, the original text was preserved to ensure that all linguistic cues relevant to abusive language detection and fine-grained classification remained available to the classifiers.

By executing these cleaning procedures, the dataset was significantly refined, resulting in a high-quality corpus ready for the next phases of preprocessing, feature extraction, and model training.

Feature extraction was then performed using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. TF-IDF was selected due to its simplicity, interpretability, and proven effectiveness in text classification tasks, particularly for abusive content detection in resource-constrained or context-specific settings. TF-IDF assigns a numerical weight to each term based on its frequency within a document and its inverse frequency across the entire corpus [23], thereby emphasizing terms that are discriminative while down-weighting commonly occurring words. The calculation of TF-IDF for a given term or word is obtained by multiplying the word's Term Frequency (TF) with the Inverse Document Frequency (IDF). The TF for a term t within document d (denoted as TFt,d) is determined by the raw count of the term t in the document d, i.e., the number of times that t occurs in document d, by the total number of terms in the document. Formally, the TF component represents the normalized frequency of a term t in a document d, as defined in Eq. (1). This process is applied to all terms, resulting in a vectorized numerical representation of the text data. The IDF component measures the informativeness of a term across the corpus and is computed as the logarithm of the inverse proportion of documents containing the term, as shown in Eq. (2). The final TF-IDF weight for each term is obtained by multiplying its TF and IDF values, as expressed in Eq. (3).

$$TF_{t,d} = \frac{n_{t,d}}{\sum_k n_{t,d}} \tag{1}$$

$$IDF_{t,D} = \log\left(\frac{N}{|\{d \in D: t \in d\}|}\right) \tag{2}$$

where

- N is the total number of documents in the corpus; $N = |D|$
- $|\{d \in D: t \in d\}|$ is the number of documents where the term t appears $(i.e., \mathrm{TF}_{t,d} \neq 0)$. If the term is not in the corpus this will lead to a division-by-zero. It is therefore common to adjust the numerator $1 + N$ and denominator to $1 + |\{d \in D: t \in d\}|$

$$TF - IDF = TF * IDF \tag{3}$$

Following feature extraction, each dataset was partitioned into training and testing sets using a 70:30 split ratio. The resulting subsets were then used for model training, evaluation, and result visualization. For the ESV model, the balanced dataset D-4 was used for both training and testing. In contrast, the EBS model leveraged all four balanced datasets: MNB was trained on D-1, LSVM on D-2, and LR on D-3, after which the pre-trained base models were evaluated using D-4. This strategy enabled diversity among base learners while improving generalization performance through ensemble learning.

The proposed models were implemented in Python 3.11 using the Scikit-learn machine learning library (version 1.4.1). Supporting libraries included NumPy, Pandas, SciPy, and Matplotlib for data manipulation, numerical computation, and visualization. Text preprocessing and feature extraction were performed using the TF-IDF vectorizer available in Scikit-learn. Table 4 summarizes the main hyperparameter settings adopted during model training. Unless otherwise specified, default Scikit-learn parameters were retained.

Table 4: Hyperparameter configurations used for training.

Component	Library/Model	Hyperparameter
Programming Language	Python	Version 3.11
Machine Learning	Scikit-learn	Version 1.4.1

(Continued)

Table 4 (continued)

Component	Library/Model	Hyperparameter
Feature Extraction	TF-IDF Vectorizer	CountVectorizer() (default), TfidfTransformer()(default)
Multinomial Naïve Bayes	MultinomialNB	Alpha = 1.0
Linear SVM	LSVM (SGD-based)	Loss = 'hinge', penalty = 'l2', alpha = 0.001, max_iter = 5000, random_state = 42
Logistic Regression	LogisticRegression	C = 1e5, solver = 'lbfgs', max_iter = 100, n_jobs = 1
Ensemble Stacking (ESV)	VotingClassifier	Voting = 'hard' Base learners = MNB, LSVM, LR
Ensemble Bagging-Stacking (EBS)	Bagging + Stacking	Voting = 'hard' Base learners = MNB, LSVM, LR

4 Results and Discussion

This section reports the experimental results obtained using the proposed models. We first analyze the performance of the individual base classifiers, followed by the ensemble approaches. Model evaluation was conducted using overall accuracy, per-class F1-scores, training time, and confusion matrices. Accuracy provides a global measure of classification performance, whereas class-wise F1-scores offer a balanced assessment by combining precision and recall into a single metric, making them particularly suitable for multi-class classification problems with potentially imbalanced classes. Confusion matrices complement these metrics by providing detailed insights into class-wise prediction behavior and revealing misclassification patterns among semantically related categories. Finally, training time was considered to assess the computational efficiency of the proposed models, which is an important consideration for lightweight and interpretable ensemble approaches intended for resource-constrained deployment scenarios.

The MNB classifier achieved an overall accuracy of 86.4%, with a training time of 0.085 s, making it the most computationally efficient model evaluated. The confusion matrix, as presented in Table 5, indicates strong performance on the Abusive class, while confusions are observed between semantically related classes such as Hate Speech, Racism, and Sexism. These results suggest that MNB is effective for high-frequency and lexically explicit abusive content but less robust for nuanced distinctions. The LSVM model achieved an accuracy of 89.6% with a training time of 0.312 s, outperforming MNB in predictive performance. As shown in Table 6, LSVM demonstrates improved discrimination across all five classes, with fewer misclassifications in the Hate Speech and Racism categories. This indicates stronger generalization capabilities while maintaining reasonable computational cost.

Table 5: Performance of the MNB model (%).

True\Predicted Labels	Hatespeech	Racism	Sexism	Abusive	Others
Hatespeech	88.00	2.20	0.63	4.10	4.70
Racism	7.20	80.00	1.60	2.00	8.90
Sexism	3.80	1.00	78.00	7.60	10.00
Abusive	1.70	1.70	0.34	93.00	2.70
Others	3.30	0.67	1.70	2.00	92.00

Table 6: Performance of the LSVM model (%).

True\Predicted Labels	Hatespeech	Racism	Sexism	Abusive	Others
Hatespeech	90.00	1.30	2.30	3.30	2.70
Racism	7.30	81.00	3.70	7.00	1.00
Sexism	4.10	1.00	93.00	1.00	1.00
Abusive	1.70	2.10	2.10	91.00	3.10
Others	2.80	1.50	1.50	1.20	93.00

As for the LR classifier, it achieved an accuracy of 89.2% with a training time of 0.734 s. Its results, as illustrated in Table 7, show balanced classification behavior across most classes, with particularly strong performance on the Abusive and Others categories. Although LR performs comparably to LSVM in terms of accuracy, it requires higher computational effort during training. Among the base classifiers, LSVM achieved the highest accuracy, while MNB offered the fastest training time. All base models showed consistent strength in identifying the Abusive class, whereas confusion among Hate Speech, Racism, and Sexism remained a challenge. These observations motivate the exploration of ensemble-based solutions.

Table 7: Performance of the LR model (%).

True\Predicted Labels	Hatespeech	Racism	Sexism	Abusive	Others
Hatespeech	93.00	3.90	1.30	3.30	0.97
Racism	2.50	92.00	1.10	2.50	2.50
Sexism	2.40	4.10	90.00	1.00	2.70
Abusive	3.90	3.50	6.80	81.00	4.50
Others	2.00	2.60	2.90	1.60	91.00

For the ensemble methods, ESV model achieved an accuracy of 90.0% with a training time of 1.097 s, achieving the highest performance as compared to all individual base classifiers. As illustrated in Table 8, ESV reduces misclassification across minority classes, particularly Sexism and Hate Speech, confirming the benefit of aggregating heterogeneous classifiers through stacking. EBS model on its own achieved the best performance, with an accuracy of 92.6% and a training time of 1.217 s. The confusion matrix in Table 9 shows consistently high true positive rates across all classes, with notable improvements in the Racism and Sexism categories. These results demonstrate that combining bagging with stacking enhances robustness and reduces variance.

Table 8: Performance of the ESV model (%).

True\Predicted Labels	Hatespeech	Racism	Sexism	Abusive	Others
Hatespeech	94.00	1.60	0.97	1.30	1.60
Racism	6.20	82.00	8.90	1.40	1.00
Sexism	2.00	2.40	90.00	3.00	2.70
Abusive	1.90	1.30	0.97	96.00	0.00
Others	5.10	3.70	1.70	2.40	87.00

Table 9: Performance of the EBS model (%).

True\Predicted Labels	Hatespeech	Racism	Sexism	Abusive	Others
Hatespeech	93.30	1.11	1.11	2.23	2.23
Racism	5.55	92.48	0.65	0.32	0.98
Sexism	0.92	1.54	92.28	2.77	2.46
Abusive	5.29	1.24	1.24	89.40	2.80
Others	1.07	0.35	0.35	2.14	96.07

The ensemble models achieved the highest performance, attaining an accuracy of 92.6% and consistently superior class-wise F1-scores compared with the individual classifiers. The ESV model achieves 90.0% accuracy, while EBS attains the highest accuracy of 92.6% and consistently strong F1-scores across all five classes (≥90%). Notably, EBS shows substantial improvements for the *Others* and *Sexism* classes, with F1-scores of 94%, suggesting enhanced robustness in handling ambiguous or overlapping categories. Although ensemble models incur higher training times, the increase remains modest and acceptable for offline training scenarios. Table 10 summarizes the overall performance of all five evaluated classifiers, including F1-scores. Based on these summary results, the EBS model consistently outperforms both base classifiers and the ESV model, achieving relative improvements of 1.07×, 1.03×, and 1.02× over MNB, LSVM, and ESV, respectively. Although ensemble methods incur higher training costs, all models remain computationally efficient, with training times below 2 s, which to us is tolerable for the improved accuracy obtained in returned. Overall, the results confirm that ensemble learning is particularly effective for fine-grained abusive content classification, where class overlap and linguistic ambiguity are prevalent.

Table 10: Summary of comparative results of all models.

Model	Accuracy (%)	Training Time (s)	F1-Score (%)				
			Racism	Hate Speech	Others	Sexism	Abusive
MNB	86.4	0.0855	86	84	85	87	89
LSVM	89.6	0.3126	87	91	88	89	93
LR	89.2	0.7342	87	90	89	89	91
ESV	90.0	1.0973	86	90	90	89	94
EBS	92.6	1.2174	90	91	94	94	94

A closer examination of class-wise F1-scores reveals that *Abusive* and *Hate Speech* are generally easier to classify across all models, likely due to the presence of more explicit lexical cues. In contrast, the *Racism* and *Others* classes exhibit lower and more variable F1-scores in baseline models, reflecting semantic overlap and contextual ambiguity. MNB, in particular, struggles with these classes, which is consistent with its strong independence assumptions and limited contextual modeling capability.

The superior performance of EBS suggests that aggregating multiple learners reduces individual model bias and variance, leading to more stable predictions across classes. Error analysis indicates that misclassifications frequently occur between *Hate Speech* and *Abusive*, as well as between *Sexism* and *Others*, where contextual intent rather than explicit keywords determines the correct label. Ensemble learning mitigates these errors by combining complementary decision boundaries learned by different models. Overall, the results demonstrate that while lightweight baseline models offer favorable computational efficiency,

ensemble-based approaches provide significantly improved generalization and class balance, particularly in multi-class conversational settings with overlapping semantics.

This study explores two ensemble-based multi-classifier models namely ESV and EBS, for the task of detecting and classifying abusive content into five distinct categories: Racism, Sexism, Hate Speech, Abusive, and Others. Given the scarcity of well-annotated datasets for fine-grained abuse classification, we addressed this challenge by constructing a comprehensive dataset through the aggregation and relabeling of samples from four different existing datasets. This effort ensures a more balanced and diverse representation of the targeted abuse classes. The key idea behind employing ensemble methods lies in leveraging the strengths of different learning algorithms to reduce generalization error and improve classification performance on complex tasks like multi-label abuse detection. The ESV model combines the outputs of three base classifiers using a meta-classifier to enhance performance. ESV aims to capture diverse decision boundaries from different models. EBS on its own, introduces diversity through bootstrap aggregating (bagging) before stacking. Bagging increases robustness and reduces variance, while stacking further refines predictions by learning from the base models' outputs.

The results obtained in this paper point toward the enhanced performance of ensemble methods over standalone classifiers. As presented in Table 10, EBS achieved the highest accuracy of 92.6%, significantly achieving the highest performance compared with the individual base models. This yields a relative improvement that EBS was 1.07 times more accurate than the weakest base model. It also outperformed the ESV model by a margin of 1.02 times, highlighting the benefit of integrating bagging with stacking. This improvement suggests that the additional variance reduction from bagging, combined with stacking's ability to learn complex model combinations, enhances classification robustness, particularly useful in a task where context, semantics, and nuanced differences matter (e.g., distinguishing between racism and general hate speech). The high performance of the proposed models ESV and EBS confirms that EL is well-suited for nuanced NLP tasks involving multiple, overlapping classes like abuse classification. Furthermore, the constructed labeled dataset, drawn from multiple sources, played a crucial role by enabling training on a rich, diverse, and representative set of abuse examples. This approach demonstrates a scalable methodology that can be adapted to other complex text classification tasks where labeled data is limited or inconsistent; fine-grained classification is essential; and model reliability and generalization are critical for deployment (like in social media moderation systems).

The experimental validation was conducted using four independently balanced datasets (D-1, D-2, D-3, and D-4), generated from the unified corpus to mitigate the effects of class imbalance. Each balanced dataset contained an equal number of instances per class, ensuring a fair evaluation of all classifiers and reducing bias toward majority classes. A standardized implementation pipeline was applied consistently across all experiments. This pipeline included text preprocessing (decoding, lowercasing, removal of unwanted characters, and stop-word removal), followed by TF-IDF feature extraction. The use of TF-IDF provided a discriminative and interpretable representation of textual content while maintaining computational efficiency. Five classification models were evaluated, including three individual classifiers (MNB, LSVM, and LR) and two ensemble approaches (ESV and EBS). For each balanced dataset, the data were partitioned into training and testing subsets using a 70:30 split ratio. The ESV model was trained and evaluated using dataset D-4, whereas the EBS model employed a diversified training strategy in which MNB, LSVM, and LR were trained separately on D-1, D-2, and D-3, respectively, before being evaluated on D-4. This design promotes classifier diversity, which is a key requirement for effective ensemble learning and improved generalization. To comprehensively assess model performance, multiple evaluation measures were employed, including overall accuracy, per-class F1-score, training time, and confusion matrices. The confusion matrices provide detailed insights into class-wise prediction behavior, enabling the identification of strengths and

weaknesses of each model beyond overall accuracy. Furthermore, the performance of the ensemble models was systematically compared against that of the individual base classifiers, allowing the contribution of the ensemble strategies to be clearly quantified.

In contrast to deep learning-based approaches commonly reported in the literature, the proposed models rely on lightweight and interpretable techniques, including TF-IDF feature representation and classical machine learning classifiers. While deep learning models often achieve superior performance on large-scale datasets by capturing contextual and semantic information, they typically require substantial computational resources, large annotated datasets, and complex training pipelines. The proposed ensemble models offer a complementary alternative that balances performance, interpretability, and computational efficiency, making them particularly suitable for resource-constrained environments or early-stage system deployment.

5 Conclusion

This study investigated the application of ensemble learning techniques for the classification of abusive content into multiple, semantically overlapping categories. Detecting abusive content poses a significant challenge due to the uneven distribution of abusive and non-abusive content on social media platforms. Abusive content accounts for less than 3% of social content, and training models with imbalanced datasets can result in the identification of non-abusive rather than abusive content. Two ensemble-based models, namely Ensemble Stacking-Voting (ESV) and Ensemble Bagging-Stacking (EBS), were proposed and evaluated using traditional machine learning classifiers as base learners. Within the defined experimental setting and dataset, the ensemble models demonstrated consistent performance improvements over individual classifiers in terms of accuracy, precision, and F1-score, suggesting that ensemble learning can enhance robustness when dealing with nuanced classification tasks involving overlapping classes. The proposed methodology involved training the models on a consolidated dataset synthesized from four pre-existing sources. This integrated dataset underwent systematic preprocessing to ensure a balanced class distribution. Experimental evaluations revealed that the EBS achieved the highest performance compared with all other classifiers, achieving a peak accuracy of 92.6%. This performance represents a 1.07-fold increase in predictive accuracy compared to the MNB model. However, the dataset used is relatively limited in size and diversity, and the evaluation was conducted in a controlled offline environment. Consequently, the observed performance gains indicate relative improvements under the given conditions rather than definitive evidence of general superiority across all abusive content detection scenarios. The results obtained therefore serve as empirical observations rather than conclusive claims. Future work will focus on extending this study by exploring advanced feature representations such as contextual embeddings, applying data augmentation or oversampling techniques, and validating the proposed models on larger and more diverse datasets to further improve robustness and generalizability.

Acknowledgement: Appreciations to the different experts who have contributed towards enrichment of this work.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Methodology, Writing of Original Draft, Aline Z. Tsague, Elie Tagne Fute and Brice Armel D. Namou; Data Curation, Software, Validation, Brice Armel D. Namou, Aline Z. Tsague and Benoît Martin A. Quimatio; Formal Analysis, Visualization, Writing (Review and Editing), all authors; Supervision, Resources, Project Administration, Elie Tagne Fute. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Aline Z. Tsague, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EBS Ensemble Backing-Stacking
ESV Ensemble Stacking-Voting

References

1. Alrashidi B, Jamal A, Khan I, Alkhathlan A. A review on abusive content automatic detection: approaches, challenges and opportunities. PeerJ Comput Sci. 2022;8(11):e1142. doi:10.7717/peerj-cs.1142.
2. Saseendran S, Sudharshan R, Sreedhar V, Giri S. Classification of hate speech and offensive content using an approach based on DistilBERT. In: Proceedings of the FIRE 2021-Forum for Information Retrieval Evaluation; 2021 Dec 13–17; Virtual.
3. Auxier B, Anderson M. Social media use in 2021. Washington, DC, USA: Pew Research Center; 2021.
4. Christian M, Zsolt D, Jon DE, Don G, Ina K, Hans-Jürgen R, et al. Problematic social media use in childhood and adolescence. Addict Behav. 2024;153(2):107980. doi:10.1016/j.addbeh.2024.107980.
5. Thuy D. Number of X (formerly Twitter) users worldwide from 2019 to 2025 [Internet]. [cited 2026 Apr 11]. Available from: https://www.statista.com/statistics/303681/twitter-users-worldwide/?srsltid=AfmBOor5uxqWRts-U0VTALIP1CCo4S8Uk6vMrBTA1qQ6ucRsxARK6-og.
6. Bourgonje P, Moreno-Schneider J, Srivastava A, Rehm G. Automatic classification of abusive language and personal attacks in various forms of online communication. In: Rehm G, Declerck T, editors. Language technologies for the challenges of the digital age. Berlin/Heidelberg, Germany: Springer; 2018. p. 180–91. doi:10.1007/978-3-319-73706-5_15.
7. Kaur S, Singh S, Kaushal S. Abusive content detection in online user-generated data: a survey. Procedia Comput Sci. 2021;189(7):274–81. doi:10.1016/j.procs.2021.05.098.
8. Batra J, Jain R, Tikkiwal VA, Chakraborty A. A comprehensive study of spam detection in e-mails using bio-inspired optimization techniques. Int J Inf Manag Data Insights. 2021;1(1):100006. doi:10.1016/j.jjimei.2020.100006.
9. Chandrika CP, Kallimani JS. Classification of abusive comments using various machine learning algorithms. In: Cognitive informatics and soft computing. Singapore: Springer; 2020. p. 255–62. doi:10.1007/978-981-15-1451-7_28.
10. Schmidt A, Wiegand M. A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media; 2017 Dec 11; Boston, MA, USA. doi:10.18653/v1/w17-1101.
11. Waseem Z, Davidson T, Warmsley D, Weber I. Understanding abuse: a typology of abusive language detection subtasks. In: Proceedings of the First Workshop on Abusive Language Online; 2017 Aug 4; Vancouver, BC, Canada. doi:10.18653/v1/W17-3012.
12. Fortuna P, Nunes S. A survey on automatic detection of hate speech in text. ACM Comput Surv. 2019;51(4):1–30. doi:10.1145/3232676.
13. Davidson T, Bhattacharya D, Weber I. Racial bias in hate speech and abusive language detection datasets. In: Proceedings of the Third Workshop on Abusive Language Online; 2019 Aug 1–2; Florence, Italy. doi:10.18653/v1/w19-3504.
14. Sap M, Card D, Gabriel S, Choi Y, Smith NA. The risk of racial bias in hate speech detection. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. doi:10.18653/v1/P19-1000.
15. Ahuja R, Banga A, Sharma SC. Detecting abusive comments using ensemble deep learning algorithms. In: Stamp M, Alazab M, Shalaginov A, editors. Malware analysis using artificial intelligence and deep learning. Berlin/Heidelberg, Germany: Springer; 2021. p. 515–34. doi:10.1007/978-3-030-62582-5_20.

16. Hegde A, Anusha MD, Shashirekha HL. Ensemble based machine learning models for hate speech and offensive content identification. In: Proceedings of the FIRE 2021-Forum for Information Retrieval Evaluation; 2021 Dec 13–17; Virtual.
17. Mnassri K, Rajapaksha P, Farahbakhsh R, Crespi N. Hate speech and offensive language detection using an emotion-aware shared encoder. arXiv:2302.08777. 2023.
18. Mednini L, Noubigh Z, Turki MD. Natural language processing for detecting brand hate speech. J Telecommun Digit Econ. 2024;12(1):486–509. doi:10.18080/jtde.v12n1.859.
19. Wich M, Mosca E, Gorniak A, Hingerl J, Groh G. Explainable abusive language classification leveraging user and network data. In: Dong Y, Kourtellis N, Hammer B, Lozano JA, editors. Machine learning and knowledge discovery in databases: applied data science track. Berlin/Heidelberg, Germany: Springer; 2021. p. 481–96. doi:10.1007/978-3-030-86517-7_30.
20. Walsh S, Greaney P. Multiclass hate speech detection with an aggregated dataset. Nat Lang Processing. 2025;31(6):1350–66. doi:10.1017/nlp.2024.62.
21. Hussain K, Saeed Z, Abbasi R, Sindhu M, Khattak A, Arafat S, et al. Towards understanding the role of content-based and contextualized features in detecting abuse on Twitter. Heliyon. 2024;10(8):e29593. doi:10.1016/j.heliyon.2024.e29593.
22. Kapil P, Ekbal A. A unified multi-task learning architecture for hate detection leveraging user-based information. arXiv:2411.06855. 2024.
23. Alqahtani AF, Ilyas M. An ensemble-based multi-classification machine learning classifiers approach to detect multiple classes of cyberbullying. Mach Learn Knowl Extr. 2024;6(1):156–70. doi:10.3390/make6010009.
24. Zhou ZH. Ensemble methods: foundations and algorithms. Boca Raton, FL, USA: CRC Press; 2012.
25. Sagi O, Rokach L. Ensemble learning: a survey. WIREs Data Mining Knowl Discov. 2018;8(4):e1249. doi:10.1002/widm.1249.
26. HASOC. HASOC 2019 dataset: hate speech and offensive content identification in Indo-European languages [Internet]. 2019 [cited 2026 Jul 29]. Available from: https://hasocfire.github.io/hasoc/2019/dataset.html.
27. Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. Proc Elev Int AAAI Conf Web Soc Media. 2017;11(1):512–5. doi:10.1609/icwsm.v11i1.14955.
28. Davidson T, Warmsley D, Macy M, Weber I. GitHub repository for the paper 'Automated Hate Speech Detection and the Problem of Offensive Language' [Internet]. 2017 [cited 2026 Jul 29]. Available from: https://github.com/t-davidson/hate-speech-and-offensive-language.