International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

The Grasshopper Optimization Technique for Hate Speech Detection on Multimodal Dataset

Annu Dhankhar[1], Amresh Prakash[2], Sapna Juneja[3]

Submitted:11/03/2024 **Revised**: 26/04/2024 **Accepted**: 03/05/2024

Abstract: Interest in multi-modal issues has increased recently, from image captioning to addressing visual questions and beyond. Online hate speech is a huge social problem nowadays, harming both individuals and society. One new kind of hostile communication, known as a "hateful meme," has arisen among them. Hate speech affects how minorities are viewed by society, even though it is not always connected to hate crimes Despite hate crimes being a public health problem, hate speech is not one in the United States. Identifying hate speech as a public health concern degrades the effects on victims and downplays hate crimes, while clearly recognizing hate speech as such downplays the act and calls for action, such as the creation of new rules or the allocation of resources to assist victims. So, hate speech identification with optimized way helps in international cooperation. Hateful memes were constructed with both text captions and images to reflect users' intentions; therefore, it is impossible to identify them with precision by only looking at the embedded text captions or photos. Identifying hate speech in multimodal memes is the new challenge set for multimodal categorization proposed in this work. Due to the addition of challenging cases to the dataset, it is difficult to rely on unimodal signals and only multimodal models may be successful. With the help of an effective feature selection technique, a grasshopper optimization algorithm GOA, and a transfer learning model VGG16, we concentrated on the identification of hate speech in multi-modal memes in this study. We attempt to resolve the Facebook Meme Challenge, a problem of binary classification that asks if a meme is hateful or not. We also include the feature selection optimization approach in addition to the multi-modal representations derived from the pre-trained model. Our model (GOA+VGG16) outperformed the other baseline models in a public test set by achieving an accuracy of 87 percent on the hateful meme identification task after using optimization algorithms and the VGG16 model and linking to a random forest (RF) classifier.

Keywords: *Hate speech, Multimodal, Optimization, Deep Learning, Machine Learning*

1. Introduction

This As social media is becoming more popular in today's lives, users are sharing their views via various modalities like text, images, Audios, Videos and Emoji's. Due to the tremendous active number of users, hate and trolling becomes a part of one's personality. Hate is defined as one of the strong feelings of not liking someone or a group of people, may resulting in inciting violence. This thinking has drastically impacted society. Hate speech is not a public health concern in the United States, despite hate crimes being one [27]. The social media platforms have surged forward to find practical and effective solutions that can help in real time predictions. Majorly and most common social media platforms like Facebook, Instagram, Twitter, WhatsApp, YouTube etc. have their certain standards for avoiding some objectionable contents to some extent [1]. The objective of machine learning is to analyze enormous volumes of data every day and aims to keep the algorithms and the growth of the data in sync. When creating an ML model, a variety of criteria are taken into account, one of which is the quantity of input characteristics that have a major impact on the system's performance [2]. The amount of information produced nowadays combined with the input features present researchers with a significant and interesting challenge because more data producing more features has a detrimental impact on the effectiveness of the previous approaches. Because of this, preprocessing data and the creation of effective algorithms for information extraction are required to separate vital features from non-essential ones in today's datasets. Preprocessing the data is as important for effective machine learning as modelling the data. These datasets are commonly preprocessed for two purposes: (1) to reduce the size of the dataset for more effective analysis, and (2) to customize the dataset to best suit the selected analysis method. These contributions from data preprocessing aid in feature selection, enhancing the efficiency of ML algorithms. Therefore, it becomes increasingly important to compress datasets methodically as their quantity and variety increase. The majority of research publications appear to have concentrated on a specific methodology centered around feature selection methods or preprocessing strategies based entirely on optimization algorithms for numerical data [3]. Despite the extensive study of various strategies provided in these works, they frequently fall short of offering a more adaptable strategy for addressing both generalized procedures and feature selection methods, which are both important components of data preprocessing. By providing a thorough comparative examination of the results obtained by four distinct machine learning models when used with

and without an optimization technique, this paper seeks to close the gap.

The first illustration in Fig. 1 shows a caption that is non-hateful whereas the image and caption together convey a hateful message. In the second example, it is shown how the meme denigrates a minority by criticizing religion. The indications in the image, particularly the attire of the figures, make this obvious. The third example involves shaming supporters of Hillary, the Democratic candidate for president in the 2016 U.S. election, and the last one effectively expresses animosity against a particular religion by excluding people from the country.

Multimodal datasets including photos and captions are employed in this study activity to further our understanding of the topic at hand. Therefore, we examined databases of nasty memes. An area of Natural Language Processing (NLP) called "hate content analysis" deals with the automatic recognition and classification of subjective information in text data, which is frequently phrased as "hate," and "no hate." Hate speech, on the other hand, refers to any type of speech that disparages or disparages an individual or a group based on that individual's or that group's race, religion, ethnicity, sexual orientation, or other identities [4]. In many nations, it is not protected by rules governing freedom of expression because it is regarded as destructive and unpleasant. To ascertain the writer's emotional tone or viewpoint, this process entails analyzing the text data.

The strength of the proposed architecture is its ability to make the search process optimized in exploration and exploitation phase, thereby improving the model's performances. Results from experiments validate the suggested model. The followings are the work's primary contributions:

- We presented a novel twofold branch architecture in which the vision branch uses VGG-16 architecture and caption branch contains grasshopper optimization technique for making the search process optimized.
- We conduct in-depth analysis and experiments Hateful Memes [5] dataset, beating the existing state-of-the-art.

The remainder of this paper is divided into the following sections. Textual, visual, and multimodal features are taken into consideration in Section 2's description of related research in this area. The proposed work and the optimization algorithm are described in Section 3 and Section 4 respectively, and dataset description is described in Section 5. Section 6 covers the result and experimental analysis. The paper concludes with additional recommendations for future research, which are discussed in section 7 and Section 8 respectively.

Fig 1. Examples of hateful memes [5][25][26]

2. RELATED WORK

We divide the relevant work into three subsections and present them as follows: The related work on Textual Aspect and Visual Aspect are covered in section 2.1 and section 2.2, respectively. The related research on Multimodal Aspect is covered in sub section 2.3.

2.1. Textual Aspect

A natural language processing application analyses and extracts the emotions from text. Sentiment analysis, hate speech detection, emotion detection, and sarcasm detection are just a few of the many uses. Since the number of users and their opinions are growing daily, hate speech identification is the research area that concerns researchers the most. The fundamental difficulty in detecting hate speech is highlighted in [6], where the authors state unequivocally that there is no accepted definition of hate speech, making detection particularly difficult. The study linked to the identification of hate speech on the English language specifically is extensively resourced in ([7], [8], [9]). The work in this paper is mostly focused on the English language. Binary text classification is the main technique for detecting hate speech [10]. Today's study also includes methods for identifying many types of hatred, including radicalization ([11], [12]), terrorism-related ([13], [14]), cyberbullying ([15], [16]), discrimination based on religion [17], racism, and sexism . The ultimate choice regarding the class can be determined by adding the results of the pre-decision. [18] demonstrates how conventional features gleaned from news articles outperform earlier models created utilizing text embedding techniques. The majority of online publications in this

topic are only available as text, and the majority of the work has been done in English [6]. The procedure is complicated and time-consuming because traditional machine learning methods rely on feature engineering [18]. Various standard learning, deep learning, and fuzzy logic classifiers are used in the performance analysis measured in [19] for the purpose of detecting hate speech, demonstrating the remarkable performance of fuzzy classifiers.

2.2. Visual Aspect

Deep learning models are used because of the enormous rise in data across many modalities. Deep learning techniques are used to automatically extract information from memes that include both text and graphics. Much of the research done so far depends on object recognition simply using photos. Due to its advantage of processing image pixels, CNN is extensively used in machine vision jobs. It has also been recognized that during the training phase, error gradients in deep networks or recurrent neural networks can accumulate and produce very large gradients. A bidirectional long- and short-term memory (Bi-LSTM) neural network fed with the CNN features is used to study the advantages of effectively avoiding gradient explosion. Convolution Neural Networks (CNNs) models are the most popular in image-based hate content detection because they are the best at spotting objectionable material such as nudity ([20], [21], [22]), appropriate or inappropriate images for children, offensive or non-compliant logos, and pornographic web pages [21]

2.3. Multimodal Aspect

Up to now, the work in this area has concentrated on unimodal characteristics, that is, either text or image separately. Today, online memes are the most popular type of information on social media platforms, and they create text-accompanied images. In SemEval2020 [10], a common job on the analysis of emotions was already accessible. There are incredibly few datasets available in this area. Prior to working with multimodal data, the algorithms' dependability and robustness should be considered. [5] created a popular Facebook challenge dataset for hostile memes in May 2020, and after the annotation procedure, the dataset had a human accuracy score of 84.70%. [23] implemented multi-scale visual kernel and distil Bert on hateful memes dataset, giving the accuracy and AUC scores with 87.50% and 83.83% respectively. For the prediction of hate material, all the numerous methodologies discussed above have considered various machine learning and deep learning techniques. The study before it shows that there aren't many publicly accessible datasets and studies that have discussed hate memes. As a result, this is a brand-new field with a huge unresolved issue. Our proposed approach, as noted in section 3, is innovative since, to the best of our knowledge, no study has previously used bio-inspired algorithm like grasshopper optimization.

3. PURPOSED ARCHITECTURE

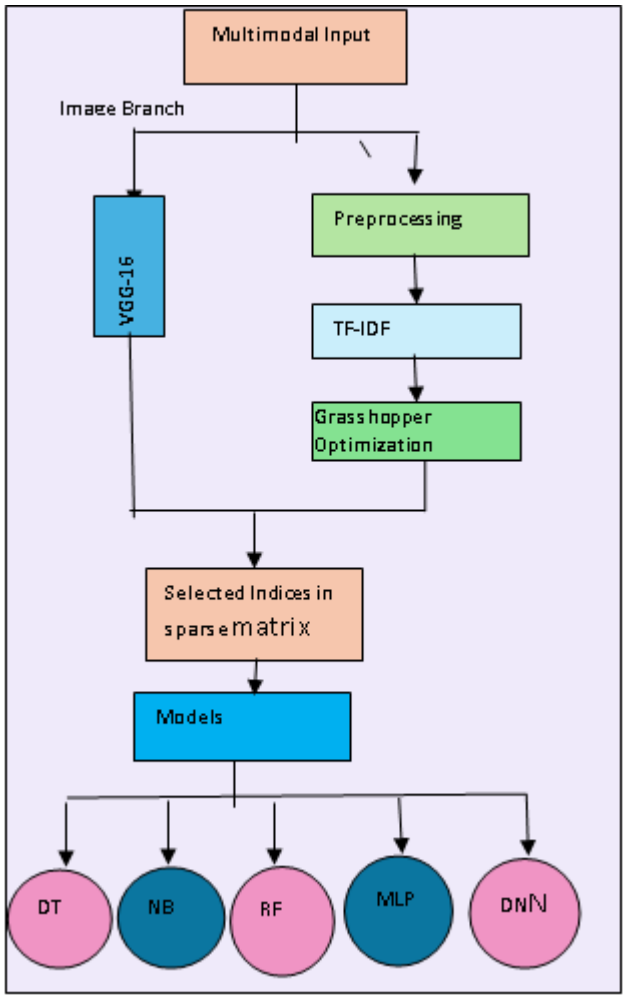

Fig 2a: Proposed Architecture with ML models

Our input dataset contains both image and text as explained in figure 2a. The resulting caption means text part was encoded using the pre-trained VGG16 model, which has a dimension of 768. The VGG model, commonly known as VGGNet is referred to as VGG16. It is a 16-layer convolution neural network (CNN) model. In ImageNet, a dataset that contains more than 14 million training photos over 1000 item classes, the VGG16 model can reach a test accuracy of 92.7%. One of the best entries in the ILSVRC competition. VGG16 enhances Alex Net by substituting sequences of smaller 33 filters for the big filters. For the first convolutional layer in Alex Net, the kernel size is 11, while for the second layer, it is 5, We are working with a text dataset on the opposite branch. To prepare the dataset, we removed special characters, changed the font size to lower case, and eliminated stop words. Then, using a statistical feature extraction method called TF-IDF, we put it into practice. Inverse Document Frequency and Term Frequency are the two statistical techniques used by TF-IDF. The phrase "term frequency" refers to the total number of times a certain term (t)

appears in the text (doc) in comparison to the total number of words in the text. The amount of information a word delivers is gauged by its inverse document frequency. The critical step came next when we extracted the feature using the grasshopper optimization technique. Grasshopper optimization algorithm GOA is a population-based algorithm and a meta-heuristics algorithm. Cross-validate five times to obtain a more reliable accuracy estimate. The two results are then combined and sent via five different classifier MLP classifier, Decision Tree DT, Naïve Bayes NB, DNN, Random Forest RF, MLP. A class called Decision Tree Classifier may do multi-class classification on a dataset. Naive Bayes is a straightforward method for building classifiers. These models assign class labels to problem cases, which are represented as vectors of feature values, and the class labels are chosen from a finite set. To do classification on related, unlabelled data, the DNN command creates a feedforward multilayer neural network that is trained with a collection of labelled data. The most fundamental style of neural network architecture is the multilayer perceptron (MLP). No single ML method outperformed all others, according to numerous research. Therefore, it is necessary to compare different ML algorithms to determine which one performs the best on the provided dataset. Using a random forest classifier, we achieved the best results.

On the same dataset we also applied deep learning model and its explanation is mentioned in figure 2b. In this data flow diagram, we explain how we split our input dataset and then also applied GOA grasshopper optimization algorithm to optimize the feature we selected. Finally, we applied three deep learning models i.e. CNN, resent and Dense net to find their accuracy, precision, Recall and f-measure

4. Optimization Algorithm

Practically every area, from engineering to economics, makes use of optimization. Because there is a limited amount of time and resources, it is crucial to make the most of what is available. Most optimizations in the real world are extremely non-linear and multimodal while also being subject to a variety of challenging constraints. Figures 3a and 3b explain how the feature selection process for the grasshopper optimization algorithm works. We are employing the Grasshopper optimization algorithm in this paper since it produces superior results. The use of GOA in our suggested architecture is explained here.

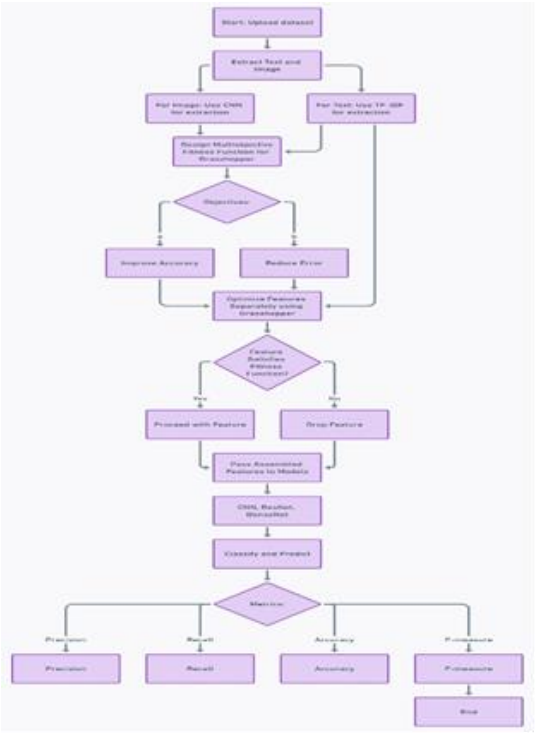

Figure 2b: Purposed Architecture with DL model

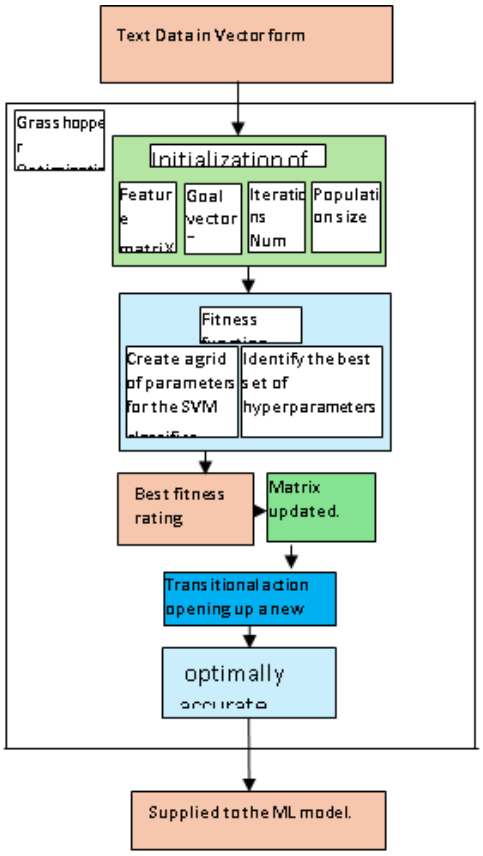

Fig 3a: Grasshopper's feature selection optimization technique

4.1. First-time setup

Provide the feature matrix A, the goal vector B, the number of iterations Num iterations, and the population size to initialize the algorithm, produce a starting

population of binary vectors, where each one represents a potential subset of features. Whether a feature is selected (True) or not (False) is indicated by each binary value construction,

4.1.1. Loop for Grasshopper Optimization

From 1 to Num iterations, for each iteration, Utilize the Fitness1 function to determine the fitness value for each binary vector in the population. This function applies cross-validation and SVM training to assess the quality of the chosen feature, Use the Update Population function to update the population. In this step, the most suitable people (binary vectors) are chosen, and new individuals are produced via cloning with some modification.

Algorithm 1 Grasshopper Algorithm for Feature Selection

```
Input: features, classLabels, maxIterations, stepSize
Output: selectedFeatures
oldbest = 0.50
grasshopperPositions = randi([0, 1], numFeatures, 1)
for iteration = 1 to maxIterations do
    Compute fitness values using calculateFitness
    Update grasshopperPositions randomly
    Identify the best grasshopper based on fitness
    Update bestGrasshopper
    Store fitness values in allitrfitness
    for i = 1 to numFeatures do
        if rand() < 0.5 then
            Update grasshopperPositions(i) positively
        else
            Update grasshopperPositions(i) negatively
        end if
    end for
    Compute new fitness values using calculateFitness
    Identify best global solution
    if bestFitness > oldbest then
        Update optimalSolution with bestGlobalGrasshopper
        Update oldbest with bestFitness
    end if
end for
Select final features based on optimalSolution
selectedFeatures = optimalSolution ≥ 0.5
```

Fig 3b: Grasshopper optimization Algorithm.

4.1.2. Fitness function

Create a grid of parameters for the SVM classifier. For the Support Vector Machine, this grid comprises various iterations of the hyperparameters (C, kernel, and gamma), Develop a classifier using SVM, use cross-validation and a grid search to identify the best set of hyper parameters that produces the maximum accuracy, Provide the grid search result with the highest cross-validated accuracy possible.

4.1.3. Update on Population function

To find the population's fittest members, sort the fitness values in descending order, a brand-new population array Perform the following operations for each person (binary vector) in the sorted list of people: Include the person in the new population possible, create a clone of the person and carry out a mutation procedure Flip the value for each feature of the individual with a low probability (0.05 in this case) by selecting it if it is not currently selected or vice versa This introduces diversity and randomness into the population. Incorporate the mutant populace.to the fresh population, Replicate the fresh populace,

4.1.4. The Best Grasshopper picks

Choose the binary vector (feature subset) from the final population with the best fitness (accuracy) when all iterations are finished,

4.1.5. Bring Back Particular Features

Provide the feature subset that corresponds to the top grasshopper (binary vector) discovered through the optimization procedure.

5. Dataset Description

Facebook AI made a data set available to identify the multimodal hate messages spread through internet memes. a set of around 10,000 PNG pictures that have been further divided into training and testing files. Three annotators who are binary labels each review a different portion of the collection. The meme photos themselves and string representations of the text in the image are the features in this collection. The dataset includes five distinct meme types[24]. Every image in both the training and validation sets has an annotation of either 1 or 0, which stands for the classes for harmful and benign memes, respectively.

6. Results and Experiment Analysis

Three publicly accessible datasets that are especially related to hate or offensive content were used in our experimentation. Most of the photographs in these databases include textual and visual information that imply different things, which is the first problem we discovered. The second problem has to do with how different image channels are sized. By using the grasshopper optimization algorithm for feature selection on a text data set and fusing that output with an image dataset's output from a pretrained VGG16 model, we may obtain the findings in table 1.

TABLE 1: Results obtained after applied Grasshopper Optimization Algorithm

Model	Precision	Recall	F-measure	Accuracy
DT	85%	85%	84%	85%
Naïve Bayes	75%	86%	85%	86%
Random Forest	88%	87%	82%	87%
MLP	79%	85%	81%	85%
DNN	75%	86%	83%	86%

Recall, f-measure, precision, and accuracy were attained by our intended architecture, VGG16 +GOA, as shown in Figures 4,5,6,7. By dividing the number of positive samples that were correctly identified as positive by the

total number of positive samples, recall—also known as sensitivity—is determined. It evaluates a model's capacity to identify positives; the higher the recall, the greater the number of positives identified. You must first decide what defines a "positive" sample before you can calculate recall. This will provide you with a precise assessment of the efficiency with which

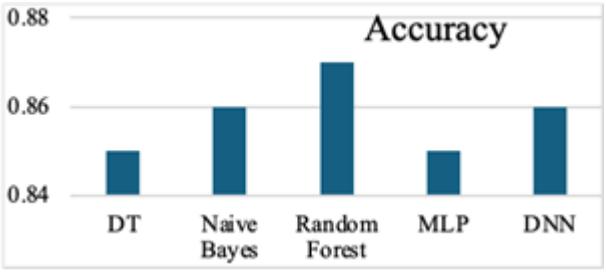

Fig 4:Accuracy for Purposed Architecture

your algorithm identified positives out of all potential positives. We accurately predicted 87% of the positive classes with RF model.The highest recall level is desirable. The statistic used to assess the effectiveness of a machine learning model is the F-score, also known as the F1 score or F-measure. A single score is created by integrating recall and precision. Identifying all instances of positivity in the data is assessed by precision, whereas recall quantifies the accuracy of making positive predictions. When precision and recall must be balanced in machine learning, the F-score is frequently utilized, and it is especially helpful when the positive class is uncommon.

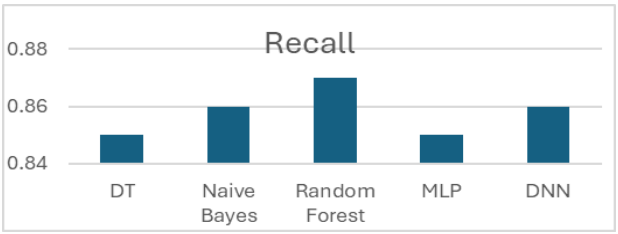

Fig 5: Recall for Purposed Architecture

The percentage of valid predictions in a dataset divided by the total number of predictions made is known as the classification accuracy. A predictive model's performance can be better understood by looking at the confusion matrix, which also reveals which classes are properly and mistakenly predicted as well as the kind of errors that are being made. The true positives and false negatives phrases are used to define the precision and recall metrics in terms of the cells in the confusion matrix. So, utilizing these performance metrics, we simply investigate the assessment of our intended model. We used optimization techniques on the text dataset because it contained a lot of features. These features were then minified, and we chose the features that would improve the model's accuracy. To add to that, we also used the VGG16 pretrained model, where transfer learning was used. I couldn't locate any papers where someone had employed an optimization algorithm to identify hate speech. We saw a significant boost in accuracy—a 20% increase.

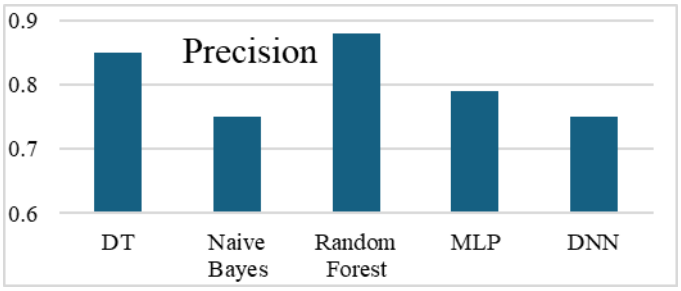

Figure 6: Precision for purposed Architecture

Next here we are showing the result we applied Deep learning model on same dataset with same optimization algorithm in table 1, accuracy is 85% and precision is 92 % and recall and F-measure both are 91% using CNN model. Out of three models we observed the best result using CNN and this performance is increased due to this optimization algorithm. In figure 7 and 8 showing ROC curve

TABLE 1a: Model Performance

	Accuracy %	Precision %	Recall %	F-measure
CNN	85	92	91	91
Dense Net	79	84	91	87
RestNet	81	86	90	88

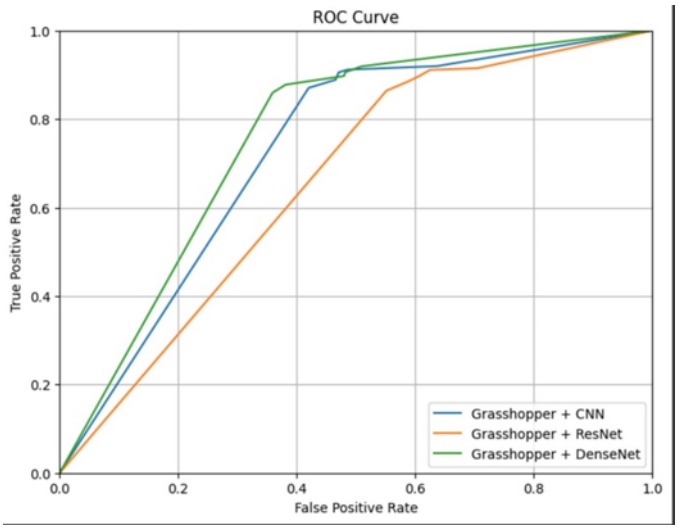

Fig 8: ROC curve

Here, we compare the effectiveness of our intended model to an existing model on the hateful meme dataset. Because the features inside the master features vector are represented as continuous data points, which make it challenging to discover the optimum threshold values needed to build a decision tree, the DT performed less well at predicting hate speech. Lack of training data is the cause

of the MLP classifier's subpar performance. When we combined the Random Forest method with the Grasshopper Optimization technique, our maximum accuracy rate was 87%, On a multimodal dataset, the outcomes following the application of GOA were assessed, and a considerable improvement for the ML models of around 19% was noted. The best accuracy we got using CNN deep learning model i.e.85%

TABLE 2: Performance comparison

Architecture	**Accuracy**
Visual Bert Coco	61%
Image captioning+visual Bert	68%
Text sentiment+visual sentiment +visual BERT	66%
VGG16+GOA(Purposed Architecture using ML model)	87%
GOA (Purposed Architecture using DL model)	85%

7. Future Work

The results of any experiment are significantly influenced by the processing of text data using feature selection and feature extraction. In this study, we demonstrated how to use GOA for the best feature selection on datasets that included both textual and visual data. Five separate machine learning models perform much better thanks to the methodology, proving it is superior to alternative methods. After evaluating the dataset and utilizing various approaches, we have identified a few topics that require more investigation in this area. As other photo captioning models create better captions, the model's ability to identify innocent text confounders will also advance, increasing classification accuracy. Fusion plays a significant part in this endeavour; hence we intend to investigate other concatenation methods using attention mechanisms, transformers, etc. As this study conducts the experiment on the 10K amount of dataset due to the restricted computer resources, the results can be improved by adding additional data. Additionally, our study can be expanded by examining related additional extensions methodologies.

8. Conclusion

The adversarial cases in the dataset for the Hateful Memes Challenge can be found using our method. Although both Image Captioning using pretrained model and optimization using Grasshopper algorithm show a positive advancement over the baseline models revealed by the Facebook Challenge, the greatest outcomes are produced by combining feature selection using optimization algorithm and Image captioning.

9. References and Footnotes

Acknowledgements

There was no specific grant from a government agency, a business, or a nonprofit group for this research. There are no studies by any of the authors in this article that used humans or animals as subjects

Conflicts of interest

The authors affirm that they do not have any competing interest's interest.

References.

[1] Gunasekara and I. Nejadgholi, "A Review of Standard Text Classification Practices for Multi-label Toxicity Identification of Online Content," 2nd Work. Abus. Lang. Online - Proc. Work. co-located with EMNLP 2018, pp. 21–25, 2018, doi: 10.18653/v1/w18-5103.

[2] Yadav and Dinesh Kumar Vishwakarma, "MRT-Net: Auto-adaptive weighting of manipulation residuals and texture clues for face manipulation detection," Expert Syst. with Appl., vol. 232, 2023, doi: https://doi.org/10.1016/j.eswa.2023.120898.

[3] S. Gite et al., "Textual Feature Extraction Using Ant Colony Optimization for Hate Speech Classification," Big Data Cogn. Comput., vol. 7, no. 1, 2023, doi: 10.3390/bdcc7010045.

[4] F. Yang et al., "Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification," no. 2017, pp. 11–18, 2019, doi: 10.18653/v1/w19-3502.

[5] D. Kiela et al., "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes," arXiv:2005.04790v3 [cs.AI], pp. 1–17, 2021, [Online]. Available: http://arxiv.org/abs/2005.04790.

[6] Chhabra and D. K. Vishwakarma, "A literature survey on multimodal and multilingual automatic hate speech identification," Multimed. Syst., no. 0123456789, 2023, doi: 10.1007/s00530-023-01051-8.

[7] M. Ali, F. A. Ghaleb, M. S. Mohammed, F. J. Alsolami, and A. I. Khan, "Web-Informed-Augmented Fake News Detection Model Using Stacked Layers of Convolutional Neural Network and Deep Autoencoder," Mathematics, vol. 11, no. 9, 2023, doi: 10.3390/math11091992.

[8] H. Aka Uymaz and S. Kumova Metin, "Vector based sentiment and emotion analysis from text: A survey," Eng. Appl. Artif. Intell., vol. 113, no. May, p. 104922, 2022, doi: 10.1016/j.engappai.2022.104922.

[9] S. Gandhi et al., “Scalable detection of offensive and non-compliant content / logo in product images,” Proc. - 2020 IEEE Winter Conf. Appl. Comput. Vision, WACV 2020, pp. 2236–2245, 2020, doi: 10.1109/WACV45572.2020.9093454.

[10] M. Zampieri et al., “SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020),” Proc. Int. Work. Semant. Eval., no. OffensEval, 2020.

[11] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, “Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines,” IEEE Intell. Syst., vol. 33, no. 6, pp. 17–25, 2018, doi: 10.1109/MIS.2018.2882362.

[12] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, “Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis,” 2016 IEEE 16th Int. Conf. Data Min., pp. 439–448, 2017, doi: 10.1109/icdm.2016.0055.

[13] S. Poria, E. Cambria, N. Howard, G. Bin Huang, and A. Hussain, “Fusing audio, visual and textual clues for sentiment analysis from multimodal content,” Neurocomputing, vol. 174, pp. 50–59, 2016, doi: 10.1016/j.neucom.2015.01.095.

[14] E. T. Niu, S. Zhu, L. Pang, “Sentiment analysis on multi-view social data,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), p. 9517, 2016, doi: http://dx.doi.org/10.1007/978-3-319-27674-8_2.

[15] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, “Image–text sentiment analysis via deep multimodal attentive fusion,” Knowledge-Based Syst., vol. 167, pp. 26–37, 2019, doi: 10.1016/j.knosys.2019.01.019.

[16] H. Ma, J. Wang, L. Qian, and H. Lin, “HAN-ReGRU: hierarchical attention network with residual gated recurrent unit for emotion recognition in conversation,” Neural Comput. Appl., vol. 33, no. 7, pp. 2685–2703, 2021, doi: 10.1007/s00521-020-05063-7.

[17] S. Poria, E. Cambria, and A. Gelbukh, “Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis,” Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process., no. September, pp. 2539–2544, 2015, doi: 10.18653/v1/d15-1303.

[18] Chhabra and D. Kumar, “A Truncated SVD Framework for Online Hate Speech Detection on the ETHOS Dataset,” pp. 1–4, 2023.

[19] Chhabra and D. K. Vishwakarma, “Fuzzy and Machine learning Classifiers for Hate Content Detection : A Comparative Analysis,” pp. 22–25, 2022.

[20] W. A. Arentz and B. Olstad, “Classifying offensive sites based on image content,” Comput. Vis. Image Underst., vol. 94, no. 1–3, pp. 295–310, 2004, doi: 10.1016/j.cviu.2003.10.007.

[21] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, “A survey of skin-color modeling and detection methods,” Pattern Recognit., vol. 40, no. 3, pp. 1106–1122, 2007, doi: 10.1016/j.patcog.2006.06.010.

[22] Tian, X. Zhang, W. Wei, and X. Gao, “Color pornographic image detection based on color-saliency preserved mixture deformable part model,” Multimed. Tools Appl., vol. 77, no. 6, pp. 6629–6645, 2018, doi: 10.1007/s11042-017-4576-2.

[23] Chhabra and D. K. Vishwakarma, “Multimodal hate speech detection via multi-scale visual kernels and knowledge distillation architecture,” Eng. Appl. Artif. Intell., vol. 126, no. PB, p. 106991, 2023, doi: 10.1016/j.engappai.2023.106991.

[24] Kiela et al., “The hateful memes challenge: Detecting hate speech in multimodal memes,” Adv. Neural Inf. Process. Syst., vol. 2020-Decem, no. NeurIPS, pp. 1–14, 2020.

[25] R. Gomez et al., “Exploring Hate Speech Detection in Multimodal Publications,” https://help.twitter.com/en/ rules-and-policies/hateful-conduct-policy.2020

[26] S Suryawanshi et al. "Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text” Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 32–41 Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

[27] American College of Physicians: American college of physicians says hate crimes are public health issue. Last Modified August 14. https://www.acponline.org/acp-newsroom/american-college-of-physicians-says-hate-crimes-are-public-health-issue (2017). Accessed 19 July 2022

[28] Annu Dhankhar et al. A Survey on Multimodal hate speech Detection published in 2023 IEEE HTC conference

[29] Annu Dhankhar et al. Feature extraction from text using grasshopper optimization algorithm for identifying hate speech published in 2023 ICAICCIT