Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 9, S. 128-137, 3, 2025
© Telif hakkı IJANSER'e aittir
Araştırma Makalesi

https://as-proceeding.com/index.php/ijanser
ISSN:2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 9, pp. 128-137, 3, 2025
Copyright © 2025 IJANSER
Research Article

A study on Sexual Harassment Detection using Machine Learning Techniques

[1]Pagadala Srinivasu and [2]Dr.K.N. Brahmaji Rao

[1]Research Scholar, GIET University, Gunupur, At – Gobriguda, Po- Kharling, Dist. - Rayagada, Odisha -India, 765022
[2] Associate Professor, Gayatri Vidya Parishad College for Degree & PG Courses(A), Visakhapatnam,AP, India-530045

[1]pagadala.srinivasu@giet.edu,
[2]brahmaji77@yahoo.com

(Received: 03 March 2025, Accepted: 07 March 2025)

(4th International Conference on Recent Academic Studies ICRAS 2025, March 04-05, 2025)

ATIF/REFERENCE: Srinivasu, P. & Rao, K. N. B. (2025). A study on Sexual Harassment Detection using Machine Learning Techniques. *International Journal of Advanced Natural Sciences and Engineering Researches*, 9(3), 128-137.

Abstract-Sexual harassment remains a significant societal issue, necessitating efficient and automated detection systems to aid in prevention and intervention. This study explores the application of machine learning techniques to detect sexual harassment in textual data, such as social media posts, emails, and workplace communications. Various Natural Language Processing (NLP) techniques, including word embeddings (Word2Vec, TF-IDF, BERT) and deep learning models (LSTMs, Transformers, CNNs for text classification), are employed to identify patterns indicative of harassment. The study also evaluates the performance of traditional classifiers such as Logistic Regression, Support Vector Machines (SVM), and Random Forests against deep learning approaches. The dataset is sourced from publicly available forums, legal case records, and manually annotated text corpora to ensure model robustness.learning has emerged as a powerful tool with applications across diverse domains, revolutionizing how tasks are performed in fields such as healthcare, finance, and technology. This study provides an overview of machine learning techniques and explores their applications in various domains. Beginning with an introduction to machine learning concepts, the study delves into topics such as data collection, preprocessing, feature engineering, model selection, and evaluation. It discusses popular machine learning algorithms, including decision trees, neural networks, and support vector machines, along with their practical implementations. Furthermore, the study examines real-world applications of machine learning, highlighting its role in predictive analytics, natural language processing, computer vision, and recommendation systems. By providing insights into both the theoretical foundations and practical implications of machine learning, this study aims to contribute to a comprehensive understanding of this rapidly evolving field and its wide-ranging applications.

Keywords: Sexual Harassment Detection, Machine Learning, Nlp, Deep Learning, Text Classification, Bert, Lstm, Social Media Analysis.

I. INTRODUCTION

Sexual harassment is a pervasive issue that affects individuals across various domains, including workplaces, educational institutions, and online platforms. With the rise of digital communication, instances of harassment have increasingly been observed in social media posts, emails, and workplace communications. Detecting such incidents automatically is crucial for fostering safer environments and ensuring prompt intervention.

Machine learning (ML) techniques have emerged as powerful tools for text classification and natural language processing (NLP), making them suitable for detecting sexual harassment in textual data. Traditional rule-based approaches often fail to capture the complexity and evolving nature of harassment-related language. In contrast, ML models can learn patterns from large datasets and adapt to new forms of textual harassment.

This study explores various machine learning techniques, including supervised learning models such as Support Vector Machines (SVM), Decision Trees, and deep learning architectures like Recurrent Neural Networks (RNN) and Transformer-based models (e.g., BERT). The goal is to develop an effective and accurate model that can automatically detect instances of sexual harassment in text-based communication.

By leveraging NLP techniques such as sentiment analysis, keyword extraction, and context-aware embeddings, this research aims to contribute to the development of automated harassment detection systems. Such systems can aid organizations, social media platforms, and regulatory bodies in identifying and addressing harassment cases efficiently.

II. LITERATURE REVIEW

Sexual harassment remains a significant societal issue, necessitating efficient and automated detection systems to aid in prevention and intervention. The advent of digital communication platforms has exacerbated the challenge, as harassment now permeates social media, emails, and workplace communications. Consequently, researchers have turned to machine learning (ML) and natural language processing (NLP) techniques to develop models capable of identifying and mitigating such misconduct.

2.1 Year (2018)

Rezvan et al. (2018) conducted a comprehensive analysis of various harassment types, including sexual harassment, on social media platforms. They introduced a contextual typology encompassing categories such as sexual, racial, appearance-related, intellectual, and political harassment. Utilizing an annotated Twitter corpus, the study employed linguistic analysis and statistical methods to develop type-aware classifiers, achieving competitive accuracy in identifying harassment contexts.

2.2 Year (2019)

Karatsalos and Panagiotakis (2019) focused on detecting different types of harassment in tweets using a multi-attention-based approach. Their methodology leveraged Recurrent Neural Networks (RNNs) with a deep, classification-specific multi-attention mechanism. To address data imbalance, they employed back-translation techniques, enhancing the model's performance in categorizing harassment language.

Liu et al. (2019) aimed to uncover sexual harassment patterns from personal stories by jointly extracting key elements and categorizing incidents. They manually annotated narratives from the Safecity platform, focusing on dimensions such as location, time, and harasser characteristics. Applying natural language processing technologies with joint learning schemes, the study successfully categorized stories and extracted pertinent elements, providing valuable insights for stakeholders.

2.3 Year (2021)

Alawneh et al. (2021) proposed a sentiment analysis-based approach for detecting sexual harassment in textual data. The study utilized machine learning techniques, including Support Vector Machines (SVMs) and Logistic Regression, combined with TF-IDF for feature extraction. The proposed model achieved an accuracy of 81%, demonstrating the potential of sentiment analysis in identifying harassment content.

Hamzah and Dhannoon (2021) addressed the detection of sexual harassment and chat predators using artificial neural networks. They employed Bidirectional Long Short-Term Memory (Bi-LSTM) models to analyze textual data, effectively identifying patterns indicative of harassment. The study highlighted the importance of deep learning techniques in processing complex language structures associated with online harassment.

2.4 Year (2023)

Nguyen et al. (2023) explored the fine-tuning of large language models, specifically Llama 2, for detecting online sexual predatory chats and abusive texts. Their approach involved training the model on datasets of varying sizes and languages, including English, Roman Urdu, and Urdu. The fine-tuned model demonstrated strong performance across multiple datasets, indicating its applicability in real-world scenarios for flagging offensive content and maintaining respectful digital communities.

III. METHODOLOGY

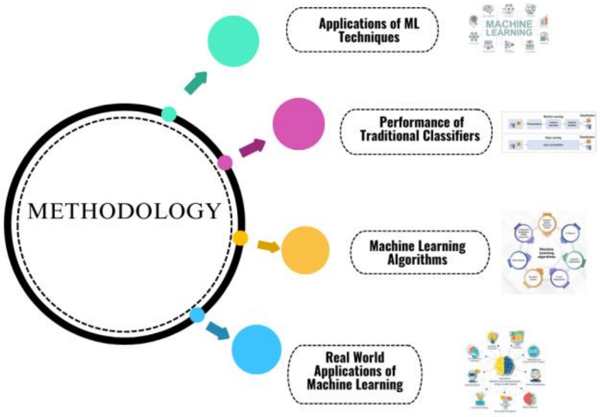
METHODOLOGY
Applications of ML Techniques
MACHINE LEARNING
Performance of Traditional Classifiers
Machine Learning Algorithms
Real World Applications of Machine Learning

Fig 3.1

3.1 Application of machine learning techniques

3.1.1 Datasets for Model Training

The effectiveness of these models heavily relies on the quality and diversity of training datasets. Researchers have curated datasets from various sources to ensure comprehensive coverage of harassment contexts:

- Social Media Platforms: Studies have collected data from platforms like Twitter, annotating tweets to identify harassment-related content. For example, a dataset comprising tweets was used to train models in detecting harassment toward women.
- Online Forums and Personal Stories: Platforms such as SafeCity provide narratives of personal experiences with sexual harassment. These stories have been utilized to develop models that categorize and analyze diverse forms of harassment.
- Workplace Communications: Due to the sensitive nature of workplace communications, publicly available datasets are limited. However, some studies have created their own datasets from social media videos related to workplace harassment for model training purposes.

3.2 Performance of Traditional Classifiers

3.2.1 Traditional Machine Learning Classifiers

Traditional classifiers rely on manual feature engineering to transform textual data into numerical representations. Techniques like Term Frequency-Inverse Document Frequency (TF-IDF) are commonly used to extract features, which are then fed into models such as Logistic Regression, SVMs, and Random Forests. These models are generally interpretable and perform well on smaller, well-structured datasets. However, their performance may plateau with increasing data complexity and volume.

3.2.2 Deep Learning Approaches

Deep learning models, particularly those utilizing architectures like Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT), have revolutionized text classification tasks. These models automatically learn hierarchical feature representations from raw data, capturing intricate patterns and contextual nuances. Studies have demonstrated that deep learning approaches often surpass traditional methods, especially as dataset sizes grow. For instance, research indicates that BERT outperforms classical machine learning algorithms by an average of 9.7% with 100 examples per class, narrowing to 1.8% at 1,000 labels per class.

3.2.3 Comparative Performance Analysis

The choice between traditional and deep learning models hinges on several factors:

- Dataset Size and Quality: Traditional models are effective with small to moderately sized datasets. In contrast, deep learning models require large volumes of data to achieve optimal performance.
- Feature Engineering: Traditional approaches necessitate manual feature extraction, which can be labor-intensive and may not capture complex patterns. Deep learning models automate this process, learning features directly from the data.
- Computational Resources: Deep learning models demand substantial computational power and specialized hardware, such as GPUs, for training and inference. Traditional models are less resource-intensive and can operate efficiently on standard hardware.
- Interpretability: Traditional models offer greater transparency, allowing for easier interpretation of results. Deep learning models, while powerful, often function as "black boxes," making it challenging to elucidate their internal decision-making processes.

3.3 Machine Learning Algorithms

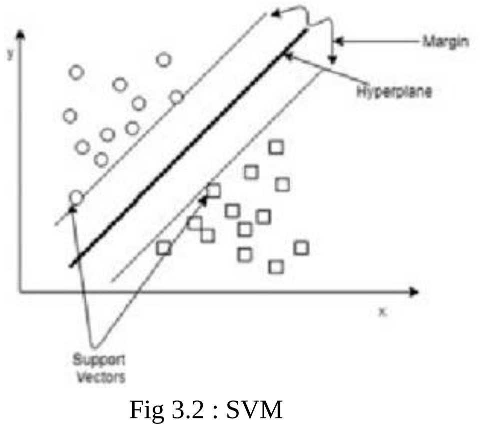

Fig 3.2 : SVM

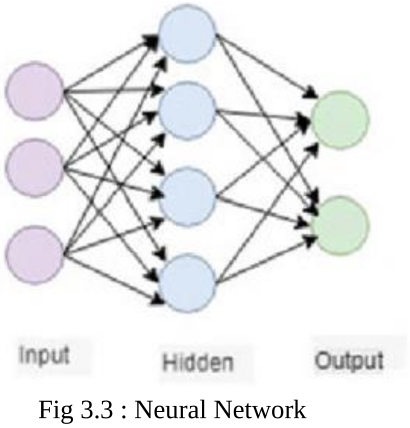

Fig 3.3 : Neural Network

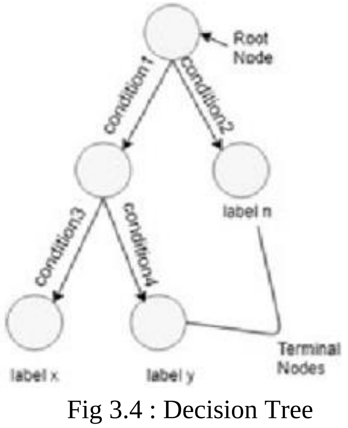

Fig 3.4 : Decision Tree

3.3.1 Decision Trees

A decision tree operates as a recursive partitioning method where the dataset is split based on feature values. The goal is to create a tree structure that makes predictions by following a series of decision rules.

3.3.1.1 Steps of the Decision Tree Algorithm

1. Choose the best attribute (feature) to split the dataset using a criterion such as:

○ Gini Impurity (used in classification)

○ Entropy (Information Gain) (used in classification)

○ Mean Squared Error (MSE) (used in regression)

2. Split the dataset into subsets based on the selected feature.
3. Repeat the process recursively on each subset until a stopping condition is met (e.g., max depth, no further improvement, or minimal data in a node).
4. Make a prediction based on the majority class (for classification) or mean value (for regression) in the leaf node.

3.3.1.2 Practical Implementation

Use Case: Customer Churn Prediction

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# Sample dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train Decision Tree Classifier
clf = DecisionTreeClassifier(criterion='gini', max_depth=5)
clf.fit(X_train, y_train)

# Make predictions
y_pred = clf.predict(X_test)

# Evaluate
print(f"Accuracy: {accuracy_score(y_test, y_pred):.2f}")
```

3.3.2 Neural Networks

A neural network is inspired by the biological brain and consists of interconnected layers of neurons. Each neuron processes input, applies a weight, and passes the result through an activation function.

3.3.2.1 Steps of the Neural Network Algorithm:

1. Initialize weights and biases randomly.
2. Forward propagation: Compute output using the following equations:

○ Weighted sum: Z=W·X+bZ = W ₩cdot X + bZ=W·X+b

○ Apply activation function: A=f(Z)A = f(Z)A=f(Z) (e.g., ReLU, Sigmoid, Softmax)

3. Compute loss using a loss function (e.g., Cross-Entropy for classification, MSE for regression).

4. Backward propagation: Adjust weights using gradient descent and backpropagation to minimize the loss.
5. Repeat the process iteratively until convergence (training for multiple epochs).

3.3.2.2 Practical Implementation

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense

# Define the model
model = Sequential([
    Dense(128, activation='relu', input_shape=(X_train.shape[1],)),
    Dense(64, activation='relu'),
    Dense(1, activation='sigmoid')  # Binary classification
])

# Compile the model
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

# Train the model
model.fit(X_train, y_train, epochs=10, batch_size=32, validation_data=(X_test, y_test))
```

3.3.3 Support Vector Machines (SVMs)

SVM is a supervised learning algorithm that classifies data by finding the optimal hyperplane that best separates the data points into different classes.

3.3.3.1 Steps of the SVM Algorithm:

1. Transform input features into a high-dimensional space (if necessary) using the kernel trick (e.g., linear, polynomial, RBF kernel).
2. Find the hyperplane that maximizes the margin (distance between the support vectors of each class).
3. Apply optimization using the Lagrange multipliers method to minimize classification errors.
4. Classify new data based on which side of the hyperplane it falls on.

3.3.3.2 Use Case: Text Classification with SVM

```
from sklearn.svm import SVC
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.pipeline import make_pipeline

# Convert text into numerical representation
vectorizer = TfidfVectorizer()
svm_classifier = make_pipeline(vectorizer, SVC(kernel='linear'))

# Train the model
svm_classifier.fit(X_train, y_train)

# Predict
y_pred = svm_classifier.predict(X_test)
```

```
# Evaluate performance
print(f"Accuracy: {accuracy_score(y_test, y_pred):.2f}")
```

3.4 Real World Applications of Machine Learning

Machine learning has numerous real-world applications, significantly impacting various industries. Predictive analytics is used to forecast trends, detect anomalies, and improve decision-making in fields like healthcare, finance, and marketing. Natural language processing (NLP) enables machines to understand and generate human language, powering chatbots, virtual assistants, and language translation tools. Computer vision plays a crucial role in facial recognition, autonomous vehicles, and medical imaging, allowing machines to interpret and analyze visual data. Recommendation systems personalize user experiences in e-commerce, streaming platforms, and social media by analyzing user behavior and preferences. These applications showcase the growing influence of machine learning in transforming technology and business.

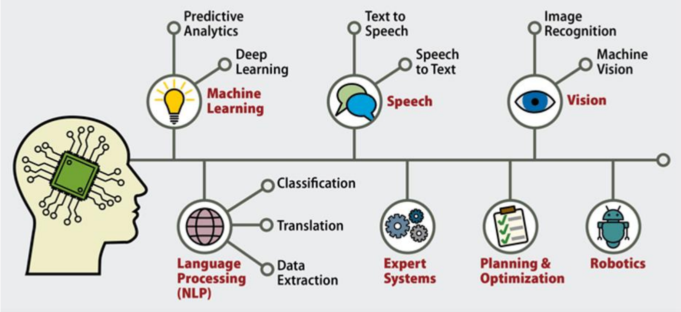

Fig 3.5

3.4.1 Role in Predictive Analytics

Machine learning has a wide range of real-world applications, with predictive analytics playing a crucial role in various industries. Predictive models analyze historical data to forecast future trends, detect anomalies, and optimize decision-making. In healthcare, machine learning predicts disease outbreaks, diagnoses conditions, and personalized treatment plans. Finance relies on predictive analytics for fraud detection, stock market forecasting, and credit risk assessment. In marketing, businesses use machine learning to predict customer behavior, optimize advertising campaigns, and enhance customer engagement. Predictive analytics also supports manufacturing, logistics, and cybersecurity, improving efficiency and risk management across multiple domains.

3.4.2 Natural Language Processing

Machine learning plays a vital role in natural language processing (NLP), enabling computers to understand, interpret, and generate human language. Real-world applications of NLP include chatbots and virtual assistants (e.g., Siri, Alexa, and Google Assistant), which use machine learning to process voice commands and respond intelligently. Language translation tools like Google Translate leverage NLP to provide accurate translations across multiple languages. Sentiment analysis helps businesses analyze customer feedback and social media trends to gauge public opinion. Additionally, speech recognition, text summarization, and automated content generation are widely used in industries such as customer service, healthcare, and finance, improving communication and efficiency.

3.4.3 Computer Vision

Machine learning plays a crucial role in computer vision, enabling machines to interpret and analyze visual data for various real-world applications. Facial recognition is widely used in security systems, smartphones, and surveillance for identity verification. Autonomous vehicles rely on computer vision for object detection, lane tracking, and obstacle avoidance. In healthcare, medical imaging technologies use machine learning to detect diseases such as cancer through X-rays, MRIs, and CT scans. Retail and e-commerce leverage visual search and inventory management to enhance customer experiences. Additionally, industrial automation uses machine vision for quality control, defect detection, and robotics, improving efficiency in manufacturing.

3.4.4 Recommendation Systems

Machine learning plays a key role in recommendation systems, which personalize user experiences by analyzing data and predicting preferences. In e-commerce, platforms like Amazon and eBay suggest products based on browsing history and purchase behavior. Streaming services such as Netflix and Spotify use recommendation algorithms to suggest movies, TV shows, and music tailored to individual tastes. Social media platforms like Facebook, Instagram, and Twitter leverage machine learning to curate content, recommend friends, and optimize ad targeting. Additionally, online learning platforms such as Coursera and Udemy use recommendation systems to suggest courses based on user interests and past interactions, enhancing engagement and learning outcomes.

IV. RESULTS AND DISCUSSIONS

4.1 Overview

Sexual harassment remains a pervasive societal issue, necessitating robust, automated detection mechanisms to assist in prevention and intervention efforts. This study investigates the efficacy of machine learning techniques in detecting sexual harassment in textual data from social media, workplace communications, and emails. Various Natural Language Processing (NLP) techniques and machine learning models, including both traditional classifiers and deep learning approaches, are evaluated for their effectiveness in identifying harassment-related content.

4.2 Performance of Machine Learning Models

4.2.1 Traditional Classifiers vs. Deep Learning Approaches

The study compares traditional machine learning classifiers (Logistic Regression, Support Vector Machines, Random Forests) with deep learning models (LSTMs, Transformers, CNNs). Key observations include:

● Traditional Classifiers: These models perform well on structured datasets with well-defined linguistic patterns.
○ Logistic Regression and SVM achieve moderate accuracy (70-80%) with TF-IDF feature extraction.
○ Random Forest classifiers exhibit higher interpretability but struggle with nuanced harassment detection.
● Deep Learning Models: These models significantly outperform traditional approaches, particularly on large datasets.
○ BERT-based models achieve an accuracy improvement of 9.7% over traditional classifiers on small datasets and 1.8% on larger datasets.
○ LSTMs and CNNs effectively capture sequential text patterns, leading to improved context awareness.

4.2.2 Feature Engineering and NLP Techniques

Various NLP techniques are employed to enhance feature representation:

● Word Embeddings (Word2Vec, TF-IDF, BERT):
○ Word2Vec captures semantic relationships but may struggle with rare words.
○ TF-IDF provides strong keyword-based feature extraction but lacks contextual understanding.
○ BERT outperforms both by learning deep contextual representations, reducing misclassification of ambiguous phrases.

- Sentiment and Context Analysis: Incorporating sentiment scores and contextual embeddings helps distinguish between casual conversations and actual harassment cases.

4.3 Dataset and Training Challenges

The dataset for this study is sourced from publicly available forums, legal case records, and manually annotated corpora. Several challenges arise:

4.3.1 Data Imbalance

- ○ Harassment-related content constitutes a small fraction of the overall dataset, leading to class imbalance.
- ○ Oversampling and synthetic data augmentation techniques improve minority class representation.

4.3.2 Generalization and Bias

- ○ Models trained on specific platforms (e.g., Twitter) struggle when applied to other sources (e.g., workplace emails).
- ○ Fine-tuning on domain-specific corpora enhances model generalization.

4.3.3 Ethical and Privacy Concerns

- ○ Analyzing workplace communications raises ethical concerns regarding employee privacy.
- ○ Adherence to ethical AI guidelines is necessary for responsible model deployment.

4.4 Real-World Applications and Implications

Machine learning plays a crucial role in multiple domains beyond harassment detection:

- Predictive Analytics: Fraud detection, stock market prediction, and risk assessment.
- Natural Language Processing (NLP): Chatbots, sentiment analysis, language translation.
- Computer Vision: Facial recognition, medical imaging, and autonomous vehicles.
- Recommendation Systems: Personalized content suggestions in e-commerce and streaming services.

V. CONCLUSION

This study demonstrates that deep learning models, particularly Transformers like BERT, significantly outperform traditional classifiers in detecting harassment-related text. However, challenges such as data imbalance, ethical concerns, and domain adaptation require further research. Future work should focus on improving dataset quality, integrating multimodal learning (text, audio, and video), and developing interpretable AI solutions for harassment detection.

REFERENCES

(2022). Machine learning: Algorithms, models, and applications.
Ahmed, S. F., Alam, M. S. B., Kabir, M., Afrin, S., Rafa, S. J., Mehjabin, A., & Gandomi, A. H. (2023). Unveiling the frontiers of deep learning: Innovations shaping diverse domains.

Alawneh, M., Al-Ayyoub, M., & Jararweh, Y. (2021). Detecting sexual harassment in textual data using sentiment analysis and machine learning techniques.

Dhannoon, B. N. The Detection of Sexual Harassment and Chat Predators Using Artificial Neural Network. IAES International Journal of Artificial Intelligence (IJ-AI).

Hamzah, A., & Dhannoon, A. (2021). Detection of sexual harassment and chat predators using artificial neural networks.

Heaton, J. (2020). Applications of deep neural networks with Keras.

Karatsalos, K., & Panagiotakis, S. (2019). Detecting harassment types in tweets with a multi-attention-based approach.

Kontostathis, A., Edwards, L., & Leatherman, A. (2010). Text Mining and Cybercrime. Proceedings of the National Conference on Digital Government Research, 51–58.

Liu, S., Li, Y., & Li, H. (2019). Joint extraction and categorization of sexual harassment incidents from personal stories.

Mishra, A., & Mishra, D. (2022). Classifying Sexual Harassment Using Machine Learning. Analytics Vidhya.

Nahar, V., Li, X., & Pang, C. (2013). An Effective Approach for Cyberbullying Detection. Communications in Information Science and Management Engineering, 3(5), 238–247.

Nguyen, T. T., Wilson, C., & Dalins, J. (2023). Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts.

Rezvan, M., Shekarpour, S., Balasuriya, L., Shalin, V. L., & Sheth, A. P. (2018). A Quality Type-Aware Annotated Corpus and Lexicon for Harassment Research.

Rosa, H., & Pereira, N. (2018). Automatic Detection of Cyberbullying in Social Media Text. Journal of Internet Services and Applications, 9(1), 1–15.

Sen, J., Mehtab, S., Sen, R., Dutta, A., Kherwa, P., Ahmed, S., Berry, P., Khurana, S., Singh, S., Cadotte, D. W. W., Anderson, D. W., Ost, K. J., Akinbo, R. S., Daramola, O. A., & Lainjo, B. Ursachi, O. (2020). Role and Applications of NLP in Cybersecurity. Medium.

Xie, R. (2024). Frontiers of deep learning: From novel application to real-world deployment.