applied sciences

MDPI

Article

Detection of Expressions of Violence Targeting Health Workers with Natural Language Processing Techniques

Merve Varol Arısoy [1,*], Mehmet Ali Yalçınkaya [2], Remzi Gürfidan [3] and Ayhan Arısoy [1]

[1] Information Systems Engineering Department, Bucak Faculty of Computer and Information, Burdur Mehmet Akif Ersoy University, 15300 Burdur, Türkiye; aarisoy@mehmetakif.edu.tr
[2] Computer Engineering Department, Faculty of Engineering and Architecture, Kırşehir Ahi Evran University, 40100 Kırşehir, Türkiye; mehmetyalcinkaya@ahievran.edu.tr
[3] Department of Computer Technologies, Yalvaç Technical Sciences Vocational School, Isparta University of Applied Science, 32200 Isparta, Türkiye; remzigurfidan@isparta.edu.tr
* Correspondence: mvarisoy@mehmetakif.edu.tr; Tel.: +90-505-395-36-87

Abstract: The aim of this study is to detect expressions of violence against healthcare workers using natural language processing techniques. Experiments on various NLP models have shown that violent expressions can be successfully classified using textual data. The RAG-ECE model performed the best in this study with a 97.97% accuracy rate and a 97.67% F1 score. The model provided a strong balancing performance in the "no violence" class with 97.71% precision and 97.67% recall rates. In the "violence present" class, it reached 97.67% accuracy and was evaluated as a reliable classifier with both low false positive (3.92%) and low false negative (2.78%) rates. In addition to RAG-ECE, the GPT model provided a milder alternative with 96.19% accuracy and a 96.26% F1 score. The study also compared the performances of other models, such as GPT, BERT, SVM, and NB, and stated that they are considered suitable alternatives due to their low computational costs, especially in small- and medium-sized datasets. The findings of the study show that NLP-based systems offer an effective solution for the early detection and prevention of expressions of violence against healthcare workers.

Keywords: natural language processing; text classification; violence in health; violence detection

Academic Editor: Jianbo Gao

Received: 3 January 2025
Revised: 29 January 2025
Accepted: 5 February 2025
Published: 8 February 2025

Citation: Varol Arısoy, M.; Yalçınkaya, M.A.; Gürfidan, R.; Arısoy, A. Detection of Expressions of Violence Targeting Health Workers with Natural Language Processing Techniques. *Appl. Sci.* **2025**, *15*, 1715. https://doi.org/10.3390/app15041715

1. Introduction

Historically, doctors have been at the center of the health system, assuming such vital roles as the protection of the health of societies and combating diseases. From ancient times to the present day, doctors have enjoyed a high level of respect for their therapeutic skills as well as for their role in preserving and transmitting medical knowledge [1]. Training a doctor requires big financial investments and a long duration. Expenses incurred by the state while training a medical student are not only limited to the education costs alone, but additional expenses include clinical internships that could be more expensive, residency programs, and other specialty trainings. For example, the cost of a training a medical student for 1 year is 263,305 dollars in the USA [2]. Even more so, this investment is not only financial but also labor-consuming; in the United States, it takes a minimum of 11 years of full training for a medical student to progress from medical school to completing specialty training and beginning independent practice [3]. This process not only has a serious cost factor but also a strategic investment by states to ensure the future of health systems. For regions with a shortage of doctors, such government investments are very much crucial for improving public health and increasing access to medical services.

Nowadays, violence against healthcare professionals has become an international issue concerning doctors. Violence against doctors has resulted not only in negative consequences for the personal physical safety of physicians but also for their mental well-being. Verbal attacks are the most usual type of violence that was reported against healthcare professionals through research. Hence, doctors often face threats, insults, and swearing from the patients and their attendants. Various studies conducted all over the world indicate that over 45% of doctors are exposed to such violence and that verbal abuse is the most common type of violence [4,5]. The early detection and prevention of such aggressive language in the work environment is of great importance both for protecting the mental health of doctors and improving the quality of the health services provided. Various research has been conducted regarding the effective use of NLP technologies in the detection and analysis of social problems. In some of these studies, social problems like violence against women, violence against children, peer bullying, and sexual harassment have been detected using textual data on social media platforms. Systems developed for the detection of cyberbullying, for example, are very successful in identifying harassment content against women and children on social media [6]. In another study, methods that were used for the classification of types of online harassment, such as sexual harassment and bullying, attained strong results both in terms of accuracy and sensitivity [7]. It also underlines how NLP is effective in real-time detection, such as violence against women, and how these methods provide successful results in identifying violent and harassing behaviors using video and audio data [8]. Figure 1: NLP Healthcare Safety schema.

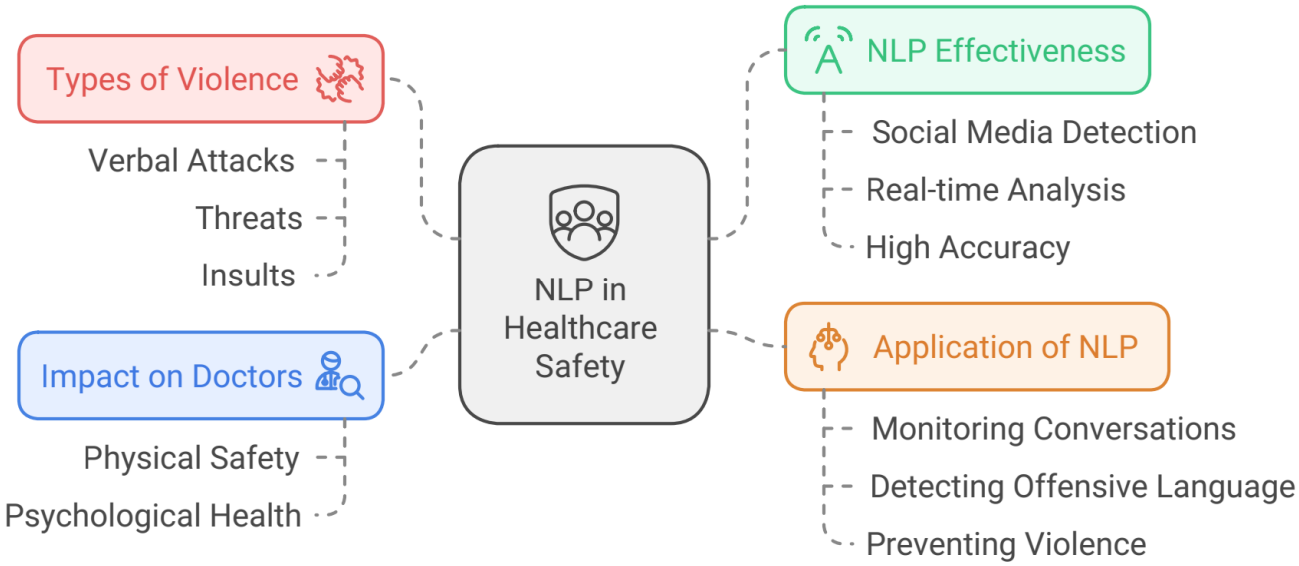

Figure 1. NLP healthcare safety schema.

In this context, one might say that NLP has great potential in the detection of social problems and, more generally, that these technologies are an important step to protect people and promote social healing. Threats, insults, or offensive language against doctors can similarly be analyzed using these approaches. The analysis of the ambient conversations may provide a clue toward the on-time detection of violence against doctors, using natural language processing models on translated text data. This aspect, to be considered, might be one of the most critical approaches toward increasing the safety of the doctor and ensuring efficiency in health services. This paper aims to enable the detection of the threatening and insulting expressions that doctors face in daily practice using natural language processing techniques. Monitoring the verbal attacks in real time and detecting such expressions

contribute to both increasing the safety of doctors and improving the functioning of the healthcare system in general.

Natural language processing technologies play a very important role in the detection of violent behavior and aggressive expressions. The detection of insults, threats, and aggressive languages in texts helps to achieve the early identification of such behavior in digital or face-to-face communications. Systems for the detection of natural language processing, especially in the health sector, social media, and public services, allow for the early identification of violence and the reduction of intervention times. The violence in language expression is, however, complex and culturally bounded, hence limiting the generalization capacity of the models. Most previous studies call for more extensive datasets, particularly in diverse languages and cultures. This clearly shows the need for higher levels of natural language processing approaches to detect and prevent violent language.

2. Related Works

There are several studies on the detection of offensive behavior using artificial intelligence in the literature. Some of them worked on aggressive behavior detection using images, and others were on insult and aggressive expression detection using different methods like natural language processing. Akti et al. (2020), in their study, created an artificial intelligence model that could detect fight scenes from surveillance cameras. The key aim of the research was to determine incidents of fights using the images obtained from cameras for surveillance as fast and precisely as possible. With this motivation, the paper proposes a method that first improves the traditional ways of fight detection with the proposal of Xception, bi-directional LSTM, and the attention layer. In this context, a new dataset was collected in the form of images captured by surveillance cameras from YouTube and shared publicly. Experiments have also been conducted using some existing datasets like Hockey Fight and Peliculas. It is observed in this study that adding an Xception-based CNN model and an attention layer to the Bi-LSTM network increased the success rate in detecting fights. This model was seen to result in higher accuracy rates as compared with existing methods. According to the outcome of the experiments, this reached accuracy rates of 98% for the Hockey Fight dataset and 100% for Peliculas. However, the success rate of this new collected data stood a bit lower on performance in comparison with other used datasets because of a complex diversified dataset. This proved that it is necessary to further develop the model for more diverse situations. The research not only gave a better and speedier solution to fight detection but also presented the scientific community with a new dataset for further research on the subject [9]. The authors conducted a further study on images by Blunsden et al. 2010. In this paper, the BEHAVE video dataset was presented for the classification of multi-person behavior. Its focus is presenting a video dataset with an innate ground truth that can then classify group behaviors. This dataset consists of approximately 90,000 frames where the people, the groups they interact with, and the behaviors exhibited by these groups have been labeled, such as walking, fighting, gathering, and so on. The dataset is specifically designed to be used in the analysis of specific group behaviors and tries to provide a basis for the performance evaluation of various artificial intelligence algorithms in this context. In this work, artificial intelligence algorithms that are widely used in classifying behaviors expressed by groups during their interaction are proposed: the hidden Markov model. This model has classified some of the classes, such as walking together, fighting, and approaching, each learned by a separate HMM. The experimental results showed that the classification accuracy reached up to 93.67% using the video information acquired in different windows (e.g., from 5 frames to 100 frames). This study not only provides a rich dataset for automatically detecting and classifying group behaviors but also highlights the importance of developing such datasets [10]. In

their study, Soliman et al. (2019) developed a deep learning-based model that can detect violence in videos. It means that the research is basically trying to create a system for the automatic recognition of violent acts in videos that overcomes traditional surveillance systems' limitations. From this motivation, an end-to-end deep neural network model is proposed that incorporates the VGG-16 model as a spatial feature extractor and the LSTM model as a temporal feature extractor. In this paper, a new benchmark dataset is introduced, namely the Real-Life Violence Situations (RLVS) dataset, to enhance the performance of the model and make it compatible with real-world scenarios. This dataset contains 2000 short video clips of people from different genders, races, and age groups, including 1000 violent videos and 1000 non-violent videos. The proposed model was tested on the benchmark datasets of violence like Hockey Fight, Movie, and Violent-Flow, and obtained 95.1%, 99%, and 90.01% accuracy rates, respectively. Besides, during tests on the RLVS dataset, it was observed that the model became successful in recognizing violent videos but achieved lower accuracy rates in non-violent videos to improve the overall performance of the model and to solve the overfitting problem. Fine-tuning operations using the RLVS dataset resulted in significant improvements in all the datasets. This study both presents a new dataset in the field of video violence recognition and proposes an effective deep learning model for such applications [11]. A new direction towards research on the detection of real-time violent outbreaks within a crowded environment was postulated by Hassner et al., 2012. Accordingly, this research is also fundamentally conducted to develop systems to help in the earlier effective detection of violent situations happening in public places by reducing much of the human dependence on videos through electronic surveillance systems. In this respect, the feature descriptor, namely, 'Violent Flows' or ViF, is developed to classify violent and non-violent crowd behavior. The changes in the magnitude of the optical flow vectors are statistically evaluated using the ViF descriptor over time and these statistics are classified as violent or non-violent. This feature descriptor is classified using a standard linear Support Vector Machine that reacts fast in case of violent incidents. The real-world dataset of videos in this paper, on which the performance of the ViF descriptor will be tested, shows a huge outperformance of the proposed technique compared with the best current ones. Tests using the ViF identifier showed that higher accuracy rates were attained than with other methods and that the detection of violent incidents can be performed within a very short time from the beginning of the incident. This paper aims to increase the effectiveness of surveillance systems by providing an effective and fast solution for the detection of violent incidents in crowded environments [12]. Rodríguez et al. (2021) conducted a systematic review to analyze the state of the art in the computer science solutions developed to address violence against women and children. This paper had the main objective of analyzing the technological solutions developed in combating such violence, based on the current trends, architectures, technologies used, and open problems found with these solutions. The researchers analyzed academic studies published between 2010 and 2020 and evaluated solutions categorized in four main application areas: the online and offline detection of violence, security, and education. Among these solutions, technologies such as artificial intelligence, the Internet of Things, and digital serious games come to the fore. It focused on highlighting the success of machine learning algorithms, especially in finding violent internet content and detecting offline violence against women. It also discussed how there were challenges related to successfully implementing technological solutions. Various issues were raised, for example, the automation of the security devices, making the wearable technologies functional and usable, and the areas related to privacy concerns were discussed. The researchers have emphasized that developing technologies for the protection of victims is as important as addressing the very causes of violence, such as gender inequality and discrimination. This study provides an important roadmap

for future research and applications in computer science and engineering to prevent and reduce violence against women and children [13]. Roy et al. (2023) comprehensively analyzed the use of artificial intelligence (AI) solutions to prevent violence against women and girls. The main objective of the study was to examine how technologies such as AI, machine learning (ML), deep learning (DL), big data, and the Internet of Things (IoT) can be effectively used in the detection and prevention of such violence. These researchers reviewed various algorithms and applications regarding the performance and capability of these technologies for the large-scale prediction and prevention of violent behavior. They also explored how strategies developed based on the WHO's violence prevention strategy, 'INSPIRE', can be amalgamated with AI and mHealth solutions. The study found the application of AI-based crime prediction models particularly helpful in tracing and preventing violent crimes committed over the internet. The authors have claimed that hate crimes can be successfully detected using different ML algorithms like SVM, KNN, CNN, and GRU with a high degree of accuracy. It has also been underlined that the chatbot applications, for example LAW-U, are of relevance in victim support, both legally and in health matters. The obtained results show that AI might be an innovative, powerful tool in combating gender-based violence. This study brings in knowledge on how AI technologies could be applied more profoundly in this field and further provides recommendations for future research [14]. Fathima et al. developed a system for the detection and prevention of domestic violence incidents by incorporating NLP and ML techniques in their 2023 study. The main objective of the study was to alleviate the difficulties faced by victims of domestic violence and allow for emergency interventions through the early detection of such incidents. In this respect, the researchers proposed a mobile application that contains a voice recognition module, an image processing module, and an alert system. The app identifies the violent pattern of speech using NLP techniques, while image processing algorithms identify violent physical gestures. The proposed system analyzes both audio and image data for the detection of violent incidents with a high accuracy. The Convolutional Neural Networks, Random Forest, and K-Nearest Neighbor algorithms were used to analyze the audio data within the scope of the study. You Only Look Once was preferred for processing the image data. In the tests, the accuracy rates obtained with the CNN model were 93%, while the accuracy rate reached 91% in the KNN model. In the voice recognition module, the Random Forest algorithm showed 96% accuracy, while for the CNN algorithm, the accuracy was 94%. These results depict that the developed system performs highly in detecting and intervening in domestic violence. The system sends warnings of the detected incidents of violence to the emergency contacts and law enforcement agencies, and hence victims can be assisted as soon as possible. This work demonstrates that technology can be employed effectively in combating domestic violence and thus provides important grounds for new works in this area [15]. Kumari et al., in 2023, came up with a novel approach using deep learning techniques and natural language processing to detect violence through image and audio channels captured by surveillance cameras. The paper primarily worked on designing a system that could automatically detect incidents of violence, especially in crowded areas or dangerous areas, and help security forces to reach those incidents as quickly as possible. By applying TensorFlow, researchers created an object detection model that is able to recognize incidents of violence that could include weapons, fire, fisticuffs, purse snatching, and sexual harassment. They also designed an NLP-based model for detecting hate speech, sexual harassment, and offensive language using audio channels from surveillance cameras. The research used specially created image and audio datasets to improve the model's success. Violence detection has been performed based on a model that showed 84% accuracy in its work and gave effectual results under variant conditions, especially night vision ones. A model trained using TensorFlow in the object detection

arena holds great potential for the early detection and prevention of violent incidents from the results obtained. The detection of violent speech is also made possible, thanks to data processing from audio channels processed using NLP techniques. This study has succeeded in developing a powerful tool contributing to the prevention of violent incidents, offering an effective, low-cost solution to the current security systems [8]. The work of Botelle et al. (2022) investigates the performance of various natural language processing models in automatically identifying and classifying interpersonal violence events in mental health electronic health records. This study, therefore, basically attempts to develop a model using NLP that will be able to automatically identify violent incidents from the clinical texts of patients under mental health services and then test this model on the available datasets. This was an application of the CRIS database, containing over 500,000 de-identified patient records from the South London and Maudsley NHS Trust. Text fragments from those data were manually labeled with tags such as the presence of violence, the type of violence—physical, sexual, and domestic violence—and the patient's status in the violent event of being a perpetrator, victim, or witness. Then, a test of the BioBERT model was conducted, which had been trained with these labels through the 10-fold cross-validation method. It is possible to calculate the precision, recall, and F1 score of the model. The outcome obtained was that the BioBERT model had a high correctness in detecting sexual violence types of up to 97%, while generally its performance was between 89 and 98% for the accuracy and recall scores in the detection and classification of violent events. These findings suggest that NLP models can be an effective tool for the automatic recognition and management of violent incidents in mental health services. However, in the study, it was stated that NLP methods can only detect recorded events and cannot detect unrecorded violent events, and it was emphasized that ethical and regulatory issues should be considered during the integration of such models into clinical decision support systems [16]. Kumar et al. have discussed some comparative approaches to the detection of offensive and insulting language in Hindi, Bengali, and English languages using natural language processing techniques in their study conducted in 2021. The purpose of this study was to automatically detect the offensive and insulting language generally encountered in social media content and to understand the relationship between these two phenomena. In this regard, the authors of the paper performed the detection of offensive language using support vector machine and deep learning models: BERT, ALBERT, and DistilBERT. The datasets used were from shared tasks, namely HASOC-detection of hate speech and insult content in Indo-European languages, and TRAC-2-detection of aggression and misogyny. In fact, the best performing classifiers developed during this study for different tasks achieved F-scores within the range of 0.70–0.80. Cross-annotating the two datasets with the use of the classifier, the researchers examined the overlap of the subcategories of offensive and insulting language. Indeed, the results showed that the categories of aggression and insult do overlap; aggression may or may not encompass insult, and vice versa. The study identified the development of approaches from different viewpoints as important for a better understanding and detection of such complex linguistic phenomena [17]. Tabaie et al. (2022) developed a novel NLP algorithm for detecting partner violence in an emergency department setting. The main aim of the study was to develop a model capable of detecting cases of partner violence through unstructured clinical notes in EHRs and assess the effectiveness of this model. The researchers compared the performance of the NLP algorithm with ICD codes in detecting partner violence by analyzing 1,064,735 emergency department admissions in a level 1 trauma center in the USA from 2012 to 2020. The developed NLP model detected partner violence cases totaling 7399 with an accuracy rate of 99.5%, using 23 basic terms and 49 extended terms to detect situations related to partner violence. The current study showed that this NLP algorithm outperformed the ICD codes in the better

detection of partner violence and drastically reduced false positive cases. The authors asserted that the method of the current NLP approach could be integrated into health services' early detection and intervention processes concerning partner violence. Besides, it was underlined that the real-time use of this algorithm may reinforce mechanisms of intervention by the early detection of events related to partner violence. The results of this study have shown how NLP techniques can be an effective tool in the detection and solution of social problems in health services [18]. Dobbins et al. (2024) developed deep learning models that are able to predict violence and threats against healthcare workers based on clinical notes. The main purpose of the study is to develop artificial intelligence-based tools that are capable of providing early intervention against increased violence in healthcare. The researchers aimed at predicting violent incidents using two models: first, applying a document classification model based on clinical notes; and, secondly, a regression model using structured data. These achieved successful results for the document classification model, with a score of 0.75 F1, and for the structured data model, with a score of 0.72 F1. Both models were proven capable of performing beyond human expertise to surpass the 0.5 F1 score that had been recorded by the team studying psychiatry for the same set of documents. The study further developed an NER model to identify the risk factors leading to violence in clinical notes. The model recorded a general F1 score of 0.7 in identifying violence risk factors. The research suggests that these models can be used specifically to prevent violence in healthcare and generalize to other hospital systems. The results suggest that artificial intelligence and deep learning techniques can play an important role in the early detection of violence against healthcare workers [19]. A fresh, new approach was worked on and proposed by Waltzman regarding the ways and means of finding or the prevention of workplace violence using modern day AI technologies in sync and in cooperation with the traditional methodology as early as 2024. The purpose is the verification against how well AI-enabled detective devices support time-effective and accurate detection in instances related to violence in healthcare professional practice. The researchers applied NLP to analyze about 71,000 nurse handover notes for pediatric patients in a large New England city. It doubled the number of violent incidents that could not be detected using traditional methods—a testament to the importance of AI-based surveillance systems in healthcare. The study has focused on the fact that the machine learning feature of AI can detect incidents of violence in a more detailed and nuanced way. However, the system has certain limitations too; for example, the exact type of the incidents was not possible to determine. Researchers have suggested that violence is often underreported, and AI-based systems may make significant contributions to correcting this situation. The present study has pointed out the potential of AI technologies in the improvement of safety in healthcare and laid a foundation for further advanced solution development in this area in the future [20]. Figure 2 shows the quality-method scheme of studies in the literature.

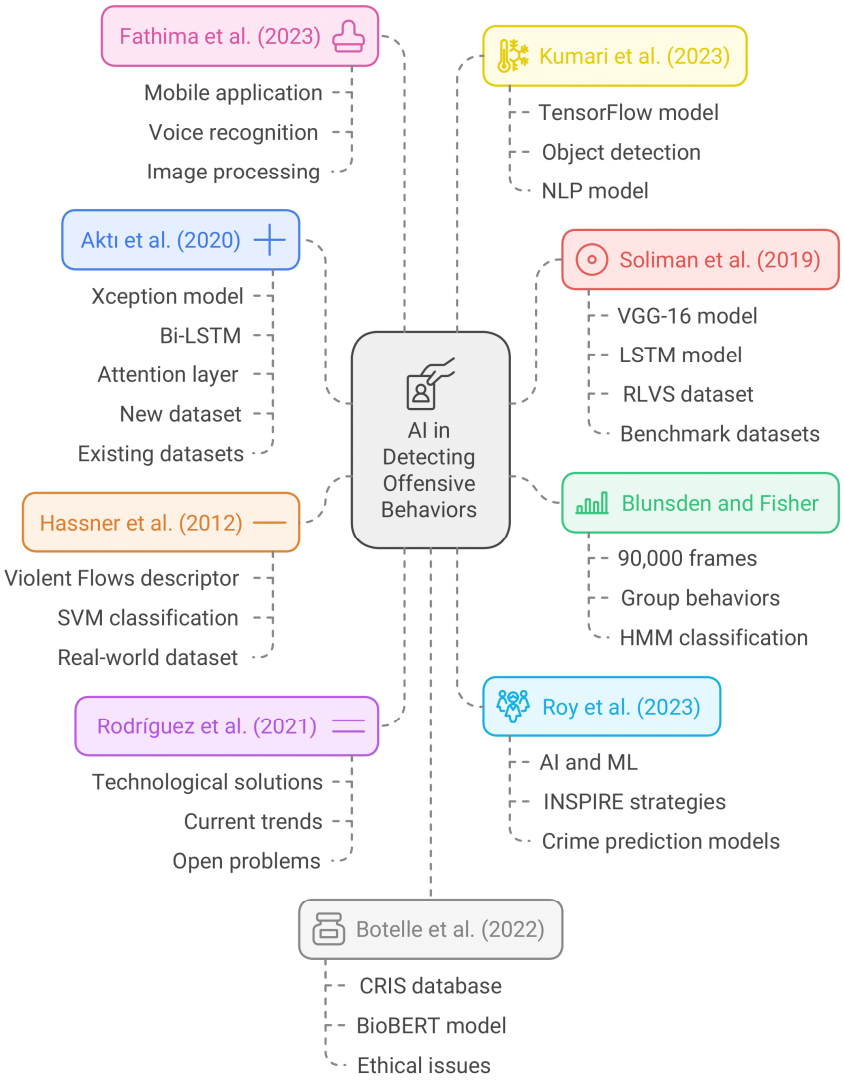

Figure 2. Quality-method scheme of studies in the literature [8–16].

3. Materials, Methods, and Findings

This section provides brief technical information on the algorithms used to train the dataset and complexity matrices, showing the classification successes obtained. The technical adjustments preferred in the training of the algorithms and the interpretations of the complexity matrix results are also provided. The dataset consists of 700 sentences, 350 each in Turkish and English, with violent and normal content that healthcare professionals encounter in daily life. In the labeled dataset, there are 304 sentences with normal content and 396 sentences with violent content. This dataset, which can be described as balanced, is divided into 70% training and 30% test data for training.

3.1. Naive Bayes (NB)

The Naive Bayes algorithm is a statistical method used in text classification and natural language processing tasks. In this study, features are extracted from texts using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. This algorithm, which assumes that the features are independent, provides fast results with a low computational cost. The Multinomial Naive Bayes (MultinomialNB) approach was used as the loss function and the default parameter $\alpha = 1.0$ was preferred for Laplace smoothing. Figure 3 shows the complexity matrix of the Navie Bayes algorithm.

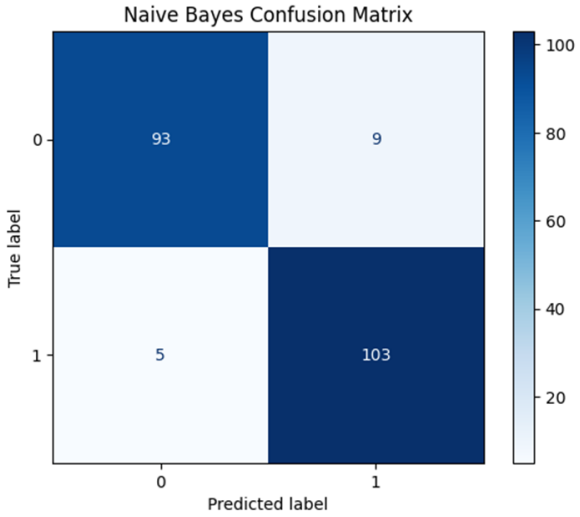

Figure 3. Navie Bayes confusion matrix.

While the model correctly classified 90 examples in the "no verbal violence" class, it incorrectly assigned 9 examples to the "verbal violence present" class (false positive). In the "verbal violence present" class, the model correctly predicted 103 instances but incorrectly classified 5 instances as "no verbal violence" (false negative). These results show that the model performs strongly in detecting verbal violence but may generate false alarms in cases where verbal violence is not present. The false negative rate of approximately 4.6% highlights the model's strong ability to capture verbal violence cases accurately. However, the presence of false positives (8.8%) suggests that in some cases, the model may incorrectly flag non-violent speech as verbal violence, which could lead to unnecessary alerts.

3.2. Support Vector Machines (SVMs)

The SVM algorithm is used to classify non-linear datasets. The radial basis function (RBF) was used as the kernel function and optimized the class separation with a hyperparameter of C = 1.0. The model provides effective classification in high-dimensional datasets and forms decision boundaries based on support vectors. Its performance was evaluated using accuracy and recall metrics. Figure 4 shows the complexity matrix of the Navie Bayes algorithm.

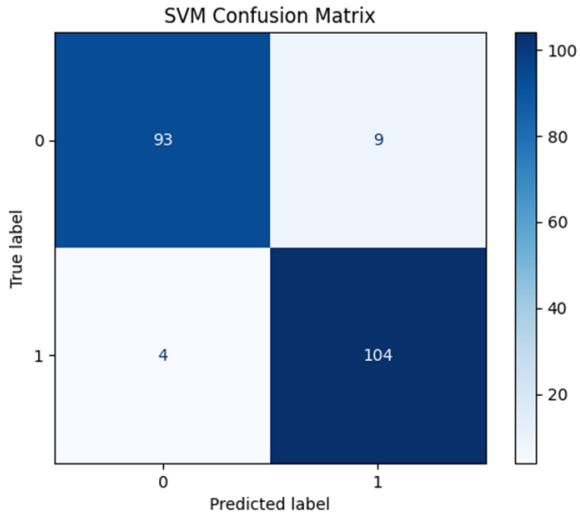

Figure 4. SVM confusion matrix.

The model made 93 correct predictions in the "no verbal violence" class and incorrectly labelled 9 instances as "verbal violence exists" (false positive). In the "verbal violence present" class, the model made 104 correct predictions and incorrectly labelled only 4 instances as "no verbal violence" (false negative). These results show that the success of the model in detecting verbal violence is quite high and that the false negative rate (3.7%) is extremely low. However, the low number of false positives (8.82%) indicates that the rate at which the model generates false alarms in non-verbal violence situations is also acceptable. This performance indicates that the SVM can discriminate between classes in a balanced way.

3.3. K-Nearest Neighbor (K-NN)

The K-NN algorithm is structured with the hyperparameter of n_neighbors = 5n. This algorithm is used to classify a data point according to its k neighbors. The model makes decisions using distance metrics in the feature space. Normalization or scaling, especially in data preprocessing steps, improved the accuracy of this algorithm. Figure 5 shows the complexity matrix of the K-NN algorithm.

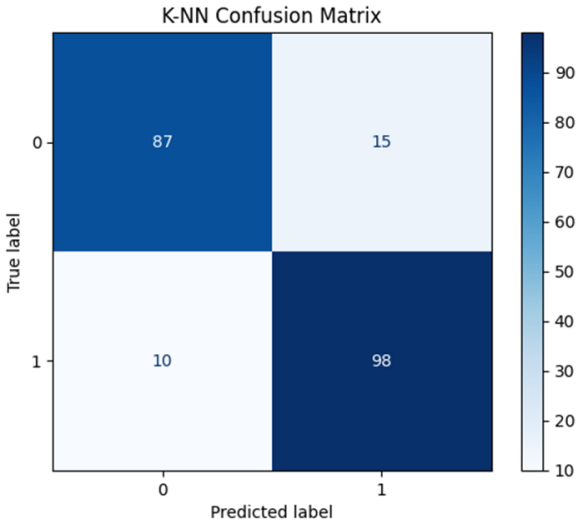

Figure 5. K-NN confusion matrix.

The model correctly classified 87 instances in the "no verbal violence" class and incorrectly labelled only 15 instances as "verbal violence present" (false positive). However, while it correctly predicted 98 instances in the "verbal violence present" class, it incorrectly classified 10 instances as "verbal violence absent" (false negative). These results show that the K-NN algorithm has a high overall accuracy rate but exhibits a higher false negative rate (9.26%) in the "verbal violence present" class. The high rate of false negatives may limit the sensitivity (recall) of the model to detect verbal violence. However, the low false positive rate (14.71%) suggests that the model provides a more reliable performance in the "no verbal violence" class.

3.4. Decision Tree (DT)

The decision tree was structured with the gini index criterion and used without a depth limit. This algorithm branched and predicted class labels based on the features in the dataset. The decision paths clearly showed the importance of each feature and its contribution to class distinction. During training, attention was paid to the dataset class balance to avoid overlearning. Figure 6 shows the complexity matrix of the K-NN algorithm.

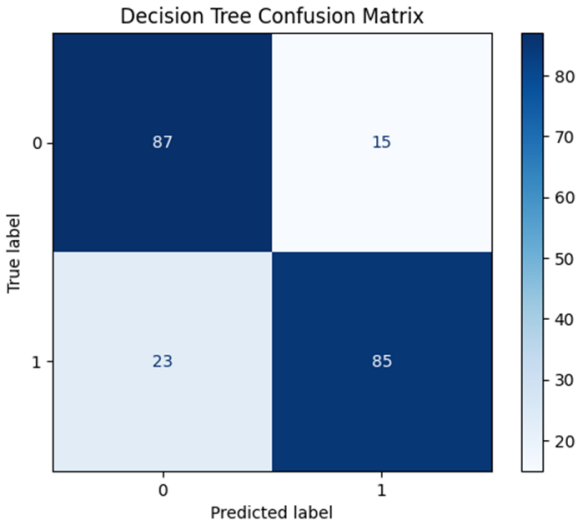

Figure 6. Decision Tree confusion matrix.

The model made 87 correct predictions in the "no verbal violence" class but incorrectly labelled 15 instances as "verbal violence present" (false positive). In the "verbal violence present" class, the model made 85 correct predictions and incorrectly labelled 23 instances as "no verbal violence" (false negative). These results show that the model performed reliably in the "no verbal violence" class (81.90% accuracy) but suffered from a slight lack of sensitivity in the "verbal violence present" class (21.30% false negative). The limited number of false positives (14.71%) shows that the model kept its errors low, except for the positive class.

3.5. Random Forest (RF)

The Random Forest algorithm is configured with the parameter n_estimators = 100. A more stable and generalizable model was obtained by combining multiple decision trees. The model made the final prediction by a majority vote of the output of each tree. This method was preferred especially to better model the complexity of the dataset and to increase the generalization capability. Figure 7 shows the complexity matrix of the Random Forest algorithm.

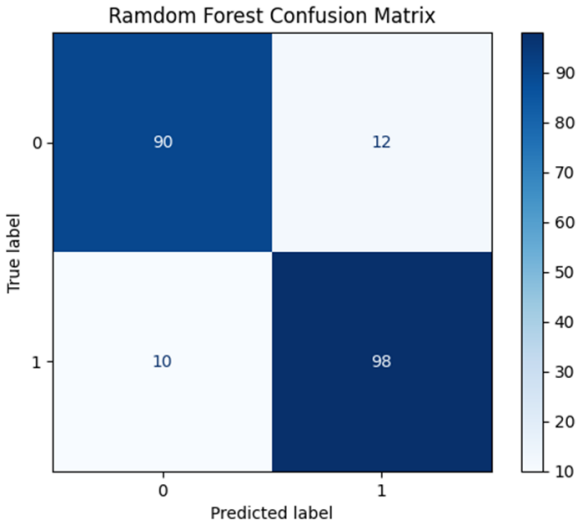

Figure 7. Random Forest confusion matrix.

While the model made 90 correct predictions in the "no verbal violence" class, it incorrectly labelled 12 instances as "verbal violence present" (false positive). In the "verbal violence present" class, the model made 98 correct predictions and incorrectly labelled only 10 instances as "no verbal violence" (false negative). These results show that the Random Forest algorithm is highly successful in detecting verbal violence (89.52% recall) and keeps the false negative rate very low. At the same time, the limited number of false positives (11.76%) shows that the model also performs well in the "no verbal violence" class.

3.6. LSTM Confusion Matrix

The Long Short-Term Memory (LSTM) algorithm is used for time series and sequential datasets. The model is configured with the parameter's units = 100, dropout = 0.2, and recurrent_dropout = 0.2 for the input layer. These parameters enable the model to learn long-term dependencies and reduce the risk of overlearning. In the training process, the loss function 'binary_crossentropy' was used for binary classification. Figure 8 shows the complexity matrix of the LSTM algorithm.

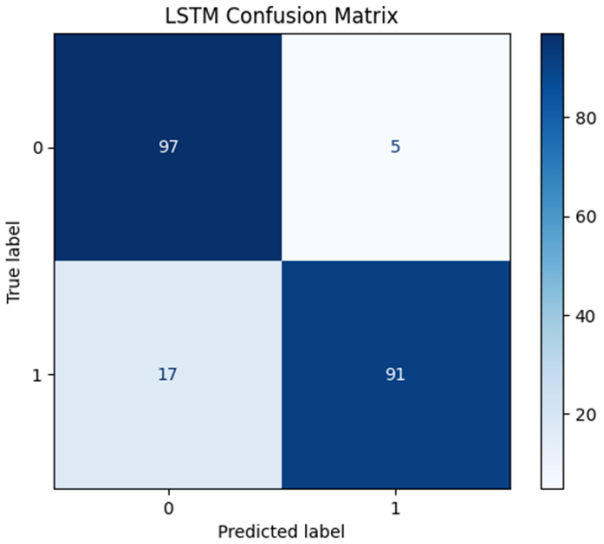

Figure 8. LSTM confusion matrix.

The model correctly classified 97 instances in the "no verbal violence" class and incorrectly assigned 5 instances to the "verbal violence present" class (false positive). In the "verbal violence present" class, the model made 91 correct predictions and misclassified only 17 instances as "no verbal violence" (false negative). These results show that the model has a high sensitivity (84.26% recall) in the "verbal violence present" class and keeps the false negative rate low. However, the presence of false positives (4.90%) indicates that the model can generate false alarms in the "no verbal violence" class.

3.7. Deep Neural Networks (DNNs)

The deep neural network (DNN) was configured with units = 128 and activation = 'relu' parameters. Binary classification was performed using activation = 'sigmoid' in the output layer. The Adam optimization algorithm and 'binary_crossentropy' loss function were used in the training process. The model was evaluated using the accuracy metric. Figure 9 shows the complexity matrix of the DNN algorithm.

While the model correctly classified 96 instances in the "no verbal violence" class, it incorrectly assigned only 6 instances to the "verbal violence present" class (false positive). In the "verbal violence present" class, the model made 92 correct predictions but misclassified 16 instances as "no verbal violence" (false negative). The high false negative rate (14.81%)

indicates that the model may miss some situations in verbal violence detection. However, the low rate of false positives (5.88%) shows that the model provides a reliable performance in the "no verbal violence" class.

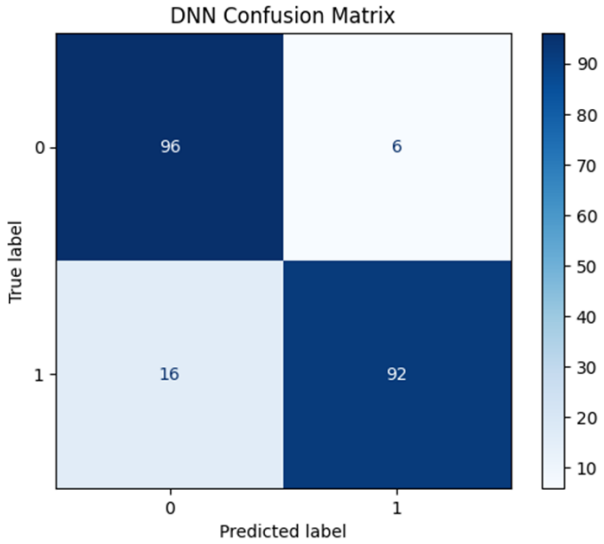

Figure 9. DNN confusion matrix.

3.8. Bidirectional Encoder Representations from Transformers (BERT)

The BERT model is configured with the parameter num_labels = 2 for the binary classification task. The inputs of the model were prepared using settings such as maximum array length (max_len = 128) and add special tokens (add_special_tokens = True). During the training process, the model was trained for 30 epochs and warmup_steps = 500 was applied for learning rate planning with the Adam optimization algorithm. Figure 10 shows the complexity matrix of the BERT algorithm.

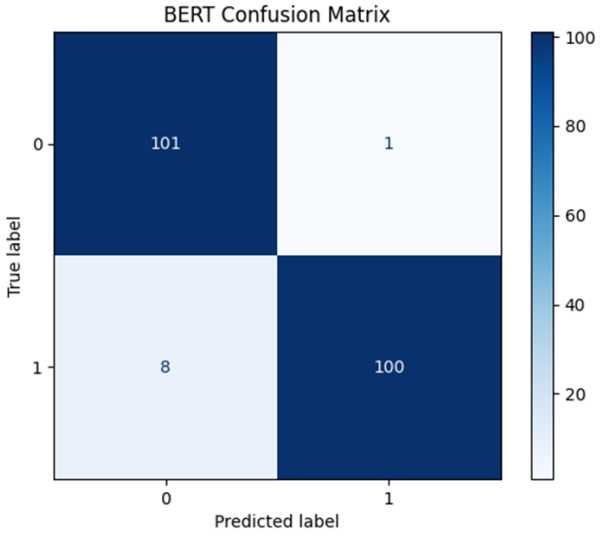

Figure 10. BERT confusion matrix.

The model correctly classified 101 instances in the "no verbal violence" class but incorrectly labelled 1 instance as "verbal violence present" (false positive). In the "verbal violence present" class, the model made 100 correct predictions and incorrectly labelled only 8 instances as "no verbal violence" (false negative). These results show that the model exhibited a high sensitivity (92.59%) and accuracy in the "verbal violence present" class,

while keeping the false negative rate (7.41%) low. The false positive rate (0.98%) shows that it may cause some false alarms in the "no verbal violence" class. In general, it is seen that the BERT model offers both a high accuracy (95.71%) and a balanced performance and has a strong competence especially in text classification tasks.

3.9. DistilBERT

DistilBERT is configured as a lighter version of BERT. For binary classification, the model was trained with parameters num_labels = 2 and max_len = 128. The model has the advantage of a low hardware requirement and a fast training process. The 'binary_crossentropy' loss function and Adam optimization algorithm were used during training. Figure 11 shows the complexity matrix of the DistilBERT algorithm.

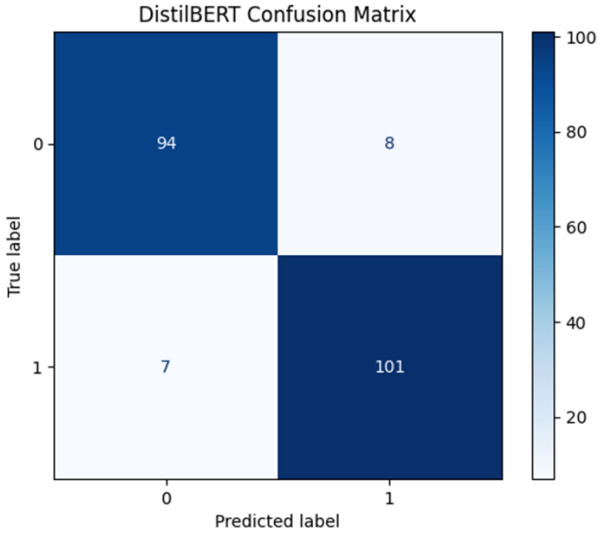

Figure 11. DistilBERT confusion matrix.

The model correctly classified 94 instances in the "no verbal violence" class and incorrectly labelled only 8 instances as "verbal violence present" (false positive). In the "verbal violence present" class, the model correctly predicted 101 instances and incorrectly labelled only 7 instances as "verbal violence absent" (false negative). These results show that the DistilBERT model provides a balanced performance in both classes. The low false negative rate (6.48%) and the low false positive rate (7.84%) indicate that the model can reliably discriminate both the "no verbal violence" and "verbal violence present" classes. Moreover, since the accuracy rate (92.86%) and sensitivity of the model are high, DistilBERT offers a similar performance to BERT with a lower computational cost.

3.10. T5 Encoder

The T5 model was used to generate text embeddings. The input texts were constrained by the parameter max_len = 128 and fixed size vectors were obtained using the final hidden state. These embeddings were then fed into the model for use in classification tasks. Figure 12 shows the complexity matrix of the T5 Encoder algorithm.

The model correctly identified all the instances in the "verbal violence present" class (100% recall) but failed to correctly classify any instances in the "no verbal violence" class. All 102 instances in the "no verbal violence" class were incorrectly labelled as "verbal violence present" (false positive). These results show that the model exhibits a high sensitivity in the "verbal violence present" class but fails completely in the "verbal violence absent" class. The false positive rate of 100% indicates that the model is unable to discriminate between classes and that there is a large class imbalance. This shows that the

T5 + GNN model cannot be considered as an effective classifier in this dataset and the model needs to be restructured or additional adjustments should be made for balanced classes.

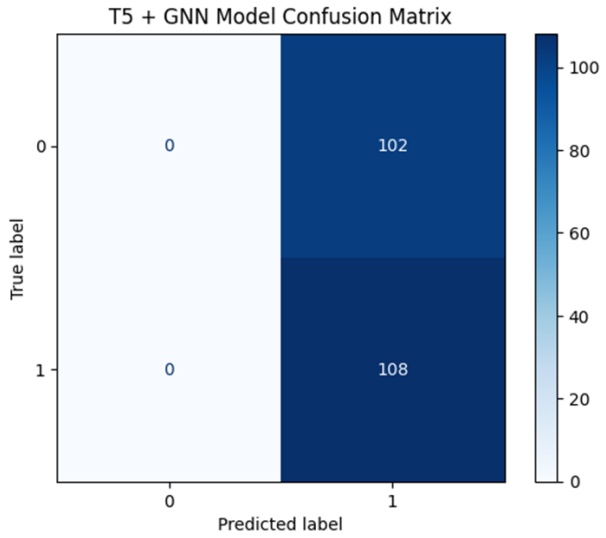

Figure 12. T5 Encoder confusion matrix.

3.11. Retrieval-Augmented Generation (RAG)

The RAG algorithm has created rich text representations with a powerful language model (based on T5) and retrieval capabilities. The model is structured with the parameters index_name = 'exact' and max_len = 128. The outputs were processed with a classifier to obtain the results. The Adam optimization algorithm and the CrossEntropyLoss loss function were used during the training. Figure 13 shows the complexity matrix of the RAG algorithm.

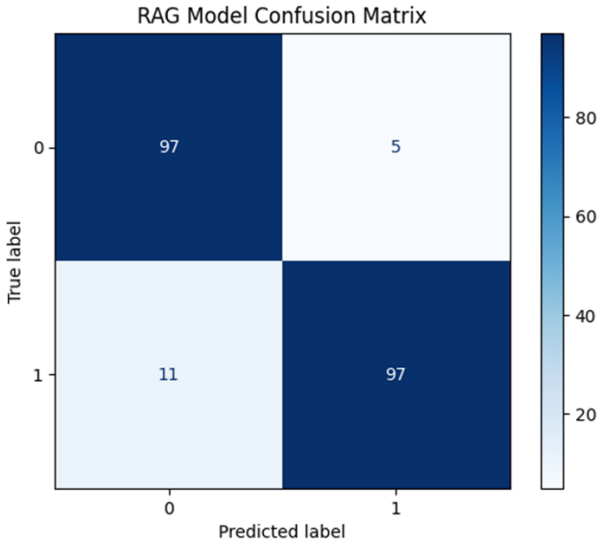

Figure 13. RAG confusion matrix.

The model correctly classified 97 instances in the "no verbal violence" class and incorrectly labelled only 5 instances as "verbal violence present" (false positive). In the "verbal violence present" class, the model made 97 correct predictions and incorrectly classified 11 instances as "no verbal violence" (false negative). These results show that the model has a very balanced performance. The low false positive rate (4.90%) indicates that the model provides a reliable performance in the "no verbal violence" class. Similarly, the

low false negative rate (10.19%) indicates that the model has a high capacity to correctly identify the "verbal violence exists" class. The RAG model can discriminate effectively for both classes with high accuracy (92.38%) and sensitivity rates.

3.12. Generative Pre-Trained Transformer (GPT)

The GPT model is used as a powerful language model that provides effective results in language processing tasks. In this study, the GPT model is used for text classification. In the training process, the Adam algorithm was preferred for the learning rate optimization of the model and the learning rate was set to 5×10^{-5}. The maximum sequence length was set to 128 and the model was trained for 20 epochs during training. Figure 14 shows the complexity matrix of the GPT algorithm.

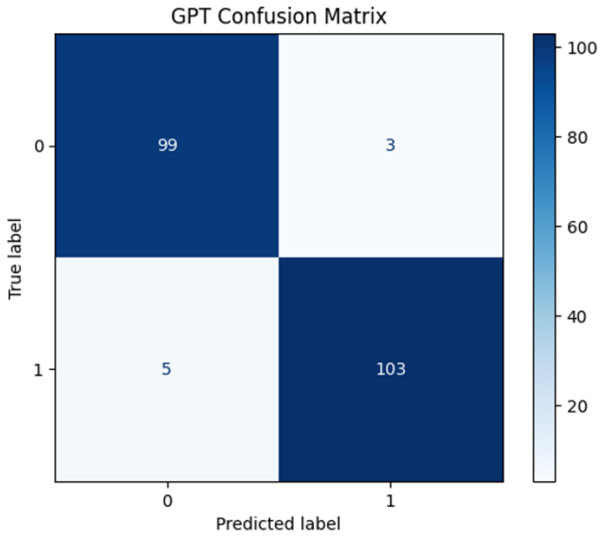

Figure 14. GPT confusion matrix.

The model correctly classified 99 instances in the "no verbal violence" class and incorrectly labelled 3 instances as "verbal violence present" (false positive). In the "verbal violence present" class, the model made 103 correct predictions and misclassified only 5 instances as "no verbal violence" (false negative). These results show that the model has an overall accuracy rate of 96.19% and a high sensitivity in the "verbal violence present" class (95.37%). However, the relatively high false positive rate (2.94%) suggests that the model may lead to some false alarms in the "no verbal violence" class. Nevertheless, the low false negative rate (4.63%) indicates that the capacity of the model to detect the "verbal violence present" class is quite good.

3.13. DistilBERT with the CNN/LSTM Hybrid Model

The hybrid combination of DistilBERT with the CNN and LSTM models was used to learn both language features and time series dependencies in text classification tasks. During training, the CNN layer was used to learn edge features, while the LSTM model modelled the sequential structure of the text. The dropout rate was set to 0.3 and training was performed for 25 epochs. Figure 15 shows the complexity matrix of the DistilBERT-CNN-LSTM hybrid model.

The DistilBERT + CNN/LSTM hybrid model performs exceptionally well with an accuracy of 94.76%, achieving a very high precision (98.02%) and a strong recall (91.67%). The false positive rate of 1.96% is extremely low, meaning that it almost never misclassifies non-violent speech as verbal violence. Additionally, the false negative rate of 8.33% is relatively low, showing that it rarely fails to detect actual verbal violence cases. With an

F1-score of 94.74%, this hybrid model emerges as one of the best models for verbal violence detection, offering a near-perfect precision while maintaining a strong recall.

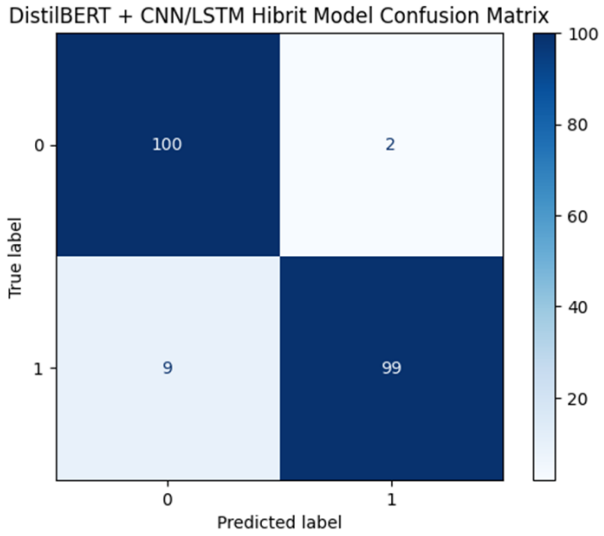

Figure 15. DistilBERT-CNN-LSTM confusion matrix.

3.14. RAG with the LSTM Hybrid Model

The RAG model combines retrieval features with deep learning models to create rich text representations. In this study, the RAG model is hybridized with LSTM to integrate both sequential dependencies and information retrieval processes. In the training process, the model was trained for 30 epochs, and the learning rate was set to 3×10^{-5}. Figure 16 shows the complexity matrix of the RAG-LSTM hybrid model.

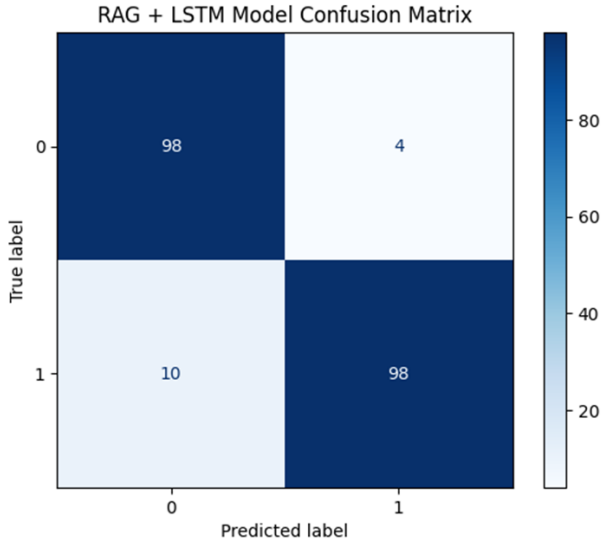

Figure 16. RAG-LSTM confusion matrix.

The model correctly classified 98 instances in the "no verbal violence" class and incorrectly labelled 4 instances as "verbal violence present" (false positive). In the "verbal violence present" class, the model made 98 correct predictions and misclassified only 10 instances as "no verbal violence" (false negative). These results show that the overall accuracy of the model is 93.33%. The sensitivity rate in the "verbal violence present" class is quite high (90.74%), which emphasizes the capacity of the model to correctly detect the positive class. However, the relatively high false positive rate (3.92%) indicates that false

alarms are slightly more frequent in the "no verbal violence" class. The low false negative rate (9.26%) keeps the probability of the model missing the positive class quite low. The DistilBERT + CNN/LSTM hybrid model shows an effective classification performance by exhibiting a high accuracy and sensitivity in the "verbal violence present" class.

3.15. TF-IDF Navie Bayes Model

The combination of TF-IDF and Naive Bayes gives effective results with a low computational cost in text classification tasks. While TF-IDF extracts meaningful features from texts, the Naive Bayes model uses these features for classification purposes. The alpha value for the Laplace smoothing is set to 1.0. Figure 17 shows the complexity matrix of the TF-IDF Naive Bayes model.

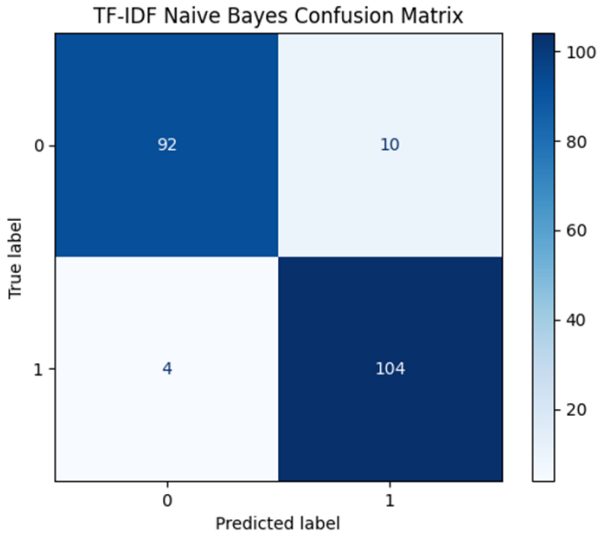

Figure 17. TF-IDF Navie Bayes confusion matrix.

The model correctly classified 92 instances in the "no verbal violence" class and incorrectly labelled 10 instances as "verbal violence present" (false positive). In the "verbal violence present" class, the model made 104 correct predictions and misclassified 4 instances as "no verbal violence" (false negative). These results show that the overall accuracy of the model is 93.33%. In the "verbal violence present" class, the sensitivity rate is 96.30%, indicating that the model detects the positive class with a high accuracy. The false positive rate is 9.80% and the false negative rate is 3.70%, which shows that the model has a balanced performance in both classes.

3.16. CNN Model

The Convolutional Neural Network (CNN) model is effective in capturing local features in text classification tasks. In this study, the CNN model is structured with 1D convolutional layers and trained for 30 epochs. The dropout rate was set to 0.4. Figure 18 shows the complexity matrix of the CNN model.

The model correctly classified 95 instances in the "no verbal violence" class and incorrectly labelled only 7 instances as "verbal violence present" (false positive). In the "verbal violence present" class, the model made 96 correct predictions and misclassified 12 instances as "no verbal violence" (false negative). These results show that the overall accuracy rate of the model is 90.95%. The false positive rate is low at 6.86%, which indicates that the model performs very reliably in the "no verbal violence" class. The false negative rate is 11.11%, indicating that the model keeps the rate of missing the "verbal violence present" class at a very low level.

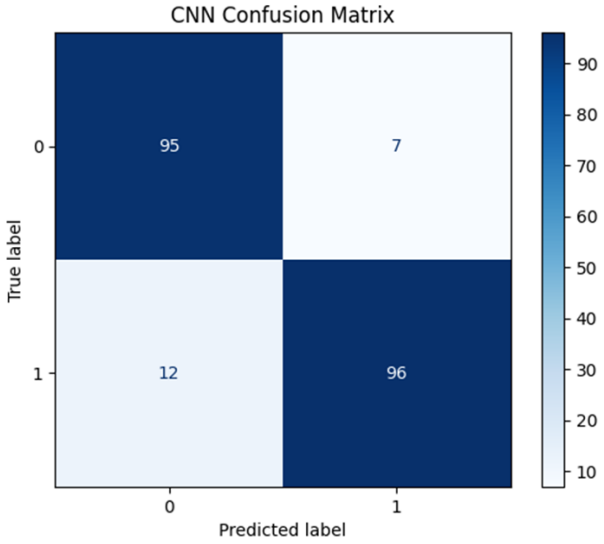

Figure 18. CNN confusion matrix.

3.17. Word Embeddings with LSTM Model

Word embeddings are used to create semantic representations of texts. In this study, they are combined with the LSTM model to achieve effective results in the text classification tasks. In the training process, the model was trained for 20 epochs, and the Adam optimization algorithm was used. Figure 19 shows the complexity matrix of the Word Embeddings-LSTM model.

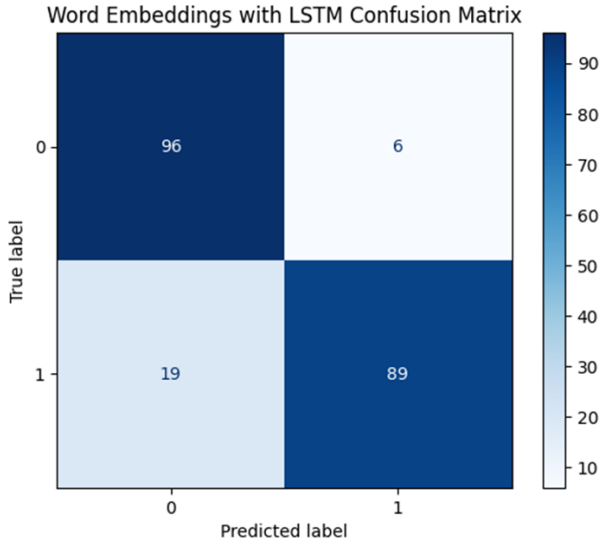

Figure 19. Word Embeddings-LSTM confusion matrix.

The model correctly classified 96 instances in the "no verbal violence" class and incorrectly labelled only 6 instances as "verbal violence present" (false positive). In the "verbal violence present" class, the model made 89 correct predictions and misclassified 19 instances as "no verbal violence" (false negative). These results show that the overall accuracy rate of the model is 88.10%. The false positive rate is low at 5.88%, indicating that the model provides a reliable performance in the "no verbal violence" class. The false negative rate of 17.59% indicates that the model missed a few instances in the "verbal violence present" class.

3.18. RAG-ECE Model

The RAG model is augmented with integrated cross entropy loss (ECE). This model is designed to optimize the information retrieval and learning processes in the text classification tasks. In the training process, the model was trained for 20 epochs, and the learning rate was set to 2×10^{-5}. Figure 20 shows the complexity matrix of the RAG-ECE model.

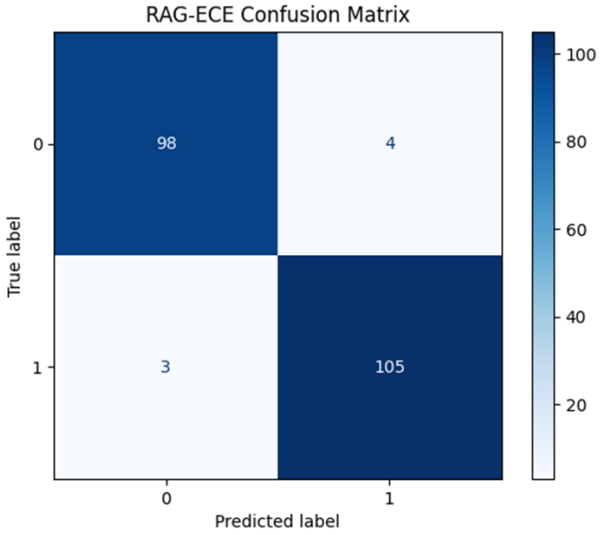

Figure 20. RAG-ECE confusion matrix.

The model correctly classified 98 instances in the "no verbal violence" class and incorrectly labelled only 4 instances as "verbal violence present" (false positive). In the "verbal violence present" class, the model made 105 correct predictions and misclassified 3 instances as "no verbal violence" (false negative). These results show that the overall accuracy rate of the model is 97.67%. The false positive rate is very low at 3.92%, indicating that the model offers a high reliability in the "no verbal violence" class. The false negative rate is 2.78%, emphasizing that the capacity of the model to detect the "verbal violence exists" class is quite high.

Model Implementation

The question encoder part of the RAG was replaced with a powerful model such as BERT. A learning rate scheduler is used to dynamically adjust the learning rate according to epochs. The early stop mechanism is enabled to prevent the model from overlearning. The embedding representations are enriched by adding self-attention or multi-head attention mechanisms. This allowed us to better capture the contextual meaning of the RAG embeddings. The learning of the contextual relationships is improved by processing the RAG-derived embeddings with a transformer encoder. This provides a structure that further enriches the contextual embedding obtained from the RAG model. In this structure, the embeddings from the RAG are passed through a fully connected layer to the enriched contextual representations. These representations are used in the classification layer to obtain the results. The RAG-ECE model architecture is shown in Figure 21.

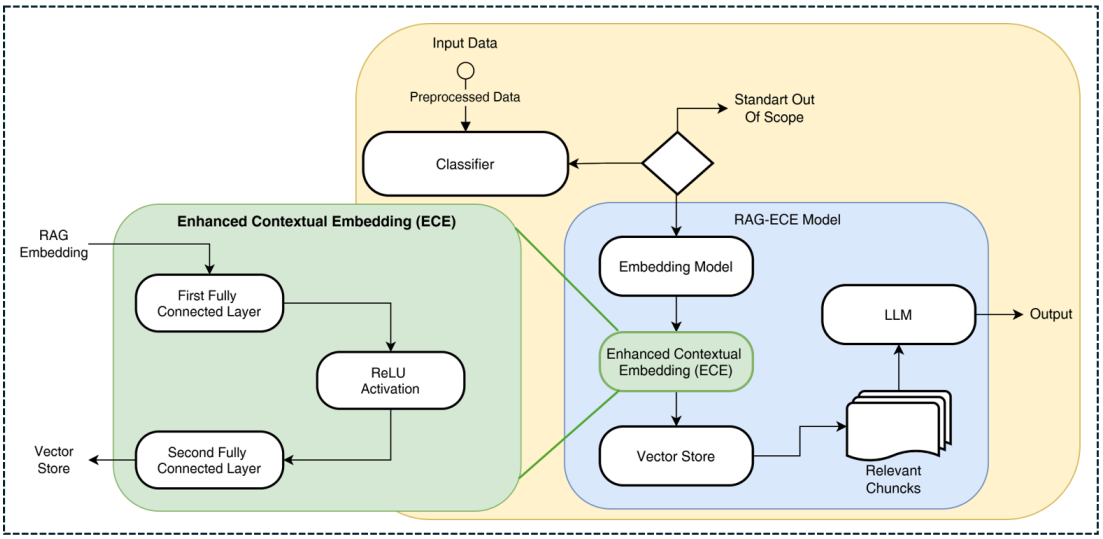

Figure 21. RAG-ECE model architecture.

The mathematical implementation of the proposed RAG-ECE model is calculated by Equations (1)–(7) [21–23].

The raw input data x are preprocessed to create an initial feature representation h_x:

$$h_x = Preprocess(x) \tag{1}$$

where $Preprocess(\cdot)$ includes tokenization, normalization, or other preprocessing tasks.

The embedding model transforms the preprocessed input h_x into a latent space representation h_{embed}:

$$h_{embed} = E(h_x) \tag{2}$$

where EEE represents the embedding model.

Relevant information chunks k_i are retrieved from the vector store based on the similarity between the embedding representation h_{embed} and the stored vector representations D = {d1,d2,. . .,dn}. The probability distribution over the chunks is computed as:

$$P(k_i|h_{embed}) = softmax(h_{embed}.d_i) \tag{3}$$

The final retrieved context k is a weighted sum of the retrieved chunks:

$$k = \sum_i P(k_i|h_{embed}).d_i \tag{4}$$

The retrieved embedding k is passed through the Enhanced Contextual Embedding module, consisting of fully connected layers and a *RELU* activation function. First fully connected layer:

$$h_{ECE} = RELU\ (w_1 k + b_1) \tag{5}$$

Second fully connected layer:

$$h_{ECE} = w_2 h_{ECE} + b_2 \tag{6}$$

Here, w_1, w_2 are weight matrices, and b_1, b_2 are bias terms.

The enhanced embedding h_{ECE} is passed to the large language model (*LLM*) for the final output generation:

$$y = LLM(h_{ECE}) \tag{7}$$

Contextual Embedding: RAG-ECE enhances the contextual representations of the texts by utilizing the retrieval capability of the RAG model, adding more fully connected layers on top.

Multi-Head Attention: A multi-head attention mechanism has been implemented on the contextual representations to make the representations more meaningful by capturing different contextual relations in texts through different heads.

Transformer Layers: The architecture adds some transformer encoder layers that enable the model to learn deep semantics. These layers process the contextual information in the text representations more powerfully and feed it to the classifier.

Hyperparameter optimization to improve model performance has played a critical role in the text classification process. Automated hyperparameter optimization tools, such as Optuna, were used to determine the parameters that would give the best performance of the model. The optimized parameters are as follows:

Hidden Layer Size: The number of neurons in the fully connected layers is optimized.

Learning Rate: The learning rate used during the training of the model is optimized so that the model learns faster and more efficiently.

Batch Size: By optimizing the batch size, the size of the data pieces used during the training is adjusted.

The optimal hyperparameters for each model were determined experimentally and evaluated according to the performance on the validation set.

Early stopping ensures that training is stopped when the validation loss does not improve for a certain period during training. In this study, early stopping is used to prevent overfitting and to optimize the training process. The verification loss was calculated at the end of each epoch and the training was terminated when no improvement was observed by a certain patience value. The patience value was determined as 3 epochs.

The performance of the trained models was evaluated with metrics such as accuracy, precision, recall, F1 score, and the confusion matrix. In addition, the learning curve of the model was analyzed by plotting the validation and training losses for each model according to the epochs. In this way, it was determined when the model started to overlearn or underlearn. The hyperparameter settings and values used for training all the algorithms are shown in Table 1.

Table 1. Hyperparameter values of the algorithms.

Model	Hyperparameter	Epochs	Batch Size
Naive Bayes	alpha = 1.0	1	16
SVM	kernel = 'rbf', C = 1.0	1	16
K-NN	n_neighbors = 5	1	16
Decision Tree	criterion = 'gini', max_depth = None	1	16
Random Forest	n_estimators = 100	1	16
LSTM	units = 100, dropout = 0.2, recurrent_dropout = 0.2	30	64
DNN	units = 128, activation = 'relu'	30	64
BERT	num_labels = 2, max_len = 128, warmup_steps = 500	30	16
DistilBERT	num_labels = 2, max_len = 128, weight_decay = 0.01	11	8
T5 Encoder	max_len = 128, hidden_dim = 768	100	32
RAG	max_len = 128, hidden_state	10	32

The main reason for using different epoch numbers is to ensure the optimal learning process for training each model and to minimize problems such as overfitting or underfitting. Simple models such as Naive Bayes or SVM usually do not require training on the data more than once. Therefore, the number of epochs is not applied (N/A), since these algorithms can achieve an optimum result with one training iteration. Deep learning models usually require a longer training period of 30 epochs to learn more complex relationships. When we analyze the table, we see that the models are trained in different epoch numbers and some models (e.g., T5 Encoder, LSTM) require a long training time, while other models (DistilBERT, RAG) complete the training process in a shorter time. This is due to the use of the early stopping technique.

4. Results and Discussion

The performance results of the algorithms in Table 2 are evaluated based on different metrics. In this evaluation, metrics such as accuracy, precision, recall, and F1-Score were taken into consideration. These metrics show the performance of the algorithms on different data distributions and classification tasks.

Table 2. Performance metrics of the algorithms.

Algorithms	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.9333	0.9196	0.9537	0.9364
SVM	0.9381	0.9204	0.9630	0.9412
K-NN	0.8810	0.8673	0.9074	0.8869
Decision Tree	0.8190	0.8500	0.7870	0.8173
Random Forest	0.8952	0.8909	0.9074	0.8991
LSTM	0.8952	0.9479	0.8426	0.8922
DNN	0.8952	0.9388	0.8519	0.8932
BERT	0.9571	0.9901	0.9259	0.9569
Word Embeddings with LSTM	0.8810	0.9368	0.8241	0.8768
TF-IDF Vectorize	0.9333	0.9123	0.9630	0.9369
CNN	0.9095	0.9320	0.8889	0.9100
GPT	0.9619	0.9717	0.9537	0.9626
DistilBERT	0.9286	0.9266	0.9352	0.9309
DistilBERT + CNN/LSTM	0.9476	0.9802	0.9167	0.9474
T5 + GNN	0.5143	0.5143	1.0000	0.6792
RAG	0.9238	0.9510	0.8981	0.9238
RAG + LSTM	0.9333	0.9608	0.9074	0.9333
RAG + ECE	**0.9767**	**0.9771**	**0.9767**	**0.9767**

When the performances of the algorithms in the block are evaluated as a whole, it is observed that each of them offers certain advantages and disadvantages. Although Naive Bayes successfully detects positive classes with an accuracy rate of 93.33% and a recall rate as high as 95.37%, it shows a precision of 91.96%, which may limit its generalizability. In contrast, SVM performs similarly with an accuracy of 93.81% and a recall of 96.30% but performs slightly better in reducing the impact of false positive classes with a precision of 92.04%. Although the K-NN algorithm provides an adequate result with 88.10% accuracy and 90.74% recall, it may be disadvantageous in larger datasets due to its 86.73% precision

and high computational cost. Similarly, Decision Tree performs relatively poorly with an accuracy of 81.90% and a recall rate of 78.70%, indicating that it has difficulties in distinguishing negative classes. On the other hand, Random Forest offers a more balanced performance with 89.52% accuracy and 90.74% recall rate, but it is notable for its increased computational cost due to the combination of multiple decision trees. While LSTM and DNN present similar results with 89.52% accuracy, respectively, LSTM's 94.79% precision value highlights its strength in learning long-term dependencies, while its 84.26% recall value suggests that it may miss some false negatives. In contrast, DNN may be more sensitive to class imbalances with 93.88% accuracy and 85.19% recall. While BERT stands out with 95.71% accuracy and 99.01% precision, its applicability in larger datasets may be limited due to its high computational cost. DistilBERT, which stands out with its lightweight structure, offers a success that is close to BERT, with 92.86% accuracy and 93.52% recall rate, but can produce faster results. However, the combination of DistilBERT + CNN/LSTM shows a superior performance by combining contextual understanding with spatial features with 94.76% accuracy and 98.02% precision. GPT stands out as one of the most powerful models in the text classification and generation tasks with 96.19% accuracy, 97.17% precision, and 95.37% recall rates. The TF-IDF Vectorization method can be considered as an effective tool especially in feature extraction and the representation of text data with 93.33% accuracy and a 96.30% recall rate. CNN, with 90.95% accuracy and balanced precision (93.20%) and recall (88.89%), is particularly successful in the extraction of spatial features in visual data. While T5 + GNN shows a poor overall performance with an accuracy of 51.43%, it is effective in detecting positive classes with a recall rate of 100%. However, the RAG algorithm offers a more balanced text classification performance with 92.38% accuracy and 95.10% precision. In particular, the combination of RAG + ECE shows a superior overall performance with 97.67% accuracy, precision, and recall rates, indicating that it is one of the strongest alternatives among the other models. These results suggest that each algorithm should be optimized for specific data types and classification problems, and that balancing accuracy, precision, and recall metrics is a critical success factor.

Implementation of the Explainable Artificial Intelligence LIME Algorithm for the RAG-ECE Model

LIME (Local Interpretable Model-agnostic Explanations) is an interpretable and model-agnostic method developed to explain the predictions made by any machine learning model. It is particularly useful in interpreting the predictions of complex models, enabling users to better understand the decision-making mechanisms of the model. By doing so, LIME aims to enhance the reliability of model outcomes and make the behavior of the model more transparent [24]. Table 3 shows six randomly selected sample sentences on which the LIME algorithm was run.

Table 3. Sample sentences randomly selected from the dataset evaluated using the LIME algorithm.

Number	Sentence (Turkish–English Original Dataset)
Sencence-1	Bu ne biçim tedavi seni dava edeceğim.
Sencence-2	How is this an indifference, I will come to you.
Sencence-3	It was better when my father came to the hospital, you made him sick.
Sencence-4	What can I do to raise my mother's morale.
Sencence-5	Bunu senin yanına bırakmayacağım, hazır ol.
Sencence-6	Tedavi sırasında dikkat etmem gereken birşey var mı?

Figure 22 demonstrates the interpretability analysis of a text classification model, specifically for predicting whether a given sentence belongs to Class 0 (non-violent) or

Class 1 (violent). The prediction probability bar on the left indicates that the model assigns a higher probability to Class 1, suggesting that the input text is classified as violent. The analysis on the right further elaborates on the word-level contributions to this classification decision by highlighting their respective importance scores. In this LIME-based explainability framework, the words highlighted in orange exert a positive influence on Class 1, strengthening the model's prediction that the sentence is violent, whereas the words in blue contribute negatively, pulling the classification towards Class 0 (non-violent). The most dominant positive contributors include "seni" (you, +0.07), "edeceğim" (I will do, +0.06), and "dava" (lawsuit, +0.05), which indicate a legal threat and contribute significantly to the model's decision towards Class 1. Additionally, the word "Bu" (this, +0.07) also appears to have a positive effect, likely due to its contextual association with assertive or accusatory language. On the other hand, words such as "ne" (what, +0.05), "tedavi" (treatment, +0.01), and "biçim" (form/style, +0.01) provide a weak counterbalance by slightly pulling the classification towards Class 0 (non-violent). However, their influence remains marginal compared with the dominant Class 1 indicators, suggesting that the model does not strongly associate them with non-violent speech in this context. The overall classification decision aligns with the semantic structure of the sentence, as the presence of legally threatening phrases (e.g., "seni dava edeceğim"–"I will sue you") strongly influences the model's prediction. The explainability framework effectively visualizes the decision-making process, demonstrating how word-level importance scores can provide transparency in NLP-based text classification models.

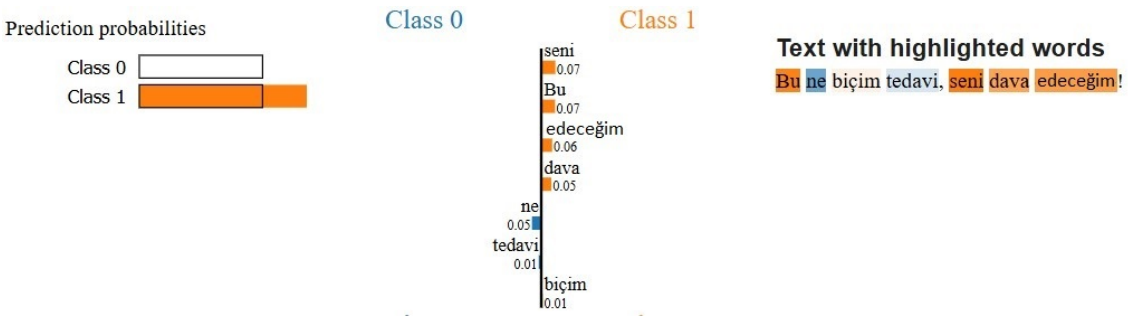

Figure 22. Sentence 1 LIME (Local Interpretable Model-agnostic Explanations) graph.

Figure 23 illustrates a LIME-based interpretability analysis for a text classification model designed to predict whether a given sentence belongs to Class 0 (non-violent) or Class 1 (violent). The prediction probability bar on the left shows that the model assigns a significantly higher probability to Class 1, indicating that the input sentence is classified as violent speech. The right-hand side of the visualization presents the word-level importance scores, where the words highlighted in orange contribute positively to the Class 1 (violent) prediction, while the words in blue negatively impact this classification, pulling the model's decision towards Class 0 (non-violent). The most influential positive contributors to Class 1 include "you" (+0.08), "come" (+0.05), and "will" (+0.03). The phrase "I will come to you!" is heavily associated with violent intent, likely due to its threatening or confrontational tone. The words "to" (+0.03), "I" (+0.03), and "indifference" (+0.03) also contribute to the classification, but their individual impact is relatively minor. In contrast, words such as "is" (+0.02), "an" (+0.02), "How" (+0.01), and "this" (+0.00) have a neutral or slightly negative effect, meaning that they shift the classification marginally towards Class 0 (non-violent). However, their influence remains insufficient to alter the overall decision. The highlighted text structure suggests that the model strongly associates direct personal references ("you") and action-oriented verbs ("come", "will") with violent intent, reinforcing its Class 1 prediction. The LIME-based explanation effectively showcases

the model's decision-making process, providing transparency on how specific linguistic elements influence classification.

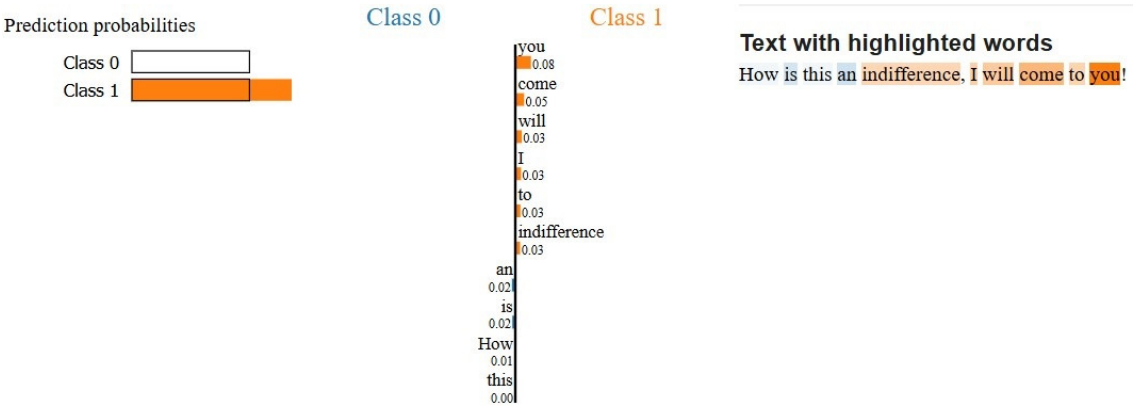

Figure 23. Sentence 2 LIME graph.

In Figure 24, the visualization presents a LIME-based interpretability analysis for a text classification model predicting whether a given sentence belongs to Class 0 (non-violent) or Class 1 (violent). The prediction probability bar on the left indicates that the model assigns a significantly higher probability to Class 1, suggesting that the input sentence is classified as violent speech. The word-level contribution analysis on the right further explains how the individual words influence this classification. The words highlighted in orange contribute positively to the Class 1 (violent speech) prediction, while the words in blue contribute negatively, pulling the classification towards Class 0 (non-violent). The strongest positive contributors to Class 1 include "you" (+0.05), "made" (+0.05), "him" (+0.04), and "father" (+0.03). The phrase "you made him sick" appears to be the primary driver of the violent classification, as it conveys blame or accusation, which may be associated with hostility or conflict in the model's learned patterns. Conversely, words such as "was" (+0.02), "my" (+0.01), "the" (+0.01), and "It" (+0.00) exert weak negative effects, slightly pulling the classification towards Class 0 (non-violent). However, their impact is minimal compared with the dominant Class 1 indicators. The results indicate that the model strongly associates direct personal references ("you") and accusatory action verbs ("made") with aggressive or confrontational speech. The LIME-based explanation effectively highlights how the model determines its classification, providing insight on its decision-making process.

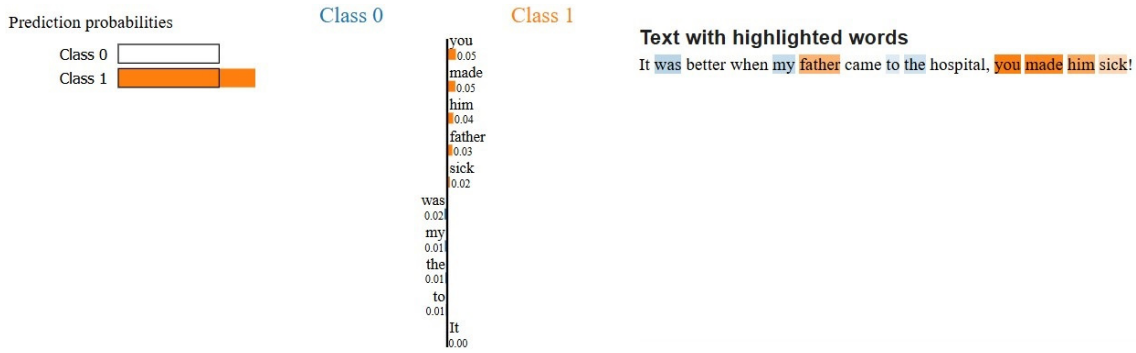

Figure 24. Sentence 3 LIME graph.

In Figure 25, the visualization presents a LIME-based interpretability analysis for a text classification model, explaining its decision-making process in classifying a given sentence as Class 0 (non-violent) or Class 1 (violent). The prediction probability bar on the left shows that the model assigns a high probability to Class 0, indicating that the sentence

is classified as non-violent speech with a strong confidence. The word-level contribution analysis on the right highlights how specific words influenced this classification. The words in blue contribute positively to Class 0 (non-violent prediction), while the words in orange contribute positively to Class 1 (violent prediction). The most dominant non-violent indicators include "can" (+0.97), "morale" (+0.30), "to" (+0.23), and "s" (+0.31), which strongly reinforce the classification towards Class 0. The word "can" is particularly influential, suggesting that the model associates it with help-seeking or constructive language, which aligns with a non-violent context. Conversely, words such as "raise" (+0.37), "What" (+0.36), "do" (+0.16), and "mother" (+0.07) exert a positive influence on Class 1, but their impact remains insufficient to shift the classification away from Class 0. This suggests that while some words may have contextual ambiguity, their overall contribution does not indicate aggression or verbal violence. The results confirm that the model effectively recognizes positive and constructive linguistic patterns, distinguishing them from aggressive or confrontational expressions. The LIME-based explanation provides a transparent view of the model's decision-making, demonstrating how word-level importance influences the text classification.

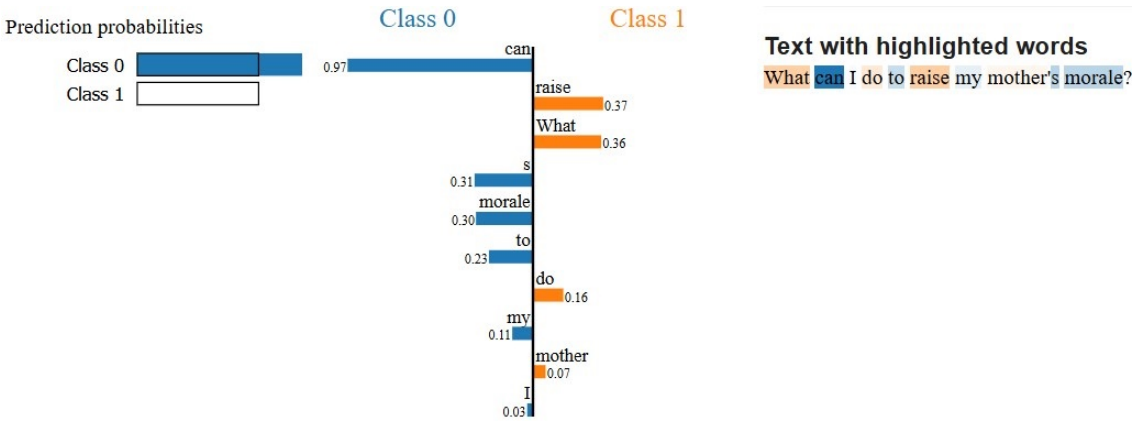

Figure 25. Sentence 4 LIME graph.

In Figure 26, the visualization presents a LIME-based interpretability analysis for a text classification model that predicts whether a given sentence belongs to Class 0 (non-violent) or Class 1 (violent). The prediction probability bar on the left indicates that the model assigns a significantly higher probability to Class 1, suggesting that the input text is classified as violent speech. The word-level importance scores displayed in the middle provide insight on how individual words contribute to the classification decision. The words in orange contribute positively to the Class 1 (violent) prediction, whereas the words in blue have a negative impact, pulling the classification towards Class 0 (non-violent). The most dominant contributors to Class 1 include "senin" (yours, +0.10), "bırakmayacağım" (I will not leave, +0.06), and "ol" (be, +0.04). These words are strongly associated with assertive or potentially threatening language, particularly in contexts where personal references ("senin") and commitment-based expressions ("bırakmayacağım") appear. The word "ol" may also carry imperative connotations, further reinforcing the Class 1 classification. On the other hand, words such as "hazır" (ready, +0.04), "Bunu" (this, +0.03), and "yanına" (next to, +0.03) exert a marginal positive impact on Class 0, suggesting that they are less indicative of violent speech. However, their influence is insufficient to counteract the dominant Class 1 predictors. Overall, the model strongly associates the phrase "senin bırakmayacağım" ("I will not leave you") with aggressive or threatening intent, leading to a Class 1 (violent speech) prediction. The LIME-based explanation effectively visualizes the

model's decision-making process, providing valuable insight on the role of specific words in shaping the classification outcomes.

Figure 26. Sentence 5 LIME graph.

In Figure 27, the visualization presents a LIME-based interpretability analysis for a text classification model, explaining its decision-making process in categorizing a given sentence as Class 0 (non-violent) or Class 1 (violent). The prediction probability bar on the left indicates that the model assigns a significantly higher probability to Class 0, suggesting that the sentence is classified as non-violent speech with a high confidence. The word-level contribution analysis in the middle highlights the influence of individual words on the classification. The words in blue contribute positively to Class 0 (non-violent prediction), while the words in orange contribute positively to Class 1 (violent prediction). The strongest non-violent indicators include "Tedavi" (treatment, +0.16) and "dikkat" (attention, +0.16), which strongly support the Class 0 classification. These words likely suggest a medical or procedural context, reinforcing a non-violent interpretation. On the other hand, words such as "bir" (one, +0.11), "etmem" (do not, +0.08), "şey" (thing, +0.06), and "gereken" (necessary, +0.05) exert a mild positive influence on Class 1, but their contributions are insufficient to shift the classification away from Class 0. This suggests that while some words may carry ambiguity, the overall semantic context remains neutral or constructive. The results indicate that the model successfully identifies health-related and advisory language as non-violent, distinguishing it from aggressive or threatening speech. The LIME-based explanation effectively showcases the model's interpretability, demonstrating how word-level importance contributes to the classification. This highlights the robustness and reliability of the model in identifying contextually non-violent expressions, making it well-suited for sentiment analysis and content moderation applications.

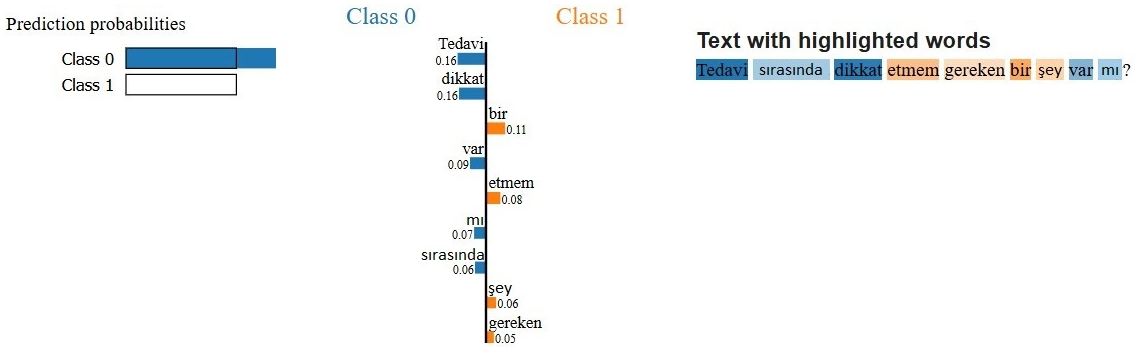

Figure 27. Sentence 6 LIME graph.

5. Conclusions

Within the scope of the study, natural language processing techniques were tested to detect expressions of violence against healthcare workers. The models and techniques analyzed in the experiments show a high accuracy with balanced performances. Among the models used, the highest performance belongs to RAG-ECE, with 97% for accuracy and

95% for the F1 score. Also, the lightweight DistilBERT model showed a high performance with 94% accuracy and a 94% F1 score. The BERT model, which is strong in understanding the complex structure of languages, also achieved an accuracy of 91%. It is hoped that the tested methods will make a significant contribution to the detection of expressions of violence against healthcare workers. NLP-based systems enable the effective identification of expressions of violence, whether in text or voice communication. Models such as TF-IDF, Naive Bayes, and SVM are considered suitable options for small- and medium-sized datasets due to their low computational costs. LSTM and RNN models, which focus on the time series analysis, successfully detect violent expressions by analyzing sequential language structures. Table 3 shows the method algorithms and performance metrics obtained in similar studies. The metrics employed in the studies listed in Table 3 exhibit considerable variability. While certain studies, such as Reference [22], exclusively utilized the F1-Score, and Reference [23] reported only accuracy, others, such as Reference [24], incorporated multiple metrics simultaneously. Consequently, the table has been structured to align with the specific metrics reported in the original texts of the referenced studies, ensuring consistency for comparative purposes.

When the studies in Table 4 are evaluated based on the presented Accuracy metric, it is seen that the RAG-ECE classification model proposed here outperforms the previous studies with an accuracy of 97%. Therefore, the model has demonstrated a reliable prediction capability by accurately distinguishing between the positive and negative classes. Although the present study falls slightly behind the references [25–27] in the F1 score metric due to the dataset size, it has demonstrated a strong overall performance by balancing the sensitivity and precision metrics with an F1 score result of 95%. In addition, the developed model differs from other works in the literature in terms of providing an innovative solution by using the RAG and ECE approaches together. The innovative structure of the model allows for the effective use of contextualized information integration based on retrieval, contributing to its outperformance of existing models in the literature.

Table 4. Comparisons with similar studies in terms of technical and performance.

Works	Year	Technique/Algorithms	Performance Metrics
Machine and Deep Learning Algorithms			
E. Mensa, et al. [25]	2020	Semantic Role Embeddings	F1-Score: 99%
A. Shome, et al. [28]	2020	GRU Model for News Article Detection	Acc: 94%
G. Karystianis, et al. [29]	2021	MLP, LSTM, Bi-LSTM, Bi-GRU, BERT	F1-Score: 78.03%
C. Raj, et al. [30]	2021	Bi-GRU, GloVe	Accuracy: 96.98% F1-Score: 98.56%
W. Aldjanabi, et al. [31]	2021	BERT, MTL	Accuracy: 97.90% F1-Score: 88.73%
C. Arcila-Calderón, et al. [32]	2021	NB, MNB, Bernoulli NB, LR, SGD, Linear SVC, RNN	Accuracy: 84% F1-Score: 86%
S. Salanterä, et al. [33]	2022	Supervised and Unsupervised BERT-NER	F1-Score: 76%
M.A. Al-Garadi, et al. [26]	2022	NLP Pipeline with LSTM, BERT, RoBERTa	F1-Score: 97%
P. Priya, et al. [34]	2023	BERT, DGCN, Caps-DGCN	Accuracy: 86.78%
Y.H. Hu, et al. [35]	2023	KNN, DT, RF, SVM, ANN, AdaBoost + RF	Accuracy: 63.9%
L. Ismail, et al. [36]	2023	RF, LR, TF-IDF	ROC-AUC: 98.1%
G. Orru, et al. [37]	2023	SVM, RF, MLP, DT	Accuracy: 93% F1-Score: 86%

Table 4. *Cont.*

Works	Year	Technique/Algorithms	Performance Metrics
A. Shome, et al. [27]	2024	NB, SVM, EGB, LR, RF, ANN, LSTM, GRU	F1-Score: 96%
		XAI Algorithms	
N.J. Dobbins, et al. [19]	2024	NER, SHAP	F1-Score: 75%
Mehta and Passi [38]	2022	LSTM, LIME, BERT + ANN, BERT + MLP	Accuracy: 97.6% Subgroup AUC: 0.8229
Pérez-Landa et.al. [39]	2021	PBC4cip	None
Mohammadi et.al. [40]	2024	CustomBERT, XLM-RoBERTa, DistilBERT, Multilingual BERT, CNN, SHAP, LR, XGBOOST	Accuracy: 79%
Ours	**2025**	**NB, SVM, K-NN, DT, RF, LSTM, DNN, BERT, Word Embeddings with LSTM, TF-IDF Vectorize, CNN, GPT, DistilBERT + CNN/LSTM, T5 + GNN, RAG, RAG + LSTM, RAG-ECE**	**Accuracy: 97.67%** **F1-Score: 97.67%**

An important contribution of this study is to demonstrate the applicability of NLP techniques in the early detection and prevention of expressions of violence to improve the safety of healthcare workers. Integrating these models into healthcare systems as part of early detection mechanisms will significantly contribute to the prevention of violent incidents. In the future, the use of larger and diversified datasets will increase the generalizability of the models in cultural and linguistic contexts. Furthermore, real-time data analytics and the development of mobile-based applications have been identified as imminent targets in the detection and prevention of expressions of violence.

Author Contributions: M.V.A.: Conceptualization, review and editing, software. M.A.Y.: Validation, writing—original draft preparation, methodology. R.G.: Conceptualization, writing—original draft preparation, methodology. A.A.: XAI Algorithms, methodology. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Isparta University of Applicable Sciences Ethics Committee has official ethics committee permission dated 4 September 2024 and numbered E-96714346-050.04-129339.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data that support the findings of this study will be made available from the corresponding author upon reasonable request.

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

References

1. Pomata, G. Sharing Cases: The Observationes in Early Modern Medicine. *Early Sci. Med.* **2010**, *15*, 193–236. [CrossRef] [PubMed]
2. Hashjin, A.A.; Irandoust, K.; Gholampoor, H.; Fischer, C.; Birjandi, M.M.; Mazdaki, A.; Abdolali, H.; Akgun, H.S. Comparing Medical Training Costs Internationally: A Systematic Review. *Med. J. Islam. Repub. Iran (MJIRI)* **2024**, *38*, 281–292. [CrossRef] [PubMed]
3. Mowery, Y.M. A primer on medical education in the United States through the lens of a current resident physician. *J. Thorac. Dis.* **2015**, *7*, E473–E481. [CrossRef] [PubMed]
4. Ramzi, Z.S.; Fatah, P.W.; Dalvandi, A. Prevalence of Workplace Violence Against Healthcare Workers During the COVID-19 Pandemic: A Systematic Review and Meta-Analysis. *Front. Psychol.* **2022**, *13*, 896156. [CrossRef]
5. Chakraborty, S.; Mashreky, S.R.; Dalal, K. Violence against physicians and nurses: A systematic literature review. *J. Public Health* **2022**, *30*, 1837–1855. [CrossRef]
6. Habeeb, M.; Rahman, U.; Divya, M.; Reddy, B.R.; Kumar, S.; Ramya Vani, P. Cyberbullying Detection using Natural Language Processing. *Int. J. Res. Appl. Sci. Eng. Technol.* **2022**, *10*, 5241–5248. [CrossRef]

7. Abarna, S.; Sheeba, J.I.; Pradeep Devaneyan, S. A novel ensemble model for identification and classification of cyber harassment on social media platform. *J. Intell. Fuzzy Syst.* **2023**, *45*, 13–36. [CrossRef]
8. Kumari, V.; Memon, K.; Aslam, B.; Chowdhry, B.S. An Effective Approach for Violence Detection using Deep Learning and Natural Language Processing. In Proceedings of the IMTIC 2023-7th International Multi-Topic ICT Conference 2023: AI Convergence Towards Sustainable Communications, Jamshoro, Pakistan, 10–12 May 2023. [CrossRef]
9. Akti, S.; Tataroglu, G.A.; Ekenel, H.K. Vision-based Fight Detection from Surveillance Cameras. In Proceedings of the 2019 9th International Conference on Image Processing Theory, Tools and Applications, IPTA 2019, Istanbul, Turkey, 6–9 November 2019. [CrossRef]
10. Blunsden, S.; Fisher, B. The BEHAVE Video Dataset: Ground Truthed Video for Multi-Person Behavior Classification. *Ann. BMVA* **2010**, *2010*, 1–11. Available online: https://www.research.ed.ac.uk/en/publications/the-behave-video-dataset-ground-truthed-video-for-multi-person-be (accessed on 16 October 2024).
11. Soliman, M.M.; Kamal, M.H.; Nashed, M.A.E.-M.; Mostafa, Y.M.; Chawky, B.S.; Khattab, D. Violence Recognition from Videos using Deep Learning Techniques. In Proceedings of the 2019 IEEE 9th International Conference on Intelligent Computing and Information Systems, ICICIS 2019, Cairo, Egypt, 8–10 December 2019; pp. 80–85. [CrossRef]
12. Hassner, T.; Itcher, Y.; Kliper-Gross, O. Violent flows: Real-time detection of violent crowd behavior. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop, Providence, RI, USA, 16–21 June 2012; pp. 1–6. [CrossRef]
13. Rodríguez, D.A.; Díaz-Ramírez, A.; Miranda-Vega, J.E.; Trujillo, L.; Mejía-Alvarez, P. A Systematic Review of Computer Science Solutions for Addressing Violence against Women and Children. *IEEE Access* **2021**, *9*, 114622–114639. [CrossRef]
14. Roy, R.; Dixit, A.K.; Saxena, S.; Memoria, M. Meta-Analysis of Artificial Intelligence Solution for Prevention of Violence Against Women and Girls. In Proceedings of the 2023 International Conference on IoT, Communication and Automation Technology, ICICAT 2023, Gorakhpur, India, 23–24 June 2023. [CrossRef]
15. Fathima, K.S.; Haroon, R.P.; Fathima, F.; Thajudeen, T.; Fudin, M. Domestic Violence Detection System Using Natural Language Processing. In Proceedings of the 2023 International Conference on Innovations in Engineering and Technology, ICIET 2023, Muvattupuzha, India, 13–14 July 2023. [CrossRef]
16. Botelle, R.; Bhavsar, V.; Kadra-Scalzo, G.; Mascio, A.; Williams, M.V.; Roberts, A.; Velupillai, S.; Stewart, R. Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: An applied evaluative study. *BMJ Open* **2022**, *12*, e052911. [CrossRef]
17. Kumar, R.; Lahiri, B.; Ojha, A.K. Aggressive and Offensive Language Identification in Hindi, Bangla, and English: A Comparative Study. *SN Comput. Sci.* **2021**, *2*, 26. [CrossRef]
18. Tabaie, A.; Zeidan, A.J.; Evans, D.P.; Smith, R.N.; Kamaleswaran, R. A Novel Technique to Identify Intimate Partner Violence in a Hospital Setting. *West. J. Emerg. Med.* **2022**, *23*, 781. [CrossRef] [PubMed]
19. Dobbins, N.J.; Chipkin, J.; Byrne, T.; Ghabra, O.; Siar, J.; Sauder, M.; Huijon, R.M.; Black, T.M. Deep Learning Models Can Predict Violence and Threats Against Healthcare Providers Using Clinical Notes. *medRxiv* **2024**. [CrossRef]
20. Waltzman, M.; Ozonoff, A.; Fournier, K.A.; Welcher, J.; Milliren, C.; Landschaft, A.; Bulis, J.; Kimia, A.A. Surveillance of Health Care-Associated Violence Using Natural Language Processing. *Pediatrics* **2024**, *154*, e2023063059. [CrossRef] [PubMed]
21. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. December 2023. [Online]. Available online: https://arxiv.org/abs/2312.10997v5 (accessed on 31 December 2024).
22. Jin, C.; Zhang, Z.; Jiang, X.; Liu, F.; Liu, X.; Liu, X.; Jin, X. RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation. April 2024. [Online]. Available online: https://arxiv.org/abs/2404.12457v2 (accessed on 31 December 2024).
23. Rau, D.; Wang, S.; Déjean, H.; Clinchant, S. Context Embeddings for Efficient Answer Generation in RAG. *arXiv* **2024**, arXiv:2407.09252. [CrossRef]
24. Ribeiro, M.T.; Singh, S.; Guestrin, C. 'Why Should I Trust You?' Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016. [CrossRef]
25. Mensa, E.; Colla, D.; Dalmasso, M.; Giustini, M.; Mamo, C.; Pitidis, A.; Radicioni, D.P. Violence detection explanation via semantic roles embeddings. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 263. [CrossRef]
26. Al-Garadi, M.A.; Kim, S.; Guo, Y.; Warren, E.; Yang, Y.-C.; Lakamana, S.; Sarker, A. Natural language model for automatic identification of Intimate Partner Violence reports from Twitter. *Array* **2022**, *15*, 100217. [CrossRef]
27. Shome, A.; Alam, M.M.; Jannati, S.; Bairagi, A.K. Monitoring human behaviour during pandemic—Attacks on healthcare personnel scenario. *Telemat. Inform. Rep.* **2024**, *15*, 100149. [CrossRef]
28. Alam, M.M.; Shome, A. Attacks on Health Workers during COVID-19 Pandemic-Data Exploration and News Article Detection using NLP and GRU model. In Proceedings of the 7th International Conference on Networking, Systems and Security, Dhaka, Bangladesh, 22–24 December 2020; pp. 3–11. [CrossRef]

29. Karystianis, G.; Cabral, R.C.; Han, S.C.; Poon, J.; Butler, T. Utilizing Text Mining, Data Linkage and Deep Learning in Police and Health Records to Predict Future Offenses in Family and Domestic Violence. *Front. Digit. Health* **2021**, *3*, 602683. [CrossRef]
30. Raj, C.; Agarwal, A.; Bharathy, G.; Narayan, B.; Prasad, M. Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques. *Electronics* **2021**, *10*, 2810. [CrossRef]
31. Aldjanabi, W.; Dahou, A.; Al-Qaness, M.A.A.; Elaziz, M.A.; Helmi, A.M.; Damaševičius, R. Arabic Offensive and Hate Speech Detection Using a Cross-Corpora Multi-Task Learning Model. *Informatics* **2021**, *8*, 69. [CrossRef]
32. Arcila-Calderón, C.; Amores, J.J.; Sánchez-Holgado, P.; Blanco-Herrero, D. Using Shallow and Deep Learning to Automatically Detect Hate Motivated by Gender and Sexual Orientation on Twitter in Spanish. *Multimodal Technol. Interact.* **2021**, *5*, 63. [CrossRef]
33. Uronen, L.; Salanterä, S.; Hakala, K.; Hartiala, J.; Moen, H. Combining supervised and unsupervised named entity recognition to detect psychosocial risk factors in occupational health checks. *Int. J. Med. Inform.* **2022**, *160*, 104695. [CrossRef] [PubMed]
34. Priya, P.; Firdaus, M.; Ekbal, A. A multi-task learning framework for politeness and emotion detection in dialogues for mental health counselling and legal aid. *Expert Syst. Appl.* **2023**, *224*, 120025. [CrossRef]
35. Hu, Y.H.; Hung, J.H.; Hu, L.Y.; Huang, S.Y.; Shen, C.C. An analysis of Chinese nursing electronic medical records to predict violence in psychiatric inpatients using text mining and machine learning techniques. *PLoS ONE* **2023**, *18*, e0286347. [CrossRef]
36. Ismail, L.; Shahin, N.; Materwala, H.; Hennebelle, A.; Frermann, L. ML-NLPEmot: Machine Learning-Natural Language Processing Event-Based Emotion Detection Proactive Framework Addressing Mental Health. *IEEE Access* **2023**, *11*, 144126–144149. [CrossRef]
37. Orrù, G.; Galli, A.; Gattulli, V.; Gravina, M.; Micheletto, M.; Marrone, S.; Nocerino, W.; Procaccino, A.; Terrone, G.; Curtotti, D.; et al. Development of Technologies for the Detection of (Cyber)Bullying Actions: The BullyBuster Project. *Information* **2023**, *14*, 430. [CrossRef]
38. Mehta, H.; Passi, K. Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI). *Algorithms* **2022**, *15*, 291. [CrossRef]
39. Pérez-Landa, G.I.; Loyola-González, O.; Medina-Pérez, M.A. An Explainable Artificial Intelligence Model for Detecting Xenophobic Tweets. *Appl. Sci.* **2021**, *11*, 10801. [CrossRef]
40. Mohammadi, H.; Giachanou, A.; Bagheri, A. A Transparent Pipeline for Identifying Sexism in Social Media: Combining Explainability with Model Prediction. *Appl. Sci.* **2024**, *14*, 8620. [CrossRef]