International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V71I10P218

Volume 71 Issue 10, 191-204, October 2023
© 2023 Seventh Sense Research Group®

Original Article

Towards Stacking Ensemble-Based Fine-Grained Hostile Class Classification (FGHCC) of Hindi Posts

Ankita Sharma[1], Udayan Ghose[2]

[1,2]*University School of Information, Communication and Technology (USICT), Guru Gobind Singh Indraprastha University, New Delhi, India.*

[1]*Corresponding Author: ankitasharma2711@gmail.com*

Received: 19 May 2023 Revised: 29 August 2023 Accepted: 16 September 2023 Published: 03 October 2023

Abstract *- Lately, there has been a phenomenal surge in Hostile Online Content (HOC). The detection and classification of HOC on Online Social Platforms (OSPs) are becoming an important research area in curbing the toxicity of OSPs. Numerous efforts have been made to address this issue in resource-affluent languages. Detecting and classifying hostile content in Hindi is still challenging due to its nature and constrained resources, like adequate multilabel hostile datasets. There has been phenomenal growth in Hindi online content (OC) due to the emergence of the UTF-8 standard. Consequently, malicious Hindi OC has also skyrocketed. There is a dire need to classify and curb Hindi maleficent content on various OSPs. This paper addresses the problem of FGHCC in Hindi (Devanagari Script) as a multilabel problem since significant overlap exists among the hostile classes. The Hindi Hostility Dataset is used in this work. This work exclusively focuses on FGHCC due to its emerging nature and the scarcity of existing research in this domain. In light of this, a two-tiered stacking ensemble of classifiers is introduced, leveraging problem transformation methods (PTMs) and various state-of-the-art Machine Learning Models (MLMs) such as GNB, DT, RF, SVM, LR, SGD with TF-IDF and unigrams as features are applied. The experimental results demonstrate that the proposed two-layered stacking ensemble based on PTMs with unigram and TF-IDF as features achieved the highest weighted F1 score of 0.60, which outperforms the MLMs used alone, based on One Vs. Rest (OVR), Binary Relevance (BR), Classifier Chains (CC), and Label Powerset (LPS) transformation approaches. Also, the proposed model performs competitively with complex models applied in the literature. Therefore, it indicates the efficacy of our proposed model in detecting fine-grained hostile classes in a resource-constraint scenario.*

Keywords *- Hindi, Hostile posts, Machine learning, Multilabel text classification, Stacking ensemble.*

1. Introduction

Internet-based Communication (IBC) is presently the most prevalent type of communication. It is becoming a mighty tool for publishing content, communicating, and expressing opinions. The textual content on various OSPs is increasing at lightning speed as it is conducive to sharing information and expressing sentiments and opinions [1, 2]. The enormous amount of data produced daily can reach millions beyond the physical boundaries [3]. Despite the numerous benefits, OSPs have boosted maleficent or inappropriate content. Despite numerous regulations, it is difficult to restrict some offensive, unpleasant posts carrying inappropriate content on OSPs. Detecting, classifying, and eradicating offensive content like hate speech and hostile posts on OSPs is a big concern. Unfortunately, OSPs serve as hotbeds for malicious content, and lack of control, anonymity, accountability, intangibility, and online disinhibition effect are the considered factors [4]. Hostile posts are multifaceted harmful content targeting people, communities, or groups and are an essential aspect of inappropriate content of OSPs [5, 6]. Hostile content is usually posted on OSPs for personal or political gain and maniac satisfaction. All this can lead to intimidating effects and make the entire OSP experience hostile. Therefore, classifying and removing offensive content like hostile posts are paramount to maintaining the OSP's hygiene.

This issue is more prevalent in resource-deficient languages like Hindi than in resource-affluent ones. With increased technology, the use of Hindi on the internet has increased exponentially [7]. India is the most populated country worldwide, and Hindi is the most spoken dialect in India, with about 366 million Hindi speakers worldwide. Due to its vast popularity and technological advancement, a large IBC on various OSPs happens in Hindi using the Devnagari script; the population feels more connected and heard when using their native language [8, 9]. People from different educational backgrounds and cultures use their native language to voice their opinions over the internet. Currently, communication and discussions are being held online in

Hindi instead of face-to-face. All of this has given people a deceptive sense of anonymity. Additionally, people do not take responsibility for the words they post online. All this results in a noticeable amount of maleficent, offensive Hindi content online, which is the plague of modern times [10].

Detection, classification, and removal of Hindi hostile posts from online platforms is the day's need since it can plant discrimination thoughts, fear, and hate throughout the communities without being noticed. The need for a hostile post-detection and classification system becomes more apparent. The conventional way of dealing with this problem was manual verification, which is infeasible today. An automated system is required to detect and classify inappropriate content in online posts.

Therefore, this work attempts the multilabel classification of hostile posts since hostile posts have overlapping classes like fake, hate, offensive, and defamation. Multilabel classification is a complex task where each instance can be associated with multiple classes. While there has been substantial research in resource-rich languages like English. As per the author’s knowledge, the field of multilabel classification in Hindi is still evolving [11].

This research is particularly concerned with FGHCC in Hindi, and its scope is somewhat constrained due to the limited resources available for the Hindi language. This study conducts a thorough analysis of the current state-of-the-art (SOTA) approaches for multilabel classification and introduces a sophisticated two-layered stacking-based model designed to perform fine-grained classification of hostile posts automatically. Moreover, it addresses the challenges posed by imbalanced label distributions in the dataset and overfitting concerns by means of the proposed model.

To the best of the author's knowledge, this represents a pioneering effort – the first-ever implementation of a two-layered stacking ensemble comprising heterogeneous classifiers for fine-grained hostile post classification based on problem transformation methods (PTMs) in Hindi (Devanagari script). This research also aimed to demonstrate that achieving success in FGHCC tasks does not necessarily require using large, intricate models like Deep Learning Models (DLMs). Instead, it suggests that exploring the potential of an ensemble consisting of MLMs based on PTMs is a valuable approach to consider.

The primary contributions of this study are outlined as follows:

- This study aims to propose a two-layered stacking ensemble of heterogenous classifiers based on PTMs for the FGHCC task.
- This paper introduces an enhanced approach to simultaneously address post-imbalance and overfitting issues. Instead of focusing solely on ensemble techniques within a multilabel learner, our approach combines cutting-edge multilabel classifiers based on Problem Transformation Methods (PTMs) into a stacking ensemble architecture.
- This integration of multilabel classifiers, each employing PTMs, offers a more diverse and independent set of predictions, which helps mitigate the imbalance problem. Furthermore, the stacking ensemble inherently addresses overfitting concerns, ultimately enhancing the overall classification performance.
- Investigate the performance of the proposed stacking ensemble and do a comparative analysis with other MLMs based on PTMs.
- The proposed model outperformed the individual models based on PTMs in terms of weighted average F1 score, a standard evaluation metric for multilabel classification; the proposed model performs comparably to the complex models applied in the literature.

The paper's organization is as follows: After an introduction in Section 1, Section 2 summarizes the related literature in the field of multilabel Hindi text classification in recent years. Section 3 gives the dataset description. Section 4 covers the proposed methodology and applied techniques. Section 5 discusses the experimental results, and section 6 concludes the paper with future directions followed by references.

2. Related Literature

Detecting online offensive content, encompassing hate speech and hostile posts, has garnered extensive attention within research circles, particularly in languages with abundant resources [12, 13]. However, in resource-constrained languages like Hindi, this area remains relatively underexplored.

The proliferation of online content in Hindi, driven by the adoption of UTF-8 standards, has also led to a notable surge in offensive textual content. The imperative to curb the dissemination of online Hindi hostile content cannot be overstated, given its potential to inflict severe harm on individuals' mental well-being, sow discord within society, and propagate fear, speculation, panic, and misinformation [14, 15].

The pursuit of fine-grained classification of Hindi hostile posts based on their specific types is in its infancy, with only a few studies addressing this challenge in the Hindi language to the author's knowledge. It is worth noting that the advent of online social platforms, offering anonymity, ease of access, and opportunities for online communities and discourse, has compounded the issue of hate speech and hostile post detection, posing a mounting challenge to society, individuals, policymakers, and researchers alike. The advancement of NLP technology has fueled substantial research in recent years in automatic hate speech and hostile post-detection. Notably, renowned competitions referenced in [16, 17], and [18] have been organized to discover improved solutions for automatic hate speech detection, primarily in resource-rich languages such as English. As far as the author is aware, only a limited number of studies have undertaken this endeavor in the context of the Hindi language. Here, we present a concise summary of the research endeavors that have leveraged Hindi datasets to delve into the intricate realm of fine-grained multilabel class classification.

The study points in [19] were to prognosticate the online spread of aggression through textual comments or posts. The dataset consists of posts in English and Hindi. The authors used an ensemble of CNN and SVM for aggression identification. Hindi Facebook and social media posts obtained an F1 score of 0.5599 and 0.3790, respectively. In the future, authors will explore hybrid MLMs and relevant linguistic features for further performance improvement.

Velankar et al. [11] have attempted offensive and hate speech detection in Marathi and Hindi texts. The HASOC 2021 dataset was employed for the same. The Hindi dataset contains binary and more fine-grained labels, while the Marathi datasets only contain binary labels. Different DLMs were deployed, and the results showed that transformer-based models performed the best and the basic models applied excelled for the fine-grained task on the Hindi dataset.

The point of work in [20] was offensive content and hate speech detection in English and Hindi using SVM. The character, word n-grams, and their combination were utilized as features. Fine-grained classification tasks in Hindi obtained a Micro$_{avg}$ F1 score of 0.4513. The researchers concluded that the small training sample and uneven corpus are responsible for the lower performance. Bhardwaj et al. [21] made a Hindi hostility detection dataset. They manually collected and annotated the online social media posts. The hostile posts were considered for multilabel tags since there is significant overlap among classes.

In [22], Sharif et al. utilized BiLSTM and SVM with unigram, bigram, and trigram using LPS. TF-IDF and word2vec were used as embedding techniques for FGHCC. The best Wgt$_{avg}$ F1 score of 50.98 is obtained with SVM with n-gram (1,3) for FGHCC. Azhan and Ahmad [23] propose LaDiff ULMFiT for FGHCC in Hindi and fake news detection in English. Along with the proposed model, LR and RF were also applied.

Results indicated that the proposed model achieved the highest F1 score of 0.53, while LR and RF achieved the same score of 42.74. In [24], Shekhar et al. made use of multiple submissions of models using an ensemble consisting of mBERT and MLMs such as XGBoost and ANN, along with this author also employed LR, SVM, RF, and MLP for Hindi Hostility detection. The fourth submission achieved the highest F1 score among all the applied models.

To sum up, there are few works concerning FGHCC in Hindi, but to the best of the author's knowledge, this work is the first work that seeks to address both post-imbalance and overfitting challenges concurrently by employing a stacking ensemble approach incorporating multilabel classifiers based on PTMs. The present work amalgamates cutting-edge multilabel classifiers utilizing the PTMs into a stacking ensemble framework instead of focusing solely on ensemble techniques within a multilabel learner.

This combination of multilabel classifiers based on PTMs effectively tackles post-imbalance and overfitting issues. The imbalance problem is mitigated through a stacking ensemble comprised of multilabel classifiers. Each classifier is rooted in PTMs, contributing to a potentially more diverse and independent set of predictions. Moreover, the inherent properties of stacking ensembles naturally alleviate overfitting problems, ultimately enhancing the overall classification performance. Experimental results indicate the efficacy of our proposed model for the FGHCC task.

3. Dataset Description

Hostile class classification is a daunting task regarding a resource-deficient language - Hindi, owing to limited resources like adequate hostility datasets [25]. This work uses the Hostility Detection Dataset in Hindi, as mentioned in [21]. It can be considered a gold standard dataset since it is used maximally in related literature. We have used the compressed version of the dataset, viewing only hostile posts. The hostile posts with overlapping hostility labels like fake, defamation, offensive, and hate are considered for FGHCC in this work. Since hostile posts have multiple overlapping labels, this problem is formulated as a multilabel text classification problem.

3.1. Dataset Statistics

A brief statistic of the distribution of Hostile posts is presented in Fig. 1. A total of 3834 hostile posts are taken, out of which 1638, 1132, 1071, and 810 are of fake, hate, offensive, and defamation labels, respectively. The dataset is divided into 70:10:20 for training, validation, and testing. The dataset is further analyzed to obtain valuable insights.

It was observed that the hostile posts contained fewer words while containing a higher average number of letters [21]. On average, fake class includes the maximum number of words per post, followed by hate, offensive, defamation, and defame labels [22]. The same pattern is observed for unique words.

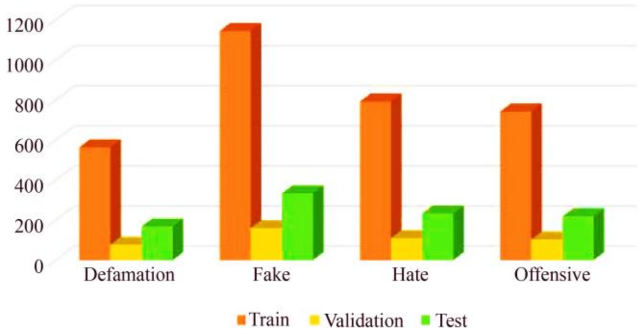

Fig. 1 Shows the statistics of the Hindi Hostility Detection Dataset for FGHCC

3.2. Challenges with the Dataset

Specific challenges with the present dataset have been addressed in this work. Firstly, the dataset is collected from OSPs, implying that the writing style in posts is vastly different from standard Hindi. The posts contain stop words, hashtags, emoticons, URLs, punctuations, misspellings, etc., which are insignificant for the classification process. If left unprocessed, then it might result in low-quality models. This issue is taken care of by the pre-processing step, as discussed in subsection 4.1 of section 4. Another problem is with the dataset size and the imbalanced class distribution, as seen in Fig.1. The number of posts in the fake class is double the number in the defamation class. Also, the dataset lacks discriminative and unique feature collection. The dataset consists of only 3834 hostile posts. Therefore, depending upon the size of the dataset, notability, and features, traditional MLMs can outperform Deep Learning Models (DLMs); thereby, MLMs and their ensemble are applied for classification. Consequently, the feature extraction step facilitates finding the apt set of features sufficient to generalize, as mentioned in a later subsection 4.2.

4. Proposed Methodology

This section explains the methodology employed in this work. The basic framework of the proposed methodology for FGHCC is shown in Fig.2. Classification is a prevalent supervised Machine Learning (ML) task that categorizes the data instances based on similarities into classes or labels defined a priori. Classification algorithms classify the data instances into ‘p’ classes based on similarities or patterns observed in data instances [22]. Text Classification (TC) is one of the fundamental NLP under supervised ML wings. TC is the process of assigning labels, classes, categories, documents, sentences, etc., to organize and structure the text automatically.

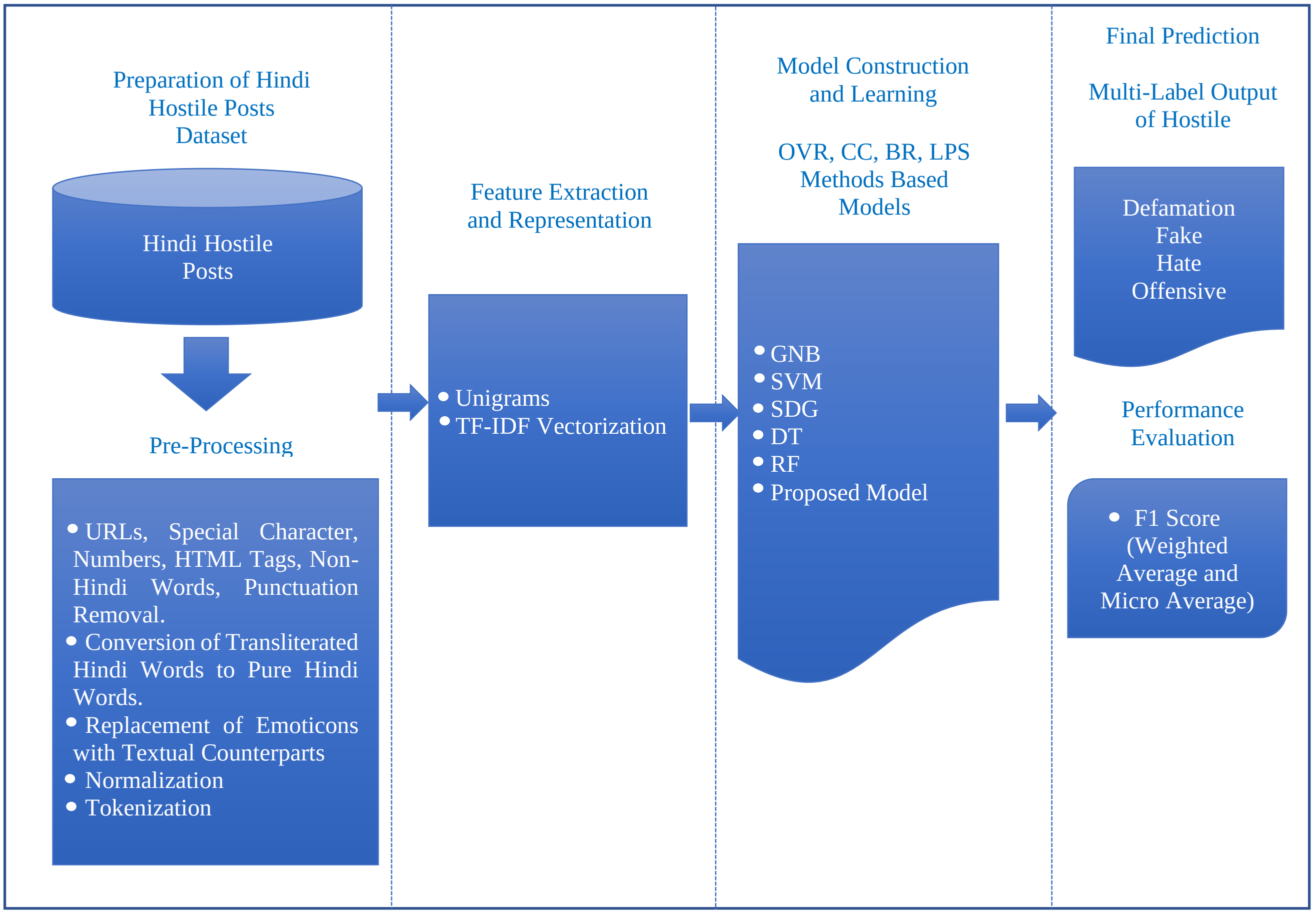
Preparation of Hindi Hostile Posts Dataset
Hindi Hostile Posts
Pre-Processing
• URLs, Special Character, Numbers, HTML Tags, Non-Hindi Words, Punctuation Removal.
• Conversion of Transliterated Hindi Words to Pure Hindi Words.
• Replacement of Emoticons with Textual Counterparts
• Normalization
• Tokenization
Feature Extraction and Representation
• Unigrams
• TF-IDF Vectorization
Model Construction and Learning
OVR, CC, BR, LPS Methods Based Models
• GNB
• SVM
• SDG
• DT
• RF
• Proposed Model
Final Prediction
Multi-Label Output of Hostile
Defamation
Fake
Hate
Offensive
Performance Evaluation
• F1 Score (Weighted Average and Micro Average)

Fig. 2 Basic framework of the proposed methodology for FGHCC

The proposed framework deals with FGHCC, and it involves four steps, namely:

- Preparation of Hindi Hostile Posts Dataset
- Feature Extraction and Representation
- Model Construction and Learning
- Final Prediction

Firstly, post-processing is accomplished under the Preparation of Hindi Hostile Posts Dataset step, as mentioned in subsection 4.1. The pre-processed posts are further entered into the feature extraction phase, wherein the relevant features from the posts are extracted, as explained in subsection 4.2. In the Model Construction and Learning step, various MLMs and proposed two-layered stacking ensembles have been applied based on OVR, BR, CC, and LPS. Lastly, in the final prediction step, multiple overlapping labels are predicted for the Hostile posts, and their performance is evaluated using the F1score metric.

4.1. Pre-processing of the Dataset

The dataset is collected from OSPs and contains a lot of incomprehensible information that needs to be removed to reduce the computational complexity. A processed corpus is created by applying custom-made functions to remove URLs, special characters, numbers, non-Hindi words, punctuation marks, and HTML tags. This dataset poses an additional challenge: along with pure Hindi words, transliterated Hindi words are also present; these were converted to actual Hindi using Indic-Trans API. Also, emoticons were replaced with their textual counterparts. Normalization is done to remove extra spacing between words, followed by tokenization [19, 26].

4.2. Feature Extraction (FE)

The mapping of words into numeric values is required to elucidate the Hindi Hostile post semantically, and FE achieves this. For PTM-based, MLMs must extract a set of features from the dataset. The features that are employed to train the model in this work are as follows [6, 19, 26]:

4.2.1. Unigrams

The n-gram consists of size 1, which is a unigram. It helps to identify which words in the posts are relevant to which hostility class. For this work, the top 2889 unigrams in the dataset are taken as one of the features for system training.

4.2.2. TF-IDF Vectorization

The models extract keywords from the posts to understand them, and this is achieved by assigning TF-IDF scores to each word in the posts. TF-IDF is used as a weighting factor. TF refers to term frequency, which measures the frequency of a word in posts, and IDF refers to inverse document frequency; it calculates the occurrence of uncommon words across all posts. It is a numerical statistic that measures a word's importance in Hindi posts. The TF-IDF considers the specificity of words and the statistical aspect of posts in the dataset, giving rare words greater weight. TfidfVectorizer can be imported from sklearn. To avoid leakage of data, TfidfVectorizer is fit_transform with training data only.

4.3. Multilabel Classification and Transformation-based Methods

Textual data is the most prominent form of data and a rich information source everywhere today [22]. It is known that social media is one of the largest sources of unstructured textual data, and getting valuable insights from it can take time and effort. TC tasks are broadly divided into three types: Binary TC [27], multiclass TC [26], and multilabel TC [28]. The number of labels or classes associated with the textual content in binary TC is two. It is common to model this problem using Bernoulli probability distribution. In Binary TC, the target label has only two possible values, which are often inverse to each other. In multiclass TC, the textual content belongs to more than two classes/categories. Multiclass classification is known as multinomial classification. Multiclass TC or multinomial classification is the classification in which the textual instance is classified into three or more classes. Multinoulli distribution is used to model this problem. Binary classification MLMs can be adapted for multiclass TC using (One vs. One) or (One vs. Rest) [29]. The motive is to predict a single class out of available classes.

In the contemporary world, conventional text label classification, also known as binary TC and multiclass TC, cannot meet the requirements of the text today as text can belong to multiple or overlapping labels. Predicting various labels associated with a single instance simultaneously is omnipresent in today's real world. This study is confined to Multilabel Classification (MLC), also known as a multi-output classifier. This kind of classification refers to the classification wherein textual posts can belong to a class greater than one and have more than one label. The textual instance can have one or more labels in multilabel textual classification. It is used when we have multiple classes or labels related to each other. MLC is a TC task that assigns a set of target labels to each training instance. It is assumed that the labels assigned are not mutually exclusive.

There are mainly two methods to deal with the MLC problem [30]. First, Problem Transformation Methods (PTMs) wherein the MLC problem is transformed into many single-label classification problems. Second is the Problem Adaptation method (PAMs), where some MLMs are adapted, meaning they are generalized to make them perform MLC. Whereas PTMs are further classified into the following categories. Binary classification transformation, similar to the One Vs. Rest method, divides the MLC problem into many independent binary classification problems; BR and CC come under this category, and multiclass classification transformation, LPS, comes under this category. It is the

MLC problem transformed into the MC classification problem. Here, the labels are combined, and one big binary classifier, namely the powerset, is formed. A detailed explanation of former and later categories is given below. The description of the multi-label-based transformation methods employed in this work are described below [31]:

4.3.1. One vs Rest (OVR)

OVR can be said to be an ensemble of binary classifiers in which the task is decomposed into several binary classification tasks in which labels are mutually exclusive. In OVR, one class is selected, and a binary classifier is trained with the samples of the selected class on one side and all the other samples on the other; thereby, we get the 'R' classifiers for 'R' labels. While evaluating, we classify the posts as belonging to the labels with the maximum score among the R classifiers.

4.3.2. Binary Relevance (BR)

BR is similar to the OVR approach of multiclass classification. This method transforms MLC with N labels into separate single-label binary classification problems. Here, every classifier predicts the membership of a class. The union of prediction by all classifiers is considered as the multilabel output. BR is a simple, popular approach. Its main drawback is that it does not consider possible class correlations. If suppose ‘n’ labels are there, this method creates ‘n’ new datasets, one for each label and every single label is trained on each new dataset.

4.3.3. Classifier Chains (CC)

This technique takes label correlation into account, which was not considered in BR. This approach shares similarities with the BR approach. It uses a chain of classifiers, where each classifier uses the prediction of all previous classifiers as input. This method is quite expensive compared to the BR method; it considers label dependencies for classification tasks. The number of classes is equivalent to the total number of classifiers, as mentioned in Eq. 1.

$$\text{Total No. of classifiers in CC} = \text{No. of classes} \tag{1}$$

4.3.4. Label Power Set (LPS)

The MLC problem is transformed into a multiclass problem. In this, a classifier is trained on all unique combinations of labels in the training dataset. It has been observed that as the number of labels in LPS increases, the number of unique label combinations also increases. This can make this approach expensive to implement. Another disadvantage is that it only predicts the label combinations in the training dataset. LPS is a PTM that matches label combinations that occur together with a combination ID and uses these combination IDs as classes, and trains classifiers accordingly. It might lead to an imbalanced dataset with label combinations. Also, it has a high evaluation complexity.

4.4. Description of the MLMs

A brief description of the MLMs employed for the FGHCC task is stated below:

4.4.1. Gaussian Naïve Bayes (GNB)

A supervised Machine Learning Algorithm (MLA) widely applied for TC tasks. It is naïve as it assumes strong independence among features and is called Bayes based on the Bayes theorem. It is mainly employed for a classification task that contains discrete features like text. GNB is imported from the sklearn naive_bayes library. In GNB, the continuous values analogous to every feature are distributed according to Normal or Gaussian distributions [30]. Following GNB, the conditional probability formula is stated below:

$$\mathrm{P}\left(\frac{\mathrm{ai}}{\mathrm{b}}\right) = \frac{1}{\sqrt{2\pi\sigma\mathrm{B}}}\exp\left(-\frac{(\mathrm{ai}-\mu\mathrm{b})_2}{2\sigma2\mathrm{b}}\right) \tag{2}$$

4.4.2. Support Vector Machine (SVM)

SVM is a discriminative model often considered one of the best “out of the box” models for the classification task. It is a generalization of the maximal margin classifier. To get better discrimination, we employed Linear SVC, which uses Squared Hinge loss for learning. The hyperparameters used are the linear kernel, C {Regularization parameter} was left default, n_jobs = -1, random_state was set to none, max_iter was set to -1 for no limit, and class_weight was balanced.

4.4.3. Decision Tree (DT)

DT is a piecewise constant approximation and a non-parametric supervised learning method widely employed for classification. The intent is to create a model that predicts the target variable value by learning simple decision rules from features in the dataset. In DT, each internal node represents a “test” on an attribute, each branch represents the test result, and each leaf node represents a class label. The classification rules are described as paths from the root to the leaf. It suffers from high variance and is extremely sensitive to the training data.

4.4.4. Random Forest (RF)

RF is an improvement over DTs that fits the number of DTs on several dataset sub-samples and utilizes averaging to enhance predictive accuracy; it also controls overfitting. For predictions, it uses the voting mechanism from an ensemble of DTs. The correlation among the features is avoided by taking a random subset of features, resulting in improved model performance. In this work, n_estimators are specified as 500, and default parameter settings are used for other parameters.

4.4.5. Logistic Regression (LR)

For an MLC problem, the LR finds the probability the posts belong to that label. Multinomial LR is used in our work; it is K-1 regression models that are combined to prognosticate labelled nominal data for supervised learning. The following parameters are taken: l1 & l2 as penalty parameters, class_weight as balanced, n_jobs as -1, and different alpha values such as {0.01, 0.1, 1, 10, 100, 1000}.

4.4.6. Stochastic Gradient Descent (SGD)

Gradient Descent (GD) is an iterative algorithm that begins with a random point on a function and plunges the slope in steps until it comes to the lowest point of that function. SGD is the modification of GD, and stochastic means random. SGD is an inexact but robust algorithm that finds the best-fit parameters between actual and predicted outputs.

This algorithm is commonly applied in NLP and text classification tasks. This algorithm randomly picks one data point from the whole dataset at each iteration to reduce the computation extensively. The number of iterations and regularization parameters are the required hyperparameters in this algorithm.

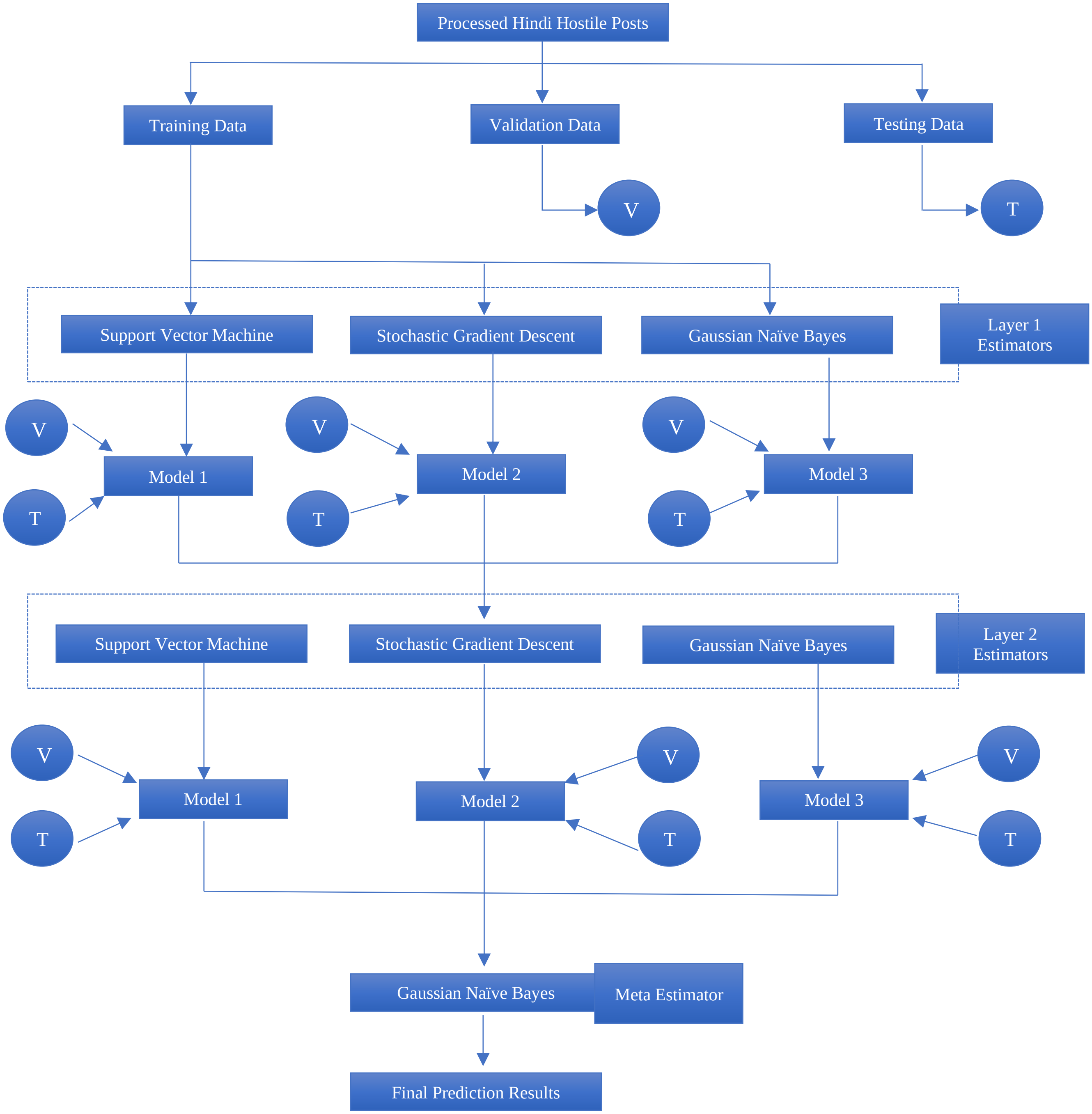

Fig. 3 Conceptual architecture of proposed two-layered stacking ensemble utilized for FGHCC.

4.5. Proposed Stacking Ensemble

It is known that the aggregated decision made by a group of people also called the crowd's wisdom, is often better than the individual's decision. The same concept is of ensembling in which multiple MLMs are applied to a problem to obtain better predictive performance, which could not have been obtained using individual MLMs alone. Stacked generalization, also known as stacking, is an ensemble technique in which predictions from several base MLMs are combined with a combiner called the meta learner in a single ensemble architecture, which is trained with the predictions from several base MLMs. The basic idea is to ensemble a robust, diverse set of base MLMs and combine them optimally by teaching a meta-learner [31]. The MLMs employed in the first layer are called base learners or weak estimators. The MLMs that are stacked on weak estimators are called meta-learners. The meta-learner is also known as a stacker or a combiner, and its role is to optimally merge the predictions made by the base learners to produce the final resultant prediction. In this work, for FGHCC, a two-layered stacking ensemble is proposed, and its conceptual architecture is given in Fig.3.

In this work, we proposed a stacking ensemble consisting of two layers and used mlxtend. In the first layer, SVM, SGD, and GNB are the employed base estimators, and they are aggregated with another layer of the exact base estimators, which constitute layer 2. Finally, the GNB is employed as the meta-estimator. Initially, the base estimators in both layers were trained with the training dataset and were tuned on the validation dataset. The predictions made by both layer 1 and 2 estimators are used as features. Finally, the trained stacking model, layer 1 & 2 estimators and meta estimators are evaluated on a testing dataset and finally predict the final overlapping hostility labels for the Hindi posts. The meta-learner GNB, in our case, has learned the strengths of base learners and complements their weaknesses. We have not used many base learners in our proposed architecture as it results in inferential latency.

The pseudo algorithm of the proposed two-layer stacking ensemble is stated beneath.

Algorithm: Two Layered Stacking Ensemble for FGHCC.

Input: Training dataset D = {(p1, l1), (p2, l2), (p3, l3), (p4, l4), ………, (pn, ln)}

Posts (p), label (l)

Testing dataset (T)

Validation dataset (V)

Meta-Learner GNB (Ml)

Output: Multilabel output predictions for the hostile posts.

```
Begin
Step1: Train the base learners on the training dataset
for posts p = 1, ……, n do
    Learn Layer 1 estimators, namely SVM, SGD, GNB
        for l do
            Learn Layer 1 estimators Lpostsl for the Training dataset
            Tune Layer 1 estimator Lpostsl on V
        end for
Step 2: for posts, do
    Learn Layer 2 estimators, namely SVM, SGD, and GNB
        for l do
            Learn Layer 2 estimators Lpostsl for the Training dataset
            Tune Layer 2 estimators Lpostsl on V
        end for
Step 3: Learn a meta-level learner M based on predictions from Layer 1 & Layer 2 estimators
M = Ml(P`)
Predict labels <l1, l2, l3, l4….> of posts p in T
Return M
end for
End
```

4.6. Performance Evaluation Metrics

Averaging-based metrics are used for multilabel classification tasks in this work [32]. We made use of Micro-Averaging and Weighted-Averaging for the performance evaluation. A confusion Matrix is a matrix that compartmentalizes correct and incorrectly classified labels into True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). We have generated a confusion matrix using sklearn's multilabel_confusion_matrix; the expected and predicted labels are passed after binarizing them.

In this work, we aim to perform FGHCC, a multilabel classification. The Weighted average (Wgt_{avg}) is calculated for the F1 score, which is the average of metric values for individual labels weighted by the support of that label. The hostile class labels for defamation, fake, hate and offensive are denoted as d, f, h, and o, respectively. F1 score is the harmonic mean of precision (Pr) and recall (Re); the formula of $\mathrm{Wgt}_{\mathrm{avg}}$ Pr and Re is given beneath.

$$\mathrm{Wgt}_{\mathrm{avg}}\mathrm{Pr} = \frac{(\mathrm{P_d.S_d+P_f.S_f+P_h.S_h+P_o.S_o})}{(\mathrm{S_d+S_f+S_h+S_o})} \tag{3}$$

$$\mathrm{Wgt}_{\mathrm{avg}}\mathrm{Re} = \frac{(\mathrm{R_d.S_d + R_f.S_f + R_h.S_h + R_o.S_o})}{(\mathrm{S_d + S_f + S_h + S_o})} \tag{4}$$

In an MLC scenario, a micro-average ($\mathrm{Micro}_{\mathrm{avg}}$) is preferable as in the employed dataset; all classes/ labels in posts are not equally distributed; that is, labels imbalance; therefore, to compute the average metric, the micro-average will aggregate the contributions of all label types, hence giving us credible results. In other words, all TPs, TNs, FPs, and FNs for each label will be summed up and averaged.

$$\mathrm{Micro}_{\mathrm{avg}}\ \mathrm{Pr} = \frac{\Sigma\ \mathrm{TP}\ (\mathrm{label}_{\mathrm{d,f,h,o}})}{\Sigma\ \mathrm{TP}\ (\mathrm{label}_{\mathrm{d,f,h,o}}) +\ \mathrm{FP}\ (\mathrm{label}_{\mathrm{d,f,h,o}})} \tag{5}$$

$$\mathrm{Micro}_{\mathrm{avg}}\ \mathrm{Re} = \frac{\Sigma\ \mathrm{TP}\ (\mathrm{label}_{\mathrm{d,f,h,o}})}{\Sigma\ \mathrm{TP}\ (\mathrm{label}_{\mathrm{d,f,h,o}}) +\ \mathrm{FN}\ (\mathrm{label}_{\mathrm{d,f,h,o}})} \tag{6}$$

$\mathrm{Wgt}_{\mathrm{avg}}$ F1 and $\mathrm{Micro}_{\mathrm{avg}}$ F1 score is the harmonic mean of Eq.3, Eq.4 and Eq.5, Eq.6, respectively.

5. Results and Discussion

This section delves into the obtained results and summarizes the findings concisely. There has been a limited amount of research concerning multilabel TC in Hindi. In addition to the nature of Hindi, there exist many other challenges, such as the dataset available being imbalanced, as in our case, for labels among some hostility classes, minimal posts are available, there are overfitting issues, and another challenge is to capture the correlation among the classes.

This study's primary focus centres on addressing the challenge of imbalanced label distributions and overfitting issues by utilizing the proposed stacked model. Previous studies have demonstrated that ensemble techniques improve the performance of individual MLMs with imbalanced class or label populations [24, 26]. Inquisitively, an ensemble of multilabel text classifiers where each MLMs belongs to different PTMs can handle the imbalance label problem in a dataset as they provide potentially more diverse, robust and independent sets of predictors.

The experimentation involved the utilization of various multilabel transformation methods such as OVR, BR, CC, and LPS on Hindi hostile overlapping classes. To perform the implementation, Python 3.11.0 was used. The main evaluation parameter for FGHCC is a $\mathrm{Wgt}_{\mathrm{avg}}$ and $\mathrm{Micro}_{\mathrm{avg}}$ F1 score; as mentioned in most works of literature [29, 32], the F1 score is a more reliable measure, as it addresses the limitation of Re and Pr when doing MLC.

All the reported results are estimated from a 10-fold CV. This paper proposes a two-layered stacking ensemble model for FGHCC on Hindi posts, as mentioned in sub-section 4.2. For Hindi multilabel TC, this idea is appealing as stacking ensembles are well known for overcoming overfitting problems and improving the performance of individual multilabel MLMs. Six widely used SOTA MLMs based on various multilabel transformation methods are applied and have been investigated along with the proposed stacking ensemble model.

Results suggest that the proposed model provides an efficient solution compared to individual multilabel MLMs. Tables 1, 2, 3, and 4 show the various multilabel transformation-based methods applied to various MLMs along with our proposed stacking ensemble methods. Individual F1 scores for all hostile classes are also reported. Fig. 4, Fig. 5, Fig. 6, and Fig. 7 shows $\mathrm{Micro}_{\mathrm{avg}}$ and $\mathrm{Wgt}_{\mathrm{avg}}$ F1 score-based comparison of OVR, BR, CC, and LPS transformation method, respectively.

When the individual MLMs are compared, SGD and SVM exhibit a good $\mathrm{Wgt}_{\mathrm{avg}}$ and $\mathrm{Micro}_{\mathrm{avg}}$ F1 score. While our proposed stacking ensemble outperformed all the individual applied MLMs and consistently performed well on $\mathrm{Wgt}_{\mathrm{avg}}$ and $\mathrm{Micro}_{\mathrm{avg}}$ F1 scores in all multilabel transformation methods, obtaining the highest $\mathrm{Wgt}_{\mathrm{avg}}$ and $\mathrm{Micro}_{\mathrm{avg}}$ F1 score of 60%, except for LPS where the proposed model performed at least as well as the best single model, SVM, and SGD in that group, this may be explained by the fact that LPS method ignores the multilabel structure of the Hindi hostility posts dataset.

As mentioned in section 4, LPS combines the multiple labels of each training post as one combined label and uses MLMs to classify the test posts. There might be some combined labels that are absent in the test set. Therefore, those samples will not be assigned any labels.

Through experiments, it is validated that the proposed stacking ensemble model yields better performance compared to SOTA MLMs applied alone.

The clustered bar graph, as shown in Fig. 8. indicates that the proposed stacking outperformed in OVR, CC, and BR transformation methods, while in LPS, it performed equivalent to the best-performing SVM and SGD in that group and Table. 5 shows the comparison of obtained $\mathrm{Wgt}_{\mathrm{avg}}$ and fine-grained average F1 score of present work with previous work.

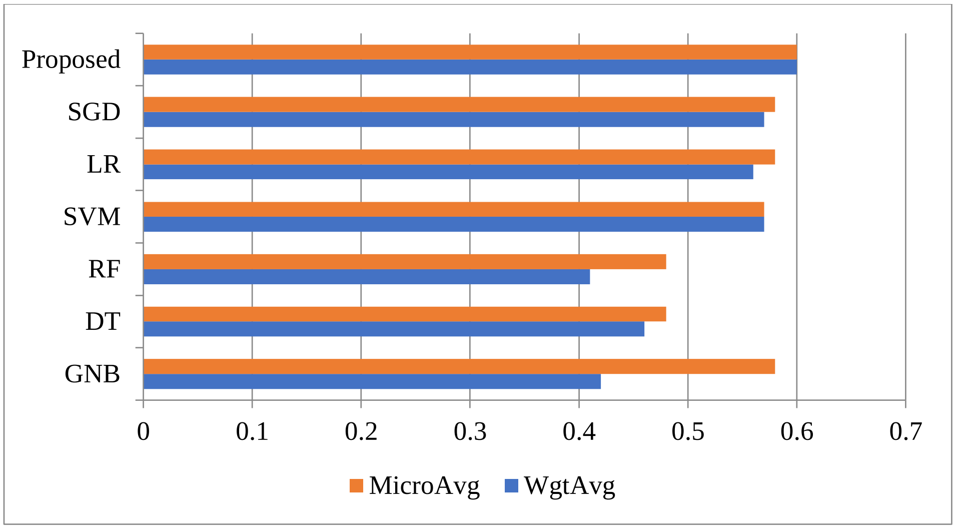

Fig. 4 Bar Graph-based comparison using OVR Transformation

Table 1. One Vs Rest transformation method

OVR	**Hostile Classes**				**F1 Score**	
	Defame	**Fake**	**Hate**	**Offensive**	**Wgt$_{avg}$**	**Micro$_{avg}$**
GNB	0.01	0.75	0.20	0.47	0.42	0.58
DT	0.26	0.62	0.34	0.50	0.46	0.48
RF	0.06	0.73	0.18	0.44	0.41	0.48
SVM	0.33	0.77	0.46	0.55	0.57	0.57
LR	0.34	0.78	0.45	0.54	0.56	0.58
SGD	0.36	0.76	0.46	0.56	0.57	0.58
Proposed	0.37	0.77	0.51	0.59	**0.60**	**0.60**

Table 2. Binary Relevance transformation method

BR	**Hostile Classes**				**F1 Score**	
	Defame	**Fake**	**Hate**	**Offensive**	**Wgt$_{avg}$**	**Micro$_{avg}$**
GNB	0.36	0.70	0.46	0.50	0.53	0.53
DT	0.31	0.63	0.35	0.55	0.48	0.48
RF	0.17	0.71	0.21	0.55	0.46	0.45
SVM	0.08	0.77	0.19	0.52	0.45	0.45
LR	0.34	0.78	0.45	0.54	0.56	0.44
SGD	0.37	0.76	0.47	0.57	0.58	0.56
Proposed	0.37	0.77	0.51	0.59	**0.60**	**0.60**

Table 3. Classifier Chains transformation method

CC	**Hostile Classes**				**F1 Score**	
	Defame	**Fake**	**Hate**	**Offensive**	**Wgt$_{avg}$**	**Micro$_{avg}$**
GNB	0.36	0.69	0.46	0.50	0.52	0.53
DT	0.31	0.63	0.40	0.50	0.49	0.49
RF	0.22	0.73	0.34	0.58	0.54	0.51
SVM	0.08	0.78	0.57	0.39	0.57	0.51
LR	0.03	0.78	0.55	0.45	0.57	0.52
SGD	0.36	0.76	0.51	0.58	0.59	0.59
Proposed	0.36	0.77	0.53	0.58	**0.60**	**0.60**

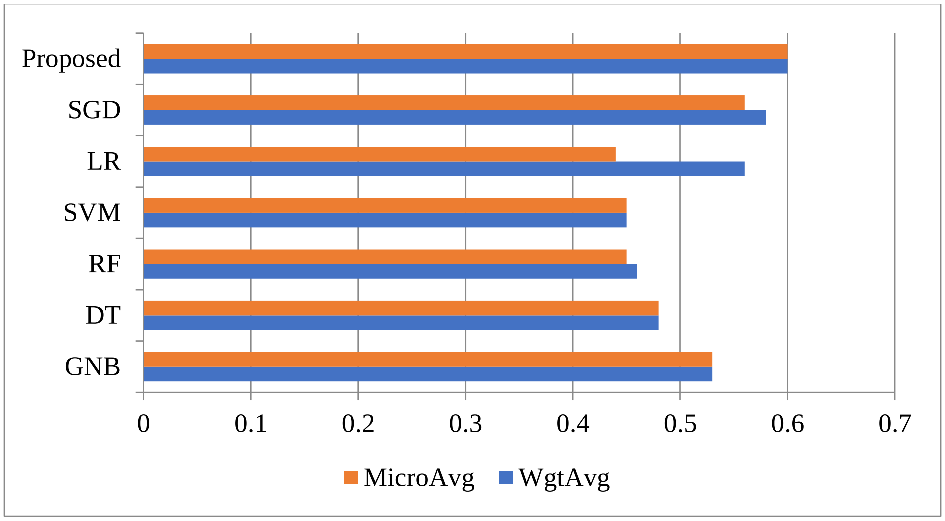

Fig. 5 Bar graph-based comparison using BR transformation

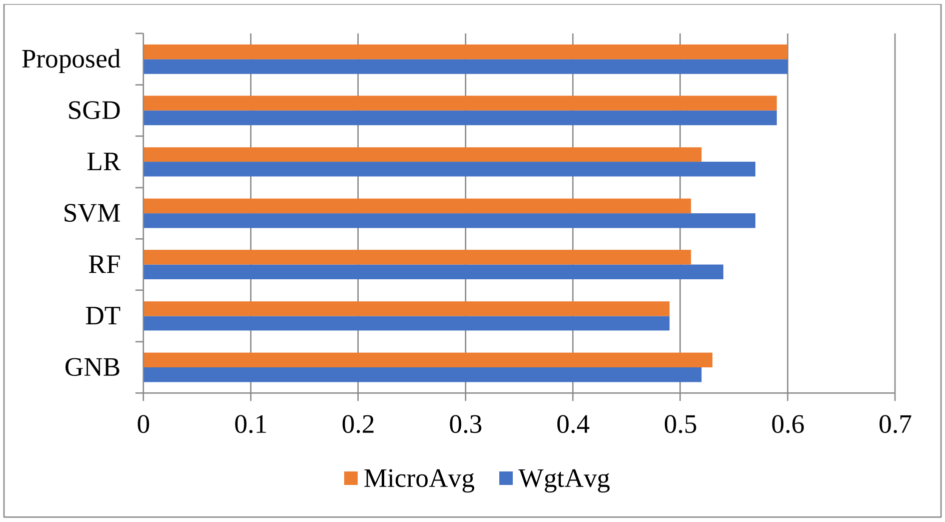

Fig. 6 Bar graph-based comparison using CC transformation

Table 4. Label Power Set transformation method

LPS	**Hostile Classes**				**F1 Score**	
	Defame	**Fake**	**Hate**	**Offensive**	**Wgt**$_{\text{avg}}$	**Micro**$_{\text{avg}}$
GNB	0.25	0.67	0.43	0.44	0.50	0.48
DT	0.23	0.65	0.37	0.49	0.47	0.47
RF	0.11	0.68	0.30	0.55	0.52	0.46
SVM	0.31	0.76	0.49	0.57	0.57	0.57
LR	0.18	0.74	0.41	0.56	0.57	0.52
SGD	0.33	0.75	0.47	0.58	0.57	0.57
Proposed	0.40	0.74	0.47	0.55	0.57	0.57

Table 5. Weighted average Fine-Grained F1 score of previous work with proposed architecture

Previous Work	**Best Weighted Fine Grained F1 Score for fine-grained hostility detection**
[22]	50.98 ~0.51
[23]	0.53
Proposed Stacking Ensemble	**0.60**

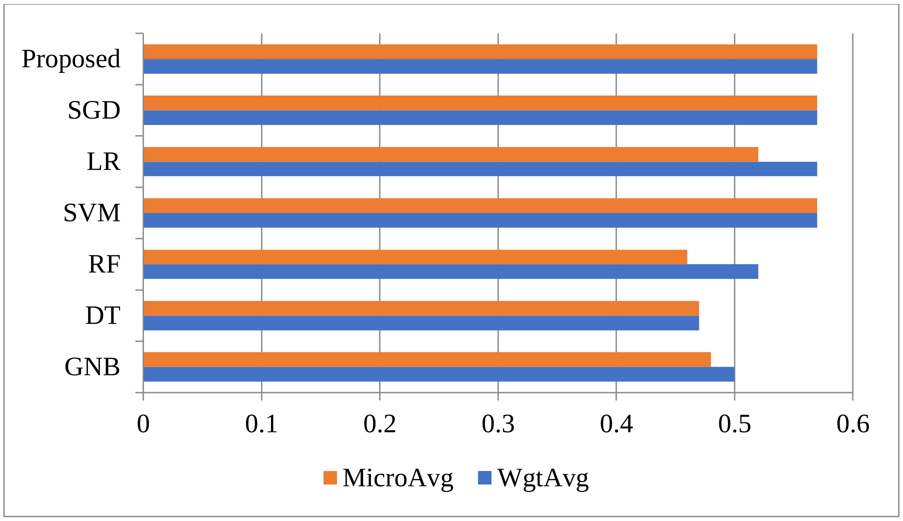

Fig. 7 Bar graph-based comparison using LPS Transformation

Table 5 shows the comparison of the proposed stacking ensemble architecture with some previous work. Our proposed stacking architecture has obtained better performance than the complex models applied, demonstrating that achieving success in FGHCC tasks doesn't necessarily require using large, intricate models like DLMs. Instead, it suggests that exploring the potential of an ensemble consisting of MLMs based on PTMs is a valuable approach to consider.

To the best of the author's knowledge, an enhanced approach to simultaneously address both post-imbalance and overfitting issues is introduced in this work. Instead of focusing solely on ensemble techniques within a multilabel learner, the proposed approach combines cutting-edge multilabel classifiers based on Problem Transformation Methods (PTMs) into a stacking ensemble architecture. This integration of multilabel classifiers, each employing PTMs, offers a more diverse and independent set of predictions, which helps mitigate the imbalance problem. Furthermore, the stacking ensemble inherently addresses overfitting concerns, ultimately enhancing the overall classification performance.

Results indicate that our proposed model outperformed the SOTA MLMs, considering both Wgt$_{avg}$ and Micro$_{avg}$ F1 scores. The results show the stacked ensemble's merit for FGHCC to overcome the overfitting and the imbalanced label distribution problem and improve performance. Previous studies have shown that ensemble techniques, such as stacking, can significantly enhance the performance of individual MLMs. Stacking is especially effective in mitigating overfitting issues and addressing class label imbalances within a dataset.

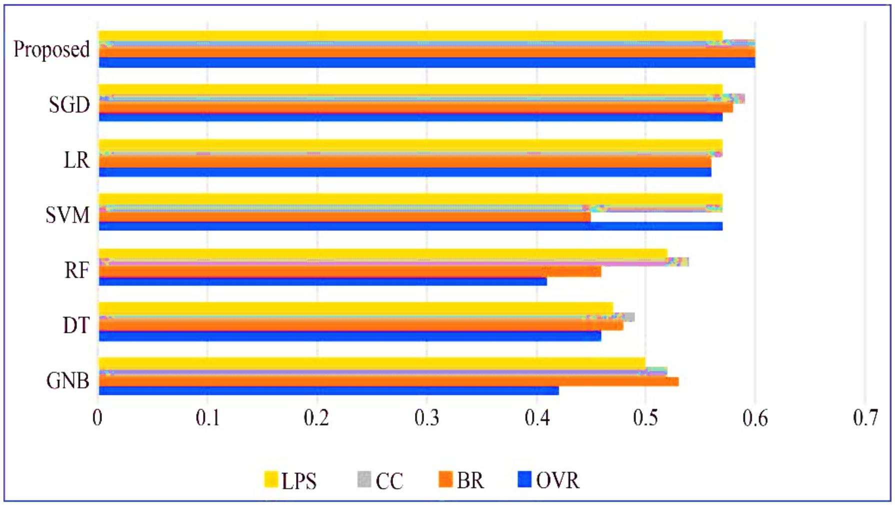

Fig. 8 Weighted average F1 Score-based comparison for all transformation methods

Consequently, incorporating multiple MLMs based on PTMs into the proposed framework can yield a more resilient predictive model. As mentioned, the suggested architecture manages class label imbalances by furnishing a potentially broader, more robust, and independently derived set of predictions. Furthermore, the stacked generalization approach effectively amalgamates diverse model types, ultimately resulting in reduced variance.

Our analysis of results showed that for Hindi FGHCC, a large, robust gold-standard dataset is required due to the wide application of hostility class detection in the present scenario. Incorporating more textual posts in the hostility dataset can enhance the performance of the current implementation, as in the future, the dataset employed would be extended to include more overlapping hostility classes. To the author's knowledge, this is the first study to combine MLMs into a two-layered stacked ensemble based on PTMs. Since Hindi multilabel TC is inherently computationally intensive and the label distribution in the dataset is imbalanced, it opens up new research challenges regarding how to choose efficient base estimators for the different layers since different combinations of base estimators may perform differently for a particular problem domain. Based on the experiments conducted, it can be argued that the proposed stacking ensemble model demonstrates promising F1 score results, provided there is adequate availability of overlapping hostile class datasets to facilitate generalization.

6. Conclusion

This paper deals with multilabel TC for FGHCC tasks in Hindi posts. Hindi FGHCC is a complex process partly due to its resource-poor nature, the unavailability of adequate datasets, and imbalanced label distribution in available datasets. To this end, a robust two-layered stacking ensemble architecture based on the PTMs is proposed for more FGHCC tasks. Firstly, data preparation and pre-processing are performed on the dataset, and afterwards, the proposed stacking ensemble and some well-known SOTA baseline classifiers such as GNB, DT, RF, SVM, LR, and SGD, with TF-IDF and unigrams as features were evaluated using the various multiclass transformation-based methods (OVR, BR, CC, and LPS). The main evaluation parameter used for evaluating the model is the Wgt$_{avg}$ F1 and Micro$_{avg}$ F1 score. The experimental outcomes demonstrated that the employed stacking ensemble model, incorporating OVR, CC, and BR, outperformed all the MLMs applied based on PTMs and performed the best. The strength of the proposed ensemble lies in its simplicity; it requires fewer computational resources and is best for overcoming over-fitting and imbalance label distribution problems and improving performance, thereby giving an efficient solution compared to the SOTA multilabel methods.

Furthermore, there is an ongoing effort to augment the number of hostile posts within the utilized dataset, intending to provide this expanded dataset to the research community eventually. Moreover, in the context of future endeavours, there is a strategic plan to broaden the applicability of the proposed model for FGHCC to encompass other low-resource Indian languages, such as Punjabi and Marathi, among others. Consequently, the stacking-based model is anticipated to provide valuable support to the social media backend team in assessing online post content to categorize and eliminate hostile posts from the internet. An additional intriguing avenue for future exploration involves the incorporation of PAMs in conjunction with weighting strategies to address label imbalance during the model training process.

Acknowledgments

The research work has been made possible with the support of the Indraprastha Research Fellowship (IPRF) from Guru Gobind Singh Indraprastha University, under award letter number GGSIPU/DRC/2022/1247.

References

[1] Gopendra Vikram Singh et al., “EmoInHindi: A Multi-Label Emotion and Intensity Annotated Dataset in Hindi for Emotion Recognition in Dialogues,” *arXiv*, pp. 1-9, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Akinbohun Folake, Akinbohun Ambrose, and E. Oyinloye Oghenerukevwe, “Stacked Ensemble Model for Hepatitis in Healthcare System,” *International Journal of Computer and Organization Trends*, vol. 9, no. 4, pp. 25-29, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ramchandra Joshi et al., “Evaluation of Deep Learning Models for Hostility Detection in Hindi Text,” *6th International Conference for Convergence in Technology*, pp. 1-5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Asif Hasan et al., “Analysing Hate Speech against Migrants and Women through Tweets Using Ensembled Deep Learning Model,” *Computational Intelligence and Neuroscience*, pp. 1-8, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] I. Gede Manggala Putra, and Dade Nurjanah, “Hate Speech Detection Indonesian Language Instagram,” *International Conference on Advanced Computer Science and Information Systems*, pp. 413-420, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Fatimah Alkomah, and Xiaogang Ma, “A Literature Review of Textual Hate Speech Detection Methods and Datasets,” *Information*, vol. 13, no. 6, pp. 1-22, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Vandana Jha et al., “Sentiment Analysis in a Resource Scarce Language: Hindi,” *International Journal of Scientific and Engineering Research*, vol. 7, no. 9, pp. 968-980, 2016. [Google Scholar] [Publisher Link]

[8] Abdalsamad Keramatfar, and Hossein Amirkhani, “Bibliometrics of Sentiment Analysis Literature,” *Journal of Information Science*, vol. 45, no. 1, pp. 3-15, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[9] Dhanashree S. Kulkarni, and Sunil S. Rodd, “Sentiment Analysis in Hindi-A Survey on the State-of-the-Art Techniques,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 1, pp. 1-46, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Pratik Joshi et al., “The State and Fate of Linguistic Diversity and Inclusion in the NLP World,” *arXiv*, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[11] Abhishek Velankar et al., “Hate and Offensive Speech Detection in Hindi and Marathi,” *arXiv*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Fabio Poletto et al., “Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review,” *Language Resources and Evaluation*, vol. 55, no. 2, pp. 477-523, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Md Saroar Jahan, and Mourad Oussalah, “A Systematic Review of Hate Speech Automatic Detection using Natural Language Processing,” *Neurocomputing*, vol. 546, pp. 1-30, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Debanjana Kar et al., “No Rumours Please! A Multi-Indic-Lingual Approach for COVID Fake-Tweet Detection,” *Grace Hopper Celebration India*, pp. 1-5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] Akshaya Gangurde et al., “A Systematic Bibliometric Analysis of Hate Speech Detection on Social Media Sites,” *Journal of Scientometric Research*, vol. 11, no. 1, pp. 100-111, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[16] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer, “Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language,” *14th Conference on Natural Language Processing - KONVENS 2018*, pp. 1-10, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[17] Marcos Zampieri et al., “SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (Offenseval),” *arXiv*, pp. 1-12, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[18] Marcos Zampieri et al., “SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval2020),” *arXiv*, pp. 1-23, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[19] Arjun Roy et al., “An Ensemble Approach for Aggression Identification in English and Hindi Text,” *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying*, pp. 66-73, 2018. [Google Scholar] [Publisher Link]

[20] Shyam Ratan, Sonal Sinha, and Siddharth Singh, “SVM for Hate Speech and Offensive Content Detection,” *CEUR Workshop Proceedings*, pp. 1-8, 2021. [Google Scholar] [Publisher Link]

[21] Mohit Bhardwaj et al., “Hostility Detection Dataset in Hindi,” *arXiv*, pp. 1-5, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[22] Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque, “Combating Hostility: Covid-19 Fake News and Hostile Post Detection in Social Media,” *arXiv*, pp. 1-11, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[23] Mohammed Azhan, and Mohammad Ahmad, “LaDiff ULMFiT: A Layer Differentiated Training Approach for ULMFiT,” *International Workshop on Combating Online Hostile Posts in Regional Languages During Emergency Situation*, pp. 54-61, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[24] Chander Shekhar et al., “Walk in Wild: An Ensemble Approach for Hostility Detection in Hindi Posts,” *arXiv*, pp. 1-10, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Ayush Gupta et al., “Hostility Detection and Covid-19 Fake News Detection in Social Media,” *arXiv*, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[26] Tanzia Parvin, and Mohammed Moshiul Hoque, “An Ensemble Technique to Classify Multiclass Textual Emotion,” *Procedia Computer Science*, vol. 193, pp. 72-81, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[27] Marcin Michal Mirończuk, and Jaroslaw Protasiewicz, “A Recent Overview of the State-of-the-Art Elements of Text Classification,” *Expert Systems with Applications*, vol. 106, pp. 36-54, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[28] Nurshahira Endut et al., “A Systematic Literature Review on Multilabel Classification Based on Machine Learning Algorithms,” *TEM Journal*, vol. 11, no. 2, pp. 658-666, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[29] Bassam Al-Salemi et al., “Multilabel Arabic Text Categorization: A Benchmark and Baseline Comparison of Multilabel Learning Algorithms,” *Information Processing and Management*, vol. 56, no. 1, pp. 212-227, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[30] Hozayfa El Rifai, Leen Al-Qadi, and Ashraf Elnagar, “Arabic Text Classification: The Need for Multi-Labeling Systems,” *Neural Computing and Applications*, vol. 34, no. 2, pp. 1135-1159, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[31] David H. Wolpert, “Stacked Generalization,” *Neural Networks*, vol. 5, no. 2, pp. 241-259, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[32] Shivang Agarwal, and C. Ravindranath Chowdary, “Combating Hate Speech Using an Adaptive Ensemble Learning Model with a Case Study on COVID-19,” *Expert Systems with Applications*, vol. 185, pp. 1-9, 2021. [CrossRef] [Google Scholar] [Publisher Link]