Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

A Review: Context-Aware Hate Speech Detection using Feature Fusion

Marwa Adnan Ahmad[1] , Ali Obied [2], Ali Hamzah Najim[3]

[1]Department Computer Science , College of Computer Science and Information Technology, AL-Qadisiyah University, Al-Diwaniyah, Iraq, master.student242@qu.edu.iq

[2]Department Computer Science , College of Computer Science and Information Technology, AL-Qadisiyah University, Al-Diwaniyah, Iraq, ali.obied@qu.edu.iq

[3]Department of Computer Technical Engineering, Imam Al-Kadhum College (IKC),Al-Diwaniyah, Iraq, alihamza@iku.edu.iq

ARTICLE INFO

Article history:
Received: 21 /02/2026
Rrevised form: 13 /03/2026
Accepted : 15 /03/2026
Available online: 30 /06/2026

Keywords:

hate speech, offensive language, machine learning, deep learning, transformers, Attention Mechanism, Transfer learning

ABSTRACT

The way people communicate has changed as a result of the proliferation of digital platforms and smart devices, which provide individuals with a means to interact, voice their opinions, and share their ideas. This trend has led to an increase in posts and comments. Some individuals have resorted to posting abusive content, including hate speech, which aims to insult or incite violence and crimes against a group or individual based on a set of protected characteristics, such as religion, race, disability, religious orientation, gender, and others. The number of hate speech posts has increased recently, creating a hostile and unsafe online environment for vulnerable groups, in addition to its psychological effects on individuals within these groups. Social media platforms have resorted to using automated methods and techniques to detect and mitigate this phenomenon. This review aims to offer an in-depth understanding of previous studies in the field of hate speech detection, describing the methodologies, frameworks, and models developed using a range of standard databases while highlighting their strengths and limitations. The paper also seeks to learn how feature fusion can be used to promote the text representation with the addition of information that would advance the model to make a distinction between the categories whether in multi- or binary classification and particularly in the complex environment. What is more, the paper explains the role of context-aware models which consider texts as a unified grammatical unit in order to enhance semantic interpretation since this strategy is essential in hate speech since meaning can be context-dependent.

MSC..

https://doi.org/10.29304/jqcsm.2026.18.22673

∗Corresponding author : Marwa Adnan Ahmad

Email addresses: master.student242@qu.edu.iq

Communicated by ‘sub etitor’

1. Introduction

With the significant increase in electronic platforms that provide freedom of expression, the possibility of anonymity, and the rapid spread of information, this has encouraged some people to misuse these platforms and spread offensive content [1]. This behavior is a growing concern for social media platforms, online communities, and government organizations [2], as it has negative effects on individuals and communities, contributing to the spread of anxiety and depression and limiting community participation [3]. Some studies suggest that social media may be acting as a medium for transferring hate speech from online to violent crimes in the real world [4].

There is an urgent need to develop effective frameworks for detecting and mitigating harmful content. Since manual solutions and fixed rules are insufficient, recent studies have turned to relying on automatic detection in this field to mitigate its harmful effects. Some studies have used traditional learning methods, but they suffer from limitations in their ability to understand context because they treat the text as a bag of words and ignore word order and grammatical relationships between words within the sentence. This reduces their ability to detect implicit hatred and sarcastic comments. As a result, other studies have employed deep learning techniques, which can handle large datasets and effectively address contextual issues. However, there are limitations regarding its reliance on fixed representations and its need for large and diverse amounts of data, as its performance deteriorates in the event of data deficiencies or distortions. In light of this, researchers have resorted to using large pre-trained models, which have proven to be highly effective. Despite this, these models fail to capture subtle and changing linguistic signals and linguistic shifts [5].

Current models still suffer from a lack of balance between classification accuracy and interpretability, given that users have developed new methods, such as coded or implicit language or language disguised with irony and humor [6]. Therefore, further research is needed to build more sophisticated models to address these limitations, given the seriousness of this issue in the social fabric, as it encourages marginalization, deepens divisions, and leads to violence and the spread of misinformation [7].

The research paper is arranged in the following way: Section Two summarizes the similar and relevant studies on automatic detection of offensive language and hate speech. Section Three shows the key distinction between hate speech and other phenomena like offensive language and cyberbullying. Section Four describes the meaning of feature engineering and the relevance of feature engineering in converting raw data into forms that can be easily manipulated by models. Section five describes the process of Attention and its significance. Section Six explains the role of transfer learning strategies in modern learning and the adaptation of models from one task to another. Section Seven is the fusion of features and explains the difference between the early and late fusion and the type of models applied in this kind of research. Section Nine is the general trends and constraints in this field and the gaps in the research and finally Section Ten is the conclusion of the research paper highlighting the challenges.

2.Related Work

Methodologies for detecting offensive language and hate speech have gone through several evolutionary stages, from traditional methods to modern transformer models. The research strategy relied on selecting the most recent research papers in this field that developed their models based on standardized datasets.

Hasan et al [8] implemented three machine learning models using preprocessing on an OLID dataset to detect abusive language, classify the type of abuse, and identify the target of abuse. The dataset was divided into 70 training and 30 test cases. The preprocessing steps included removing missing values, common stop words, tokenization, lowercasing, and lemmatization. This helped remove noise in the data and improve performance. The study indicated that using term frequency-inverse document frequency (TF-IDF) increases feature importance by highlighting relevant and distinct words, which is particularly effective for small datasets. Support vector machines outperformed Random Forest artificial neural networks on all three tasks, achieving F1 scores (57%, 88%, and 68%), precision scores (76%, 87%, and 67%), recall rates (45%, 88%, and 68%), and accuracy scores (77%, 88%, and 68%). The model's bias and limited ability to generalize.

To detect abusive language on Twitter, Zhang et al [9] implemented an ensemble model consisting of a bidirectional LSTM + a bidirectional GRU, a convolutional neural network, and a bidirectional LSTM with attention. Adding a drop layer after inclusion to randomly drop words is useful for overfitting. They used pre-processing and also manually checked the OLID database and found up to 4% misclassification errors in tweets that were considered abusive but were misclassified as non-abusive. They corrected the errors. The cluster model achieved an overall F1 score of

0.8066 and an accuracy of 0.8407. For the other task of determining whether the abuse was targeted, they used a set of inference rules derived from training data and manual observations. The use of a weak and non-scalable model to identify the target of abuse, as well as the impact of data imbalance on performance, which made the training process difficult, and on the other hand, the inherent bias due to the method of collecting data using keywords.

For the three subtasks of the semEvel 2019 competition (detecting offensive language, classifying the type of abuse (targeted vs. non-targeted), and targeting an individual, group, or other),Sridharan et al. [10] proposed using a temporal convolutional neural network to handle attention. For the first subtask, they implemented an attention-based convolutional neural network (ATT-TCN). Instead of a traditional dropout layer, they applied a simple attention mechanism at the output of the embedding layer and at the output of the temporal convolutional neural network before the classification layer. For the second and third subtasks, they introduced a self-attention-based temporal convolutional neural network (SAE-ATT-TCN). They implemented a TCN model with an attention layer at the output of each Dilated Conv1D, an input attention layer to characterize multiple token locations, and semantic feature extraction. The dataset used in the study is OLID. Additionally, they used additional data from Kaggle to refine the first subtask and TRAC to refine the second subtask. Using a layered approach leads to improved model generalization. Models suffer from poor performance in classifying imbalanced classes.

Wasi [11]proposed the XG-HSI model for detecting anti-Muslim hate speech. He used graphical neural networks (GNNs) with attentional mechanisms to capture hidden relationships and contextual patterns. Comments were represented as nodes, and edges were connected to the nodes based on contextual and semantic similarity, measured using cosine similarity. Textual representations (Embedding) were extracted using BERT and BiRNN models, and the GNNExplainer tool was used to interpret the model's decisions.GNNs are expensive to train, and limited generalization, since they are trained using a single dataset.

In this work, Qin et al. [12] took several measures to overcome FEW-SHOT environments, including training on the MindSpore platform, as it promotes effective implementation in resource-limited environments. They used two sets (HateXplain and HSOL). The experimental setup for this work consists of several steps: First, they performed encoding and synonym-based adversarial augmentation to increase data diversity and expand learning by replacing some words with their synonyms. Second, they used fast, learnable embeddings before concatenating the original text to provide the model with prior contextual information, which helps it better understand the task even with limited data. Third, they extracted local features using a CNN consisting of a Conv1D layer and a Max pooling layer to select the most important features and reduce dimensionality. Fourth, they captured context using BiLSTM + Attention. Finally, they performed final classification using a fully connected layer. The proposed model structure is complex, reliance on explainability is limited, and performance is relatively low in the complex implicit language.

Sarkar et al. [13] argue that fine-tuning or retraining domain-specific models before applying them to new tasks leads to robust results across various domains. They present fBERT, a BERT-based model retrained on the SOLD dataset of over 1.4 million offensive English phrases to detect offensive language; they used masked language modelling to recalibrate the model. 15% of the total tokens were masked for replacement, and the model was then trained to predict the original words. 80% of the selected tokens were replaced with a mask, and 10% with a random token. The resulting FBERT model was trained for 25 epochs using MLM with a probability of 0:15 to mask the tokens in the input randomly. After retraining, fBERT was tested on the OLID, HS&O, and HatEval datasets, where it outperformed both BERT and HateBERT, with Macro F1 scores of 0.596 on HatEval, 0.813 on OLID, and 0.878 on HS&O. The quality of learning is adversely affected with the use of a SOLID database, which includes errors because of semi-supervised mode of collection. In addition, the model is not tried at the token level and hence little knowledge of the exact context is developed.

Caselli et al. [14] used transfer learning, retraining the BERT model on a large dataset of Reddit comments extracted from communities banned for posting hate speech. The use of masking language modelling (MLM) is beneficial, as it makes the model biased toward abusive language contexts. They evaluated the performance of their proposed model (HateBERT) on three datasets (OLID, HatEval, and AbusEval), achieving superior performance compared to the original model on all three. In addition, they provided a large dataset (RAL-E) containing over 4.1 million messages. The model showed a clear bias towards some communities that contain a larger number of publications compared to other communities, and this affects the fairness of the predictions and limits their generalizability.

Alothman et al. [15] used the XLNet model, which combines the power of autoregression and autocoding to capture two-way context. They compared their proposed model with BERT on an OLID dataset, where BERT excelled at identifying the target of the offence, while XLNet excelled at detecting abusive content and classifying the type of abuse. The model's performance was weaker in identifying the target audience due to the model's inability to understand all the contextual aspects of the text. In addition, the model struggles with underrepresented categories such as the OTH category, which affected the overall performance of the category. On the other hand, there is reliance on a single database, as well as limited experimentation and comparisons.

Ashwin Singh and Rurarroop Ray [16] compared machine learning models with transformer models. They implemented three experimental approaches. The first approach used four types of features, including statistical, sentiment-based, TF-IDF, and abuse-based features. Then, they applied traditional classification algorithms such as SVM, logistic regression, and Naive Bayes. The researchers observed that the Random Forest classifier performed best, achieving Accuracy = 75.41 and F1 (macro) = 83.01. In the second experimental method, the researchers followed a different approach, which was to use sentence embeddings generated by BERT with CLS tokens as inputs representing features. With the same models mentioned previously, a data cleaning process was carried out that included removing hashtags, emojis and common stop words. The results showed that logistic regression was the best on the test set, as it achieved Accuracy = 83.05 and F1(macro) = 77.12. In the third approach, state-of-the-art transformative language models, such as BERT and its lighter version, DistilBERT, were fine-tuned on two OLID datasets and an additional dataset for abusive language and hate speech. DistilBERT outperformed, achieving an F1 (macro) = 78.80 and Accuracy = 83.25. Overall, the transformer models outperformed traditional models on the tasks of classifying abusive language and hate speech. Performance in this study was characterized by a variety of methods and features used. They also emphasized the importance of reproducibility, noting that attempts to reimplement the methods failed to yield the same results as reported in the study.

Ramakrishnan et al. [17] propose an ensemble model to identify abusive tweets (offensive vs. non offensive) and classify the type of abuse (targeted vs. nontargeted). The model consists of aggregating the results of five traditional machine learning models using voting, three of which are basic logistic regression models with L2 regularization: one using Bag of Words (1-4 grams), another using tweet polarity, word embedding, cuss word count, and cuss word position, and the third using character 4-grams and two tree-based models (Random Forest and XGBoost). Using multiple models contributed to improved accuracy, as the ensemble model leverages the best features of each individual model. They used the OffensEval2019 dataset and an additional TRAC dataset. They observed that using the additional dataset led to a drop in performance from an Accuracy of 0.80 and an F1 (macro) of 0.74 to an Accuracy of 0.74 and an F1 (macro) of 0.73, due to the use of heterogeneous data sources. . Despite the variety of features, they are superficial, so the models often misclassify non-offensive tweets simply because they contain obscene words

Swamy et al. [18] used machine learning and feature extraction to detect whether a tweet was offensive or non-offensive, as well as to classify the abuse as targeted or non-targeted. Their ensemble model consists of five classifiers, including: L1-Regularised Logistic Regression, L2-Regularised Logistic Regression, Linear SVC, SGD, and PA. They used various feature extraction methods, including surface-level tokens (1–3 grams) weighted by TF-IDF; POS tags obtained through the CMU tagger2; sentiment scores assigned using a pre-trained model included in TextBlob3; and count features for URLs, mentions, hashtags, punctuation marks, words, syllables, and sentences. They showed that both word n-grams and character n-grams were more predictive. Their results showed that the Ensemble model achieved F1 scores of 0.7434 and 0.7078 on the main tasks, respectively. As for setting the target, whether it is an individual, a group, or another, use LSTM with Embedded Glove5. Difficulty in classifying non-offensive tweets that contain profanity and offensive tweets that do not contain it, as well as tweets with a political motive that are incorrectly classified as offensive. On the other hand, the impact of most of the features is limited sometimes reduces performance, as well as low performance in the task of identifying the target, whether an individual, a group, or another.

In this study, Adewumi et al. [19] used a range of models to detect hate speech, including bidirectional long-term memory networks (BiLSTM), convolutional neural networks (CNNs), ROBERT, and a text-to-text converter (T5). They also employed data preprocessing, including IP address and URL removal and converting all characters to lowercase. Their proposed model combines the outputs of three models, RoBERT, T5 Small, and T5, using majority voting. Their efficient and simple data augmentation method improved the stability of their model. They also revealed the weakness of data annotations in the HASOC 2021 dataset and applied interpretive artificial intelligence (XAI) techniques. They used SHAP and Integrated Gradients (IG) to interpret the model's decisions. Short comment

of Twitter used in the model restricts its ability to generalize to longer texts and hence performs poorly. It is also very time consuming in terms of computing power.

Clarke et al. [20]combined transformer-based representations and rule-based reasoning through a dual encoding architecture. They used a text encoder and a rule encoder to learn both textual and logical representations from the data. This is useful for making the model generalizable and interpretable. They implemented this approach using the BERT model, linking rules to examples through Contrastive learning and cosine similarity. They evaluated the performance of their model, called RBE, on three databases, and it outperformed traditional models trained on confused data and rules-based models alone. Relying on expert supervision increases human costs and the need for high computing resources; on the other hand, the quality of performance is affected by the quality of the rules.

In this study to detect offensive language, Liu et al. [21] used several techniques to implement a three-stage framework for detecting offensive language, including data augmentation, interaction fusion mechanism, and semantic understanding. In the first stage, they applied preprocessing, which included data cleaning, removing URLs and hashtags (#), and converting emojis to text. Following this, they applied data enhancement based on reverse translation, a method useful for multiplying data and improving its diversity. They divided semantic understanding into two units: the first relied on Deep Semantic Modules (DSM) using the BERT model, a multi-headed attention mechanism, and a feedforward network. This step was useful for creating high-level vector representations of the inputs. The second unit was a Character Capture Network (CCN), using n-gram features at the character level and a CNN neural network to extract local features. They then merged the outputs of BERT and CNN, using interaction fusion with pyramid layers to enhance feature integration and capture the interaction between deep and character-based features. The difficulty in detecting implicit or indirect offensive language, especially in political texts that do not contain explicitly negative words, is a concern, as some sentences have been classified as offensive simply because they contain a vulgar word.

Putra et al. [22] combine advanced convolutional neural networks (Advanced CNNs) and contextual BERT representations for the detection of hate speech and abusive language. The BERT model consists of 12 layers, including a hidden layer of 768 and another of 1024, with 340 million parameters. BERT requires three types of input: token embedding, position embedding, and segment embedding. These representations are combined to form the inputs for BERT. Advanced CNN uses two to four kernels to concatenate the BERT content. After convolution, each kernel is connected to a Maxpooling function, followed by a concatenate function to combine the outputs. Each node of the output layer represents the probability of a sentence being classified as hate speech. This study used two datasets to evaluate the performance of the proposed model: the Davidson dataset, which includes the classes (hate speech, offensive language, neither1), and the TRAC-1 dataset, which includes the categories (affectively aggressive, covertly aggressive, and non-aggressive). Preprocessing, including token removal, connective deletion, case normalization, punctuation removal, and other cleaning steps were performed before the data was fed to BERT to avoid compromising the quality of the contextual representation. Their model outperformed Fast Text+CNN and BERT alone, achieving an F1 score of 73% on Davidson and 56% on TRAC-1. In addition to the above, the proposed model is highly flexible in dealing with class balance. The drawbacks of this work include the use of the traditional CNN algorithm without experimenting with deeper neural network architectures, limited diversity of input features and representations, low performance, especially on the TRAC-1 dataset, and the lack of interpretability or transparency.The drawbacks of this work include the use of the traditional CNN algorithm without experimenting with deeper neural network architectures, limited diversity of input features and representations, low performance, especially on the TRAC-1 dataset, and the lack of interpretability or transparency

Zhu et al. [23] used a deep prompt-based multi-task network (DPMN) to detect abusive language. DPMN implements Deep and Light Prompt Tuning for PLM models. The researchers examined the impact of prompt lengths, tuning strategies, and initialization methods on the identification of offensive language. Their Task Head, which combines a bi-LSTM and a feedforward network, classifies short texts. DPMN's multi-task learning improves detection metrics by sharing information between tasks. The model, tested against eight leading methods on three public datasets (OLID, SOLID, and AbuseAnalyzer), showed superior performance in abusive language detection compared to current state-of-the-art approaches. The model depends heavily on the quality of the PLM used. If the linguistic knowledge in the pre-model is weak or poor, performance decreases, and interpretability is not utilized.

Table 1: A comparison summarizing the performance of different models across Benchmark datasets discussed in Related work using performance metrics including (accuracy, recall, precision, and F1-score).

Ref	years	Models	Performance metrics				Dataset
			Accuracy	precision	Recall	F1-score	
[8]	2024	SVM+TF-IDF	A=0.77 B=0.88 C=0.68	A=0.76 B=0.87 C=0.67	A=0.45 B=0.88 C=0.68	A=0.57 B=0.88 C=0.68	OLID
[9]	2019	Ensemble (CNN+BiLSTM-Attention+BiLSTM-BiGRU)	0.84	-	-	0.80	OLID
[10]	2019	ATT-TCN (Attention-based Temporal CNN	A=0.65	-	-	A=0.46	OLID
		SAE-ATT-TCN (self AttentiveEmbedding+Attention TCNsubset)	B=0.75 C=0.61	-	-	B=0.61 C=0.51	
[11]	2024	XG-HIS-BiRNN	0.74	-	-	0.73	Hatexplain (Muslim focused subset)
		XG-HIS-BERT	0.75	-	-	0.74	
[12]	2025	MS-FSLHate	-	65.45	65.78	65.56	Hatexplain
			-	85.16	85.21	85.16	HsoL
[13]	2021	FBERT	-	-	-	0.596	HateEval
			-	-	-	0.813	OLID
			-	-	-	0.878	Davidson
[14]	2021	HateBERT	-	0 .836	0.404	0.809	OLID
			-	0 .553	0.696	0.765	AbusEval
			-	0.565	0.567	0.516	HatEval
[15]	2025	XLNet	A=0.81 B=0.83 C=0.56	A=0.62 B=0.97 C=0.44	A=0.78 B=0.84 C=0.41	A=0.69 B=0.90 C=0.36	0LID
[17]	2019	Ensemble ML (logistic BOW, semantic features, char n-gram, Random Forest, XGBoost	0.80	-	-	0.74	OLID
			0.74	-	-	0.73	TRAC-1
[18]	2019	Ensemble (L1Log Reg, L2 Reg, SVC, SGD, PA)	A=0.80 B=0.88	-	-	A=0.74 B=0.70	OLID
		LSTM+GloVe embedding5	C=0.63	-	-	C=0.50	
[19]	2023	T5-Base+Augmented data	-	-	-	A=82.54 B=62.71	Hasoc2021

[20]	2023	RBE (BERT-based Dual Encoder +contrastive Learning)	0.79	0.79	0,89	0.83	Hatexplain
		RBE (BERT-based Dual Encoder)	0.99	0.58	0.62	0.60	jigsaw (identity Hate)
			0.90	0.48	0.46	0.47	CAD(contextual Abuse dataset)
[21]	2022	BERT+CNN+interaCtivefusion+Back-Translation	0.96	-	-	0.97	Davidson
			0.89	-	-	0.86	OLID
			-	85.16	85.21	85.16	HsoL
			-	0.58	0.55	0.56	TRAC-1
[22]	2023	BERT+Advanced CNN	-	0.72	0.75	0.73	Davidson
			-	0.58	0.55	0.56	TRAC-1
[23]	2024	DPMN	-	-	-	0.83	OLID
			-	-	-	0.92	SOLID
			-	-	-	0.81	Abuseanalyae

3. Hate speech and other related concepts

Determining whether a piece of text constitutes hate speech is difficult, as the phenomenon is complex even for humans, given its overlap with other related concepts. It is important to understand the difference between hate speech and related terms such as cyberbullying and abusive language. referred to hate speech is as any form of targeting an individual's enjoyment of general civil rights or civil liberties, rather than focusing solely on the offensive content of the text[24]. Similarly, Description of hate speech as an expression that involves prejudice or incitement to violence against a group based on characteristics protected by international law, such as race, religion, disability, region, sexual orientation, or age, and is intended to insult, humiliate, offend, or threaten members of these groups[25]. On the other hand, Cyberbullying refers to repeated aggressive behavior targeting a specific individual with the intent to insult, harass, or threaten, regardless of whether the individual belongs to a specific group [26]. Offensive language is vulgar, obscene, or hurtful expressions used in the context of speech, without incitement to violence or discrimination against a protected group [27].

4.*Feature Engineering*

Feature extraction is a crucial task for improving the performance of machine learning models. It is a pivotal step in transforming raw data into meaningful numerical representations for these models [28]. This enables the models to treat text as learnable input by converting the data into numerical vectors with real-valued properties [29]. Furthermore, each model architecture exhibits a different response depending on the type of geometric feature. Feature types include: Character n-grams, Word n-grams, Cuss-word Dictionary and Profanity Checker, GloVe Embedding, Word2vec, FastText, ELMO, Part of Speech, and other features that enhance performance in text classification and predictive tasks [30].

5.Attention Mechanism

The attention mechanism refers to assigning a weight to each value based on the degree of correspondence or similarity between the query and its corresponding key. Correspondence is calculated using a correspondence function, and the weights represent the importance of each value to the current query. The main idea is that attention makes the model more intelligent, and not every instance of attention is treated with equal importance. CNNs and RNNs have been replaced by Multiheaded Attention Mechanisms for processing linguistic sequences in modern models represented by transformers, where the attention mechanism allows the extraction of relationships between any two words in a sentence without the need for chronological order [31].

6.Transfer Learning

It is a branch of machine learning that aims to reuse existing models to solve new problems instead of starting from scratch. This type of methodology is beneficial for saving resources and time on model training. Previously, in the traditional approach to machine learning, each task was processed in isolation, and the model was trained from scratch for each new problem. In contrast, the modern approach of transfer learning relies on data or previous models from similar tasks to improve performance. Transfer learning has many applications, including real-world simulations, gaming, image classification, and sentiment classification. **Figure.1 illustrates the modern transfer learning**, TL describing it as an optimization tool that contributes to enhancing a model's performance for another task by leveraging its
313 ior knowledge base. For example, retraining the BERT model for an abusive language detection task involves using a SOLID database [32].

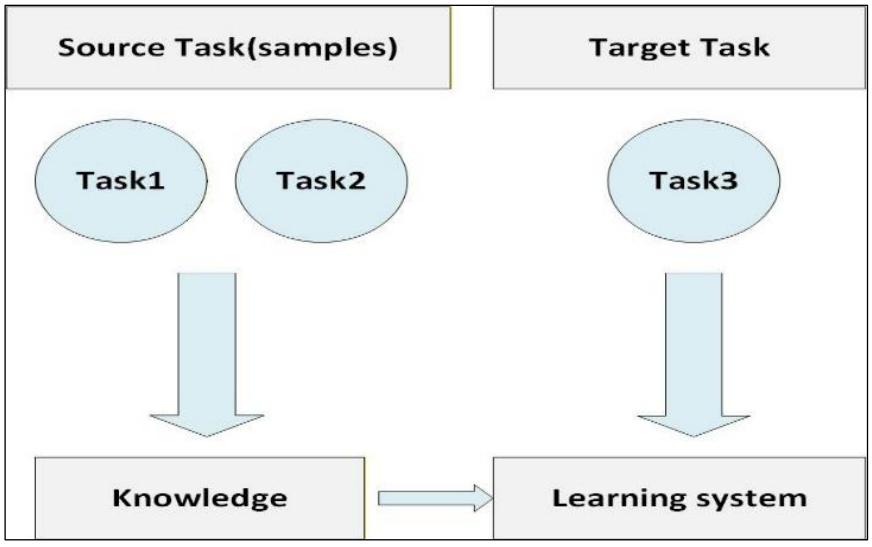

Figure1: The modern transfers learning

7.Feature Fusion

Fusion refers to the aggregation of features extracted from multiple layers or channels into a single, informative representation that reflects the complex interaction between these features. This strategy is used in machine learning and aims to improve performance and prediction while reducing redundancy [33]. There are two common methods for data fusion: early fusion and late fusion. Early fusion involves extracting features from raw data and combining them into a common representation before feeding them to a machine learning model for training. The data must be correctly aligned in this representation; this type is considered the most efficient and is called 'feature-level fusion' [34]. **Figure.2 illustrates the mechanism of early fusion of features before they enter as a vector into the model.**

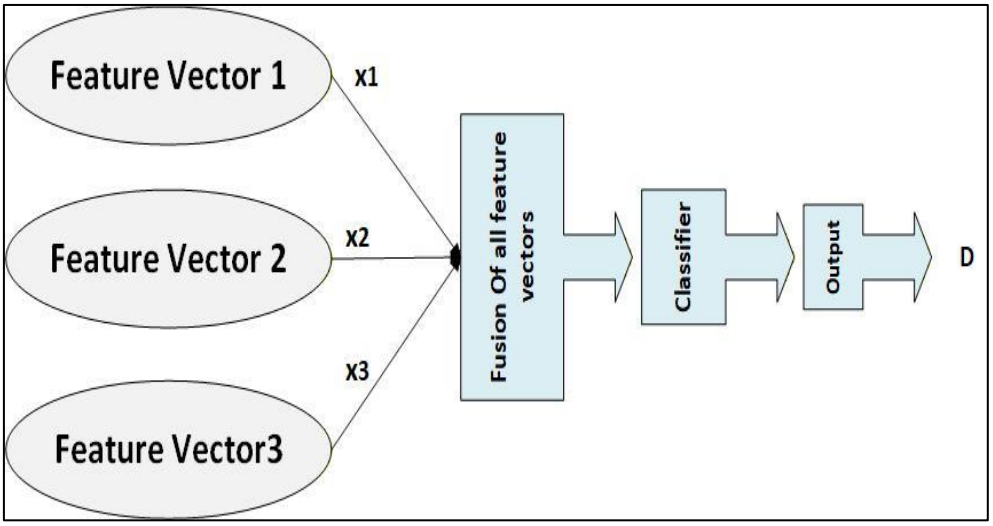

Figure 2: Early Fusion Mechanism

Late fusion involves combining results using averaging or summation methods after training individual models on the data and generating predictions. This type is the most common, but it is characterized by limited performance due to its failure to exploit relationships between single-mode data [35]. **Figure 3 illustrates the Late Fusion mechanism,** where extracted features are delivered to individual classifiers, and after the results appear, Decision-Level Fusion is used

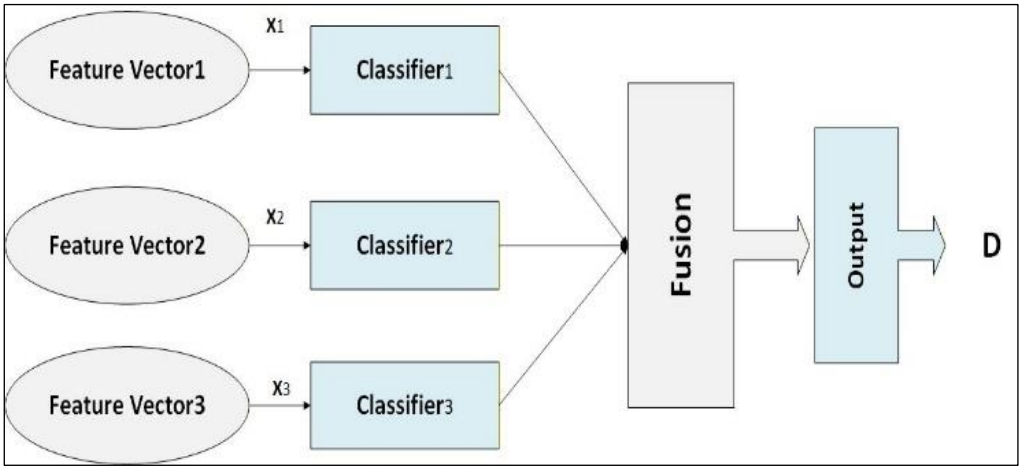

Figure3: Late Fusion Mechanism

8.Types of proposed Models in the field of Hate speech

8.1. Machine Learning-based models

Machine learning is an effective tool for the automatic detection of hate speech and abusive language. It relies on statistical methods to train algorithms to classify or predict text categories [36]. Many early studies in this field used traditional machine learning algorithms, such as Supporting Vector Machines (SVMs), Logistic Regression (LR), Random Forests (RF), Decision Trees (DT), and Naive Bayesian Algorithm (NB) (Subrata Paul and Hasan). Despite

their excellent performance, these algorithms suffer from limited ability to handle large datasets, problems in capturing precise linguistic context, and a lack of interpretability [37].

8.2. Deep Learning-based models

Deep learning techniques have demonstrated high efficiencies in natural language processing, speech processing and visual and audio data. They excel at representing complex patterns, understanding nonlinear relationships, and automatically extracting features from data without human intervention. Furthermore, as data volumes have increased and the data can be easily transferred to new tasks, many studies on detecting abusive language have utilised deep learning models such as CNN, LSTM, and RNN, which have performed well in understanding contextual and chronological sequences [38].

8.3. Transformers-based models

Transformers are among the most important recent achievements in natural language processing. They are characterized by their reliance on self-attention mechanisms and their use of parallel processing, which increases their training speed and computational efficiency. For example, the BERT model is a widely used standard in natural language processing due to its ability to understand the full context of text. Similarly, the XLNET and XLM-R models have demonstrated outstanding performance in detecting offensive language and hate speech on social media [39][40].

8.4. Ensemble model

Ensemble modelling is a proven method in machine learning and artificial intelligence, especially in cases of simulating complex functions and insufficient data problems. The basic idea is to combine the predictions of several models into a single prediction, which in turn produces a final model with high performance and few errors. There are several types of cluster models, such as bulking, boosting, stacking, voting, and averaging. Clustering methods are used in a wide range of applications, such as weather forecasting, text classification, fraud detection, and weather prediction [41].

9. Discussion

In the field of hate speech detection, Machine learning has been very instrumental in the area of hate speech detection. Conventional techniques like the Random forests, supporting vector mechanisms, and logistic regression have been found to work well on different data collections, yet they still are constrained in tricky settings such as the short and indirect texts shared by social media users. These techniques are not very effective in identifying sarcasm and other subtle contextual variations. Deep learning has come forward in order to overcome those
315 lenges and it has shown that it can deal with context, find complex patterns and implicit hate. Nevertheless, it has some difficulties, especially in terms of class balance, data quality, and computation resources. The transfer learning has also helped in enhancing detection of hate speech by using already trained models and re-training them in a specific task. The transformer-based models have attentional mechanisms that allow semantic comprehension of complex texts. That said, there are still certain issues with fine-tuning and resource consumption, and there is still no need to develop models that combine a systematic mixture of Handcrafted created linguistic features with deep contextual representations, and the consideration of bias and interpretability remains low.

10. conclusion and Future Direction

Hate speech detection has experienced an impressive growth in the field. This research paper explains the methods in this area, such as attentional mechanisms, transfer learning, and feature engineering, and how they influence the reduction of computational resources and performance improvement. It also explains the inherent distinction between hate speech and other terms and outlines models that were created and evaluated with the help of various standard databases including OLID and Hatexplain. Machine learning algorithms are good in this field, yet they have the problems with processing large volumes of data and perceiving the context. Conversely, deep learning has been found to be more effective in performance in this respect, though, it tends to reduce its performance in case of insufficiency, unbalance, and lack of diversity. Also, according to other research findings, pre-trained models including the BERT model are effective at reuse in automatic detection of hate speech and offensive language. The

primary issues related to the process of detecting hate speech and abusive language, which we have highlighted, are the disproportion between categories in databases, which causes the model to focus on the most prevalent category. This weakness has a harmful effect on its generalization and manipulation of new or rare examples. Models also find it hard to identify implicit language and coded sarcasm which people intentionally use to avoid detection systems. Moreover, lack of interpretability is also a major problem as models are more likely to act as a black box, so it is hard to comprehend the decision-making process or explain how the classification has been made. Even huge models need huge computing resources. In the future, the feature fusion and multi-level attention processes can be employed to enable a model to become more attentive to the subtle contexts and implicit patterns. Also, transfer learning as well as retraining can be used to take advantage of more generalizability of the model to other fields. To obtain a more favorable balance between accuracy and interpretability we suggest the design of internal interpretation mechanisms built into the architecture of the model instead of depending on external interpretation tools only. There is also a need to have effective methodologies of dealing with low-resource environments and data imbalances to increase model efficiency and generalizability.

Acknowledgements

I would also like to acknowledge my supervisor and professor Ali Obeid and Ali Hamza Najm who helped me and guided me in this research.

References

[1] S. Ghosh, M. Suri, P. Chiniya, U. Tyagi, S. Kumar, and D. Manocha, “CoSyn: Detecting Implicit Hate Speech in Online Conversations Using a Context Synergized Hyperbolic Network,” Oct. 2023, [Online]. Available: http://arxiv.org/abs/2303.03387

[2] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, “Predicting the Type and Target of Offensive Posts in Social Media,” Apr. 2019, [Online]. Available: http://arxiv.org/abs/1902.09666

[3] J. M. Pérez *et al.*, “Assessing the impact of contextual information in hate speech detection,” Mar. 2023, [Online]. Available: http://arxiv.org/abs/2210.00465

[4] P. Vijayaraghavan, H. Larochelle, and D. Roy, “Interpretable Multi-Modal Hate Speech Detection,” Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.01616

[5] J. Daniel and J. H. Martin, “Speech and Language Processing,” 2025.

[6] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, “HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection,” 2021. [Online]. Available: https://github.com/punyajoy/HateXplain

[7] S. Bradshaw, “Disinformation and Identity-Based Violence,” 2024.

[8] M. N. Hasan, K. S. Sakib, T. T. Preeti, J. Allohibi, A. A. Alharbi, and J. Uddin, “OLF-ML: An Offensive Language Framework for Detection, Categorization, and Offense Target Identification Using Text Processing and Machine Learning Algorithms,” *Mathematics*, vol. 12, no. 13, Jul. 2024, doi: 10.3390/math12132123.

[9] H. Zhang *et al.*, “MIDAS at SemEval-2019 Task 6: Identifying Offensive Posts and Targeted Offense from Twitter,” 2019. [Online]. Available: https://sites.google.com/view/trac1/

[10] M. Sridharan and S. T. R, “Amrita School of Engineering-CSE at SemEval-2019 Task 6: Manipulating Attention with Temporal Convolutional Neural Network for Offense Identification and Classification,” 2019.

[11] A. T. Wasi, “Explainable Identification of Hate Speech towards Islam using Graph Neural Networks,” 2024.

[12] Z. Qin, D. Wu, Y. Liu, and G. Yang, “Few-shot Hate Speech Detection Based on the MindSpore Framework,” Apr. 2025, [Online]. Available: http://arxiv.org/abs/2504.15987

[13] D. Sarkar, M. Zampieri, T. Ranasinghe, and A. Ororbia, “FBERT: A Neural Transformer for Identifying Offensive Content,” Sep. 2021, [Online]. Available: http://arxiv.org/abs/2109.05074

[14] T. Caselli, V. Basile, J. M. Mitrovi´c, and M. Granitzer, “HateBERT: Retraining BERT for Abusive Language Detection in English,” 2021. [Online]. Available: https://en.wikipedia.org/wiki/

[15] R. Alothman, H. Benhidour, and S. Kerrache, “Offensive Language Detection on Social Media Using XLNet,” Jun. 2025, [Online]. Available: http://arxiv.org/abs/2506.21795

[16] Ashwin singh and ruraroop ray, “identifying offensive content social media posts,” 2019.

[17] M. Ramakrishnan, W. Zadrozny, and N. Tabari, “UVA Wahoos at SemEval-2019 Task 6: Hate Speech Identification using Ensemble Machine Learning,” 2019. [Online]. Available: https://news.itu.int/

[18] S. D. Swamy, A. Jamatia, B. Gambäck, and A. Das, “NIT Agartala NLP Team at SemEval-2019 Task 6: An Ensemble Approach to Identifying and Categorizing Offensive Language in Twitter Social Media Corpora.” [Online]. Available: www.cs.cmu.edu/

[19] T. Adewumi, S. S. Sabry, N. Abid, F. Liwicki, and M. Liwicki, “T5 for Hate Speech, Augmented Data, and Ensemble,” *Sci*, vol. 5, no. 4, Dec. 2023, doi: 10.3390/sci5040037.

[20] C. Clarke˚: *et al.*, “Rule By Example: Harnessing Logical Rules for Explainable Hate Speech Detection,” Long Papers, 2023. [Online]. Available: https://github.com/ChrisIsKing/

[21] J. Liu, Y. Yang, X. Fan, G. Ren, L. Yang, and Q. Ning, “Offensive-Language Detection on Multi-Semantic Fusion Based on Data Augmentation,” *Applied System Innovation*, vol. 5, no. 1, Feb. 2022, doi: 10.3390/asi5010009.

[22] C. D. Putra and H. C. Wang, “Advanced BERT-CNN for Hate Speech Detection,” in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 239–246. doi: 10.1016/j.procs.2024.02.170.

[23] J. Zhu *et al.*, “Deep Prompt Multi-task Network for Abuse Language Detection,” Jun. 2024, [Online]. Available: http://arxiv.org/abs/2403.05268

[24] NockLEBy, “Encyclopedia of the American Constitution,” 2000.

[25] A. H. Zahid, M. K. Roy, and S. Das, “Evaluation of Hate Speech Detection Using Large Language Models and Geographical Contextualization,” Feb. 2025, [Online]. Available: http://arxiv.org/abs/2502.19612

[26] Cohen Raphael, “Cyberbullying and Hate Speech,” 2022.
[27] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated Hate Speech Detection and the Problem of Offensive Language,” Mar. 2017, [Online]. Available: http://arxiv.org/abs/1703.04009
[28] E. Omran, E. Al Tararwah, and J. Al Qundus, “A comparative analysis of machine learning algorithms for hate speech detection in social media,” *Online J. Commun. Media Technol.*, vol. 13, no. 4, Oct. 2023, doi: 10.30935/ojcmt/13603.
[29] S. Abro, S. Shaikh, Z. Ali, S. Khan, G. Mujtaba, and Z. H. Khand, “Automatic hate speech detection using machine learning: A comparative study,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, pp. 484–491, 2020, doi: 10.14569/IJACSA.2020.0110861.
[30] J. Heaton, “An Empirical Analysis of Feature Engineering for Predictive Modeling,” Nov. 2020, doi: 10.1109/SECON.2016.7506650.
[31] A. Vaswani *et al.*, “Attention Is All You Need,” 2023.
[32] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, “Transfer learning: a friendly introduction,” *J. Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00652-w.
[33] L. M. Pereira, A. Salazar, and L. Vergara, “On Comparing Early and Late Fusion Methods,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, 2023, pp. 365–378. doi: 10.1007/978-3-031-43085-5_29.
[34] K. Gadzicki, “Early vs Late Fusion in Multimodal Convolutional Neural Networks.” [Online]. Available: http://www.ease-
[35] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, “Early versus Late Fusion in Semantic Video Analysis,” 2005.
[36] D. B. Parase, J. I. Nandalwar, and P. S. Sable, “Machine Learning-Its Applications, Benefits, and Threats,” *International Journal of Scientific Research in Engineering and Management*, 2023, doi: 10.55041/IJSREM18228.
[37] S. Paul, “Context-Aware Hate Speech Detection: A Comparative Study of Machine Learning Models.” [Online]. Available: https://ijcnis.org/
[38] “Natural_Language_Processing_with_Transformers_by_Lewis_Tunstall”.
[39] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2019. [Online]. Available: https://github.com/tensorflow/tensor2tensor
[40] A. Shahriar, D. Pandit, and M. S. Rahman, “XLNet-CNN: Combining Global Context Understanding of XLNet with Local Context Capture through Convolution for Improved Multi-Label Text Classification,” in *Proceedings of the 2024 11th International Conference on Networking, Systems and Security, NSysS 2024*, Association for Computing Machinery, Inc, Jan. 2025, pp. 24–31. doi: 10.1145/3704522.3704540.
[41] T. G. Dietterich, “Ensemble Methods in Machine Learning,” 2000. [Online]. Available: http://www.cs.orst.edu/~tgd