Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Unmasking Hate: Deep Learning-Based Detection of Violent Incitement in Social Media

Shama Fatima
City University of Science & Information Technology, 2500, Pakistan
Email: shamafatima@cusit.edu.pk

Zulfiqar Ali
City University of Science & Information Technology, 2500, Pakistan
Email: zulfiqarali@cusit.edu.pk

Iftikhar Alam*
City University of Science & Information Technology, 2500, Pakistan
Email: iftikharalam@cusit.edu.pk

Said Muhammad
City University of Science & Information Technology, 2500, Pakistan
Email: saidmuhammadcs@gmail.com

Ghazi Jan
City University of Science & Information Technology, 2500, Pakistan
Email:- ghazijan817@gmail.com

ABSTRACT

The increasing usage of social media has amplified the spread of fake news as well as violent incitement content, causing critical concerns about societal harmony. Social media platforms are actively engaged in combating the dissemination of fake news around the clock; however, the overlooked content that incites violence, spreads stress, and fosters hatred has received insufficient attention. In connection, this study proposed machine learning techniques to detect and classify inciting violence in social media text. Leveraging the power of DistilBERT, a pre-trained transformer model, the research introduces a robust framework for identifying and classifying harmful content into various domains, including but not limited to political, religious, cultural, racist, and general categories. The dataset used underscores a versatile range of tweets, annotated for sentiment and contextual categories, allowing accurate examination of harmful communication patterns. The model's result, illustrated through precision, recall, and F1 score metrics, outlines its productivity in taking convoluted, context-sensitive data. This innovative approach leads to keen insights into the dynamics of online incitement, providing a solution for content moderation and refining strategies. Aside from its technical results, this study also ensures the ethical concerns of filtering out media content and the societal requirement of automated systems to tackle digital

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

hate expressions. These results allow policymakers, platform regulators, and researchers to refine content moderation practices and ensure safer online interactions. Future directions include the joining of multilingual datasets and filtering models for nuanced hate speech detection across vast cultural contexts. This research is a step towards leveraging artificial intelligence to mitigate the most pressing challenges in the digital age, incitement to violence, and instilling more inclusive online environments.
KEYWORDS: Violent Incitement, Hate Speech, Deep Learning, Social Media, Sentiment Analysis

Introduction

Information and Communication Technology (ICT) and the Internet have had a profound impact and have become an integral part of human life [1]. In the past, the introduction of Web 2.0 brought read/write web facilities and gave birth to the social web. The social web's positive consequences on societal morals, values, beliefs, and ways of thinking are quite redundant. Facebook, Twitter, Instagram, and platforms like these provide people with awareness about varied beliefs, cultures, and perspectives, enhancing empathy and tolerance. Besides possible advantages, there are a few drawbacks to using such publicly accessible technologies. For example, the internet is full of harmful material that can change people's perceptions. A person, including children, can be hacked, bullied, tortured, or attracted to crimes that most people are not aware of, especially in developing countries [2]. These overlooked crime factors and harmful material on the internet may promote crimes, create mobs, and create hidden networks of people that can hardly be detected. To tackle such issues, there should be some countermeasures generally advised by the tech giant and a system that fits our cultural and social needs. It is important to deal with unauthorized acts that involve electronics or technology, including the Internet [3]

Social media is used by everyone irrespective of gender, age, culture, and region. It has both positive and negative effects on people [4]. The positive consequences of social media on societal morals, values, beliefs, and ways of thinking are quite redundant. Facebook, Twitter, Instagram, and many more platforms like these provide people with awareness about varied beliefs, cultures, and perspectives, enhancing empathy and tolerance. Hence enhanced the quality and rate of collaboration [5]. Despite having positive outcomes, social media also has negative consequences on social morals, values, beliefs, and ways of thinking. Fake news is spreading like a shot that can deform public perception. Some platforms normalize harmful behavior and compromise societal values and beliefs. When using social media, we are entering into a public realm that is open to all and where an exhibition of wealth, material possessions, and views influences people's perceptions [6].

Different laws have been enacted to protect privacy and consumer rights; however, consumers should be aware of his/her rights to exercise them effectively [7]. People using social media are mostly concerned with maintaining an online facade and comparison with others, which leads to anxiety and depression. Addiction is also a factor of excessive use of social media, affecting their daily life activities and relationships. Many popular social media platforms provide opportunities for individuals to victimize and bully others with negative comments and posts [8].

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Misinformation and disinformation are two categories for the dissemination of false information, where misinformation probably misleads people due to misinterpretation, whereas disinformation is intended to mislead people [9]. Social media uses algorithms that impede public perceptions and critical thinking, as it limits exposure due to biases with users' existing beliefs. Relationships and online connections lack emotions and depth, which can lead to conflicts and cyberbullying, engaging people in harmful behaviors and psychological suffering. Romantic relationships are mutual relationships demanding balanced satisfaction, but if a relationship is established on social media, then appropriate use of this platform is very important; otherwise, it may backfire, causing emotions like jealousy, monitoring of partner, infidelity, and distraction [10].

Malicious/fake users use social media to manipulate public opinions and spread propaganda. This activity influences democratic processes and creates societal divisions. Nowadays, social media acts as a home ground for malicious users to spread fake news, rumors, conspiracies, and other forms of deception [11]. Social media escalates extremist views and polarizes political discourse. The algorithms of these platforms often promote diverse content, inflaming social disagreement. The broad reach of social media enables individuals to promote and spread extremist dialogues and content that incites violence in public. Algorithms of these platforms maximize user engagement and boost provocative content, including incitement to violence and hate speech. Social media can be used as a platform to normalize hate and extremist ideologies among youth who are more exposed to the internet. Such a sort of normalization can make racial, religious, or political violence more common. Young people tend to suffer more from the mental, physical, and psychological consequences [12].

There is a broader term known as “Cybercrime”, which means that crimes are related to the internet in general and electronic devices, such as smartphones, in specific. Many countries have developed different procedures and mechanisms to deal with such crimes, however, with the passage of time and the advancement of technology, the laws and Acts need to be revised, especially in developing countries. In some cases, the crime can be interpreted in different ways because it is difficult to measure someone’s intentions about a crime [13]. It means that a person committing a crime is sometimes unaware of the severity of the crime.

Cybercrimes have increased to the extent that they are now considered an obstacle in the path to success for so many applications, including the economic growth of a country. For instance, newspapers are filled with cybercrime news, including financial fraud, hacking, pornography, stalking, cyberbullying, false admission papers, fee receipts, certificates, unauthorized access during the examination, and much more, when most of them are still not reported and recorded. All these lead to a very hazardous situation in society. To deal with it effectively, we need to look at the sociological and technological factors influencing cybercrime and cybersecurity [14]. Specifically, inciting violence is a global issue. Verbal violence has adverse effects on people, as it plays a vital role in causing physical violence [15, 16].

Sentiment analysis focuses on understanding the given text or sentence's meaning using Natural Language Processing (NLP). It bridges the gap between humans and machines, enabling it to understand and process human languages[17]. Sentiment analysis

Online ISSN 3007-3197
Print ISSN 3007-3189

analyzes the texts to identify the emotional aspects found in them [18]. Due to variation in texts, deep learning methods are given priority as compared to traditional machine learning methods. This work proposes a deep-learning approach to detect inciting violence in textual content only. We are not considering images, audio, and videos due to their heavy computational constraints. These may be our future work. We will use the most popular deep learning method, "Transformers". The Python libraries will be used for training and testing purposes. The emotions and multiple features in a textual post will be identified using sentiment analysis [19]. By analyzing datasets, we will further understand the characteristics of textual data on social media to improve security and reduce cyber terrorism [20] .

The target is to use Twitter as our case study. This study aims to detect cybercrimes in textual posts on social media, specifically on Twitter. We have tentative deep-learning techniques, for instance, Transformers. convolutional neural networks CNN and Recurrent neural networks (RNN) [21]. Sentiment analysis is defined to be a field within NLP that focuses on evaluating the emotional tone and rhyme in a text. Recent work focuses on fine-tuned pre-trained transformers [22]. It includes marking the content as positive, negative, or neutral based on sentiments. This analysis is essential for understanding public opinion and detecting hate or violence that promotes extremism. The enhanced occurrence of incitement to violence and offense on text-based online platforms urgently requires a thorough analysis and models to handle such problems [23].

Transformers like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformers) are pre-trained on tremendous amounts of data. Data must be preprocessed before being fed into models [23]. By using transformers for sentiment analysis, researchers can improve accuracy and precision for detecting engagement in violence on social media. Fine-tuning provides a consistent way to examine the impact across data splits [24]. Platforms like Facebook, Twitter, YouTube, and Instagram, etc., have revolutionized communication with various opportunities for positive involvement but have also opened the door for significant challenges especially spreading harmful, hateful, and violent content resulting in concerns like violent incitement, hate speech and manipulation in opinions leading to serious issues like riots, social unrest, etc., [1].

Traditionally, semantic analysis was limited to binary or ternary classifications like positive, negative, or neutral [13]. However, modern society needs fine-grained classification identifying whether a post is politically, culturally, or religiously offensive, or expressing general frustration, offense, or incitement. Sentiment analysis, a subfield of NLP, acts as a significant tool for textual content categorization and understanding. Sentiment analysis analyzes the patterns in the content, whether it contains positive, negative, or neutral emotions. An individual would never radicalize and show extremist behavior without having an indirect or direct relation with such material [25] . The dependency of people on social media to achieve their life goals is leading to crimes and security concerns [26]. Such issues spiked during crises, such as COVID-19, when people were isolated and became dependent on social media, examining the effect showed that the users were psychologically affected, and the spread of panic, terror, and fear was increased [27].

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

The detection of extremist and depressive texts using different social media platforms was done through Convolutional Neural Networks (CNN) and Classical Machine Learning Algorithms [28]. Pre-trained models on large datasets were used for the detection of characteristics of textual data and to improve detection efficiency [29]. Despite the advancement of sentiment analysis and transformer models, incitement to violence detection on social media is difficult, and maintaining a respectful online environment is a challenge [30]. Therefore, this research explores DistilBERT, a compact and efficient transformer model, to perform multi-level classification of tweets, enhancing the detection of incitement to violence in textual content.

To detect "inciting violence" automatically, many options/solutions are available. However, AI and deep learning are providing state-of-the-art protection against, e.g., hacking, cracking, and incitement to violence. To the best of our knowledge, no such work that exists and uses deep learning for detecting inciting violence fulfills the compelling needs of society. The study aims to propose a solution that fits the society's needs and cultural values, and datasets used to incite violence. The existing generalized proposed solutions hardly fit and are sometimes not aligned with most of the societal needs. For example, a slang word in the text may not be considered slang in other regions of the world.

The study attempts to answer the following questions:

How do cultural and regional differences impact the detection of inciting violence on social media?

How efficient are the social media platforms in fighting cybercrime?

How can content monitoring systems be improved?

This proposed study has two main objectives: To review state-of-the-art literature for prominent issues and challenges.

To detect inciting violence in the textual content of social media using deep learning approaches.

To evaluate and compare the results with existing solutions

The rest of the paper is organized in this manner. Section 2 discusses the detailed yet comprehensive literature review. Section 3 discusses the research methodology and its components, while Section 4 presents results and discussion. Section 5 discusses and visualizes the results. Section 6 discusses the effectiveness of the model, and the last section is the conclusion and future work. References are listed at the end.

Literature Review

The overlooked nature of cybercrime is equally harmful and may cause severe damage to deteriorate society. According to the report, cybercrime cost six trillion dollars in 2021, up from 5 trillion dollars in 2015, according to the editor-in-chief of the annual cybercrime report [31]. In 2025, the world will preserve 200 zettabytes of information, and 50% of it will be stored in the cloud, predicted to Cisco's published statistics. More than 70% of trading will be for criminal activities in cryptocurrency [31]. Every citizen has the right to be treated according to the law, according to Article 4 of the Constitution. This means that any violation of a citizen's rights, whether committed by a private or governmental individual or body, must be justified following a national law [32]. An

Online ISSN 3007-3197
Print ISSN 3007-3189

individual would never radicalize and become an extremist without having direct or indirect relations with extremist material or people, either online or offline [25].

Social media is a place containing large amounts of user-generated data every day, and people are becoming dependent on the internet and social media to achieve their goals in life, leading to security concerns and crimes [26, 33]. The effect of social media was examined, and the result showed that users were psychologically affected, spreading panic, terror, and fear among people [27]. The use of social media platforms such as Facebook, Twitter, Snapchat, etc., and the detection of depressive and extremist texts through different machine learning approaches are done including CNN and classical machine learning algorithms [28]. Deep learning is an approach that utilizes artificial neural networks to imitate human brain functioning It makes the computer learn from the data given to it in large amounts and make predictions accordingly [34]. Pre-trained models on large datasets for detecting general representation and characteristics of textual data which is then improved to be used for effective detection [29]. The data created on social media is vast, and it is challenging to differentiate data that is related to extremism or that is a personal belief, political viewpoint, or joke [35].

A deep learning architecture is prepared using different algorithms such as transformers, CNN, and Recurrent Neural Networks (RNN) for apprehending patterns in textual data [36]. Patterns are identified using machine learning approaches in the textual data of online conversations to prevent crimes before they happen [37]. The rise of social media platforms has revolutionized how people communicate and work in this new and vastly growing era, we can share data and communicate globally with anyone in the world in just a single click. However, with this revolution, we are also facing serious concerns. Inciting violence has become a great issue in today's times. People spread hate speech, extremism, and calls to violence. High-profile individuals influence people, and therefore, freedom of speech in some cases may be harmful [38]. The main purpose of our research is to detect content inciting violence. It is really important for languages where tiny amendments can change the meaning and context of the discussion [39]. If social media is validated and regulated, it will have a positive effect on individuals [40].

Social Media's Evolution and Impact

Fake news on social media became popular, which caused people to doubt science, true news, and social norms, which had a great impact on societal values and opinions of important issues [41]. Facebook, Twitter, Instagram, and platforms like these now serve a paramount role in engaging global audiences and influencing everything from cultural trends to political trends [42]. This progression in Communication tools and methods has brought a significant shift in how information is used, shared, and modified. As the influence grows with time, their potential to be used for both negative and positive engagement often harmful interactions, is widened too. The insidious effect of social media on human life has demanded a thorough understanding of its pressure and influence on mental health, raising the question of whether the use of social media is harmful or addictive [43]. Research has revealed a critical effect on the polarization of public conversations, fueling violent interactions that spread negative impact, escalating offline violence and political violence with extremist eloquence [44].

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

The Trend of Online Incitement

Majorly, during political unrest, social media often intensifies tensions, which leads to real-world outcomes. Research and studies have shown that the global reach of these platforms has magnified the spread of provocative content that people avoid sharing in public settings. This issue has led to a requirement for research focused on recognizing the mechanisms behind such activities and maturing strategies to lessen their effects. NLP is advancing with each passing day, leveraging state-of-the-art models such as GPT-3 and BERT for a refined perception of abusive behavior [45].

Sentiment Analysis: A Tool for Knowledge

Sentiment analysis, a subfield of natural language processing (NLP), has surfaced as a significant tool for understanding and categorizing textual content [46]. It analyzes language patterns and determines whether a piece of text contains positive, negative, or neutral emotions. When we talk about social media, there is a vast content that needs to be monitored for indications of harmful behavior, including incitement to violence. Many fields have illuminated the growing concern surrounding online hate speech, but a standardized definition of what is considered hate speech is still sparse [47]. 'Harm Potential' is defined as the capability of an online public post to cause real-life physical harm or Violence, which is often stimulated by the web combining several online tactics and strategies of social media, resulting in targeted physical Violence [48].

The Elevation of Transformers in Textual Evaluation

The emergence of transformer models, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), has transformed the field of text analysis. Their rapid evolution in NLP (Natural Language Processing) has played a crucial role in achieving breakthroughs [49]. Making these tools ideal for sentiment analysis tasks, their ability to understand and process complex language patterns has been considerably improved. For a more sophisticated examination of textual data, Different platforms have been used by extremist groups to carry out evil activities [50].

Barriers to Detecting Incitement to Violence

Despite the advancement of sentiment analysis and transformer models still incitement to violence on social media is still a complex challenge. With the rise in the variety of content on social media due to technological advancements, maintaining a respectful online environment is a challenge [30]. One of the crucial difficulties lies in the vagueness of language, words, and phrases, as it may be harmless in one context but highly controversial in another. The cultural and linguistic diversity of social media users is an added factor in complexity, as sentiments and meanings differ across different groups.

Integrating the Elements: Prospects for Future Research

Sentiment analysis and transformers have made considerable progress in text analysis, but still, various gaps justify further investigation in this field. Future research can emphasize the cultural sensitivity of sentiment analysis tools, develop more vigorous methods for the detection of incitement to violence, and enhance the insightfulness of transformer models. Furthermore, it is important to address the ethical considerations of

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

automated content moderation and the potential for bias in such systems. If we could address these gaps, researchers could contribute to the advancement of effective ethical solutions for regulating and monitoring social media content.

Proposed Methodology

To achieve the research problem, a mixed approach is needed. Transformers and Datasets are used to deal with the issue of inciting violence. An existing literature review is utilized. Textual data on social media is analyzed semantically through semantic analysis. The focus of this research is on the detection of incitement to violence in the context of social media using sentiment analysis and Transformers models. The proposed methodology utilizes a Twitter reviews dataset. The dataset will be categorized into three main groups: Positive, Negative, and neutral reviews. Within the negative reviews, further subcategories will be formed based on political, religious, cultural, hate speech, offensive, and general content. Subsequently, the dataset will undergo training using a sentiment classification using DistilBERT, a pre-trained transformer model, used for its efficiency and effectiveness in the classification of textual data.

A resource like 'SENTIWORDNET' to assign sentiment scores or labels, namely positive, negative, and neutral For gaining insights into the primary discussion and trends within data, a classification model or algorithm, such as Transformers, will be employed. This model will be trained on 80 percent of the dataset and tested on the remaining 20 percent. Ultimately, the research will culminate in the presentation of findings related to the distribution of sentiment within the dataset.

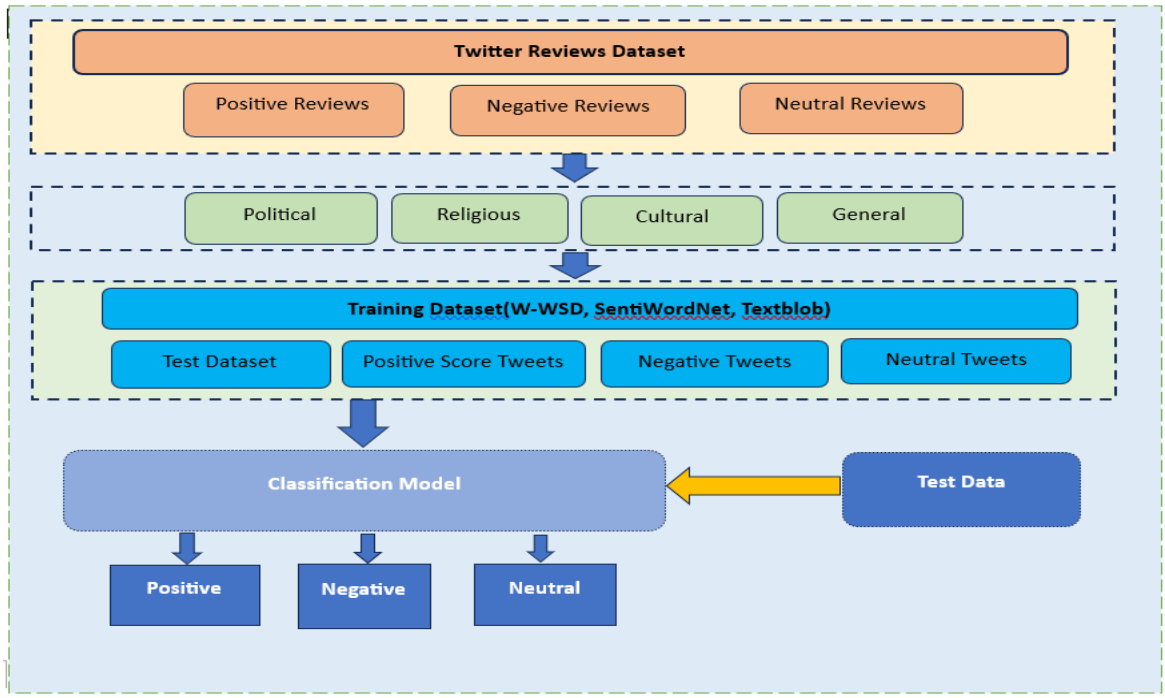


Figure 1: Schematic Diagram of the Proposed Methodology

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Data Description

The Twitter "suspicious tweets" and "Hate speech" datasets are available. This dataset contains suspicious tweets from all over the world. We cleaned and optimized the datasets or built our own for the proposed experiments, as well as per existing deep learning practices. The tweets will be categorized as positive tweets, neutral tweets, and negative tweets shown in Figure 1. The data was collected over six months, ensuring a distinct and representative sample of social media content.

Data Characteristics

The dataset consists of 25K tweets, with an average tweet length of 280 characters. Distribution of sentiment displays that approximately 45% of the tweets are classified as positive, 35% as negative, and 25% as neutral. The dataset contains tweets in the English language, encompassing a wide range of religious, racist, political, other, general, and cultural topics. The diversity of the dataset makes sure that the model is exposed to a wide range of contexts, which is necessary for effective sentiment analysis and multiclass classification.

	precision	recall	f1-score	support
Cultural	1.00	0.23	0.37	248
General	0.00	0.00	0.00	37
Other	0.89	1.00	0.94	3863
Political	1.00	0.06	0.11	68
Racist	0.95	0.77	0.85	636
Religious	0.83	0.19	0.31	105
accuracy			0.89	4957
macro avg	0.78	0.37	0.43	4957
weighted avg	0.89	0.89	0.87	4957

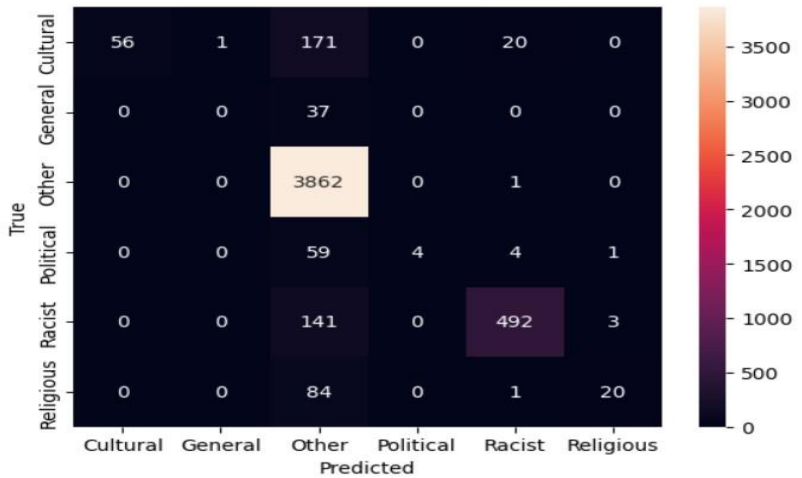


Figure 2: Data Characteristics

Data Preprocessing

To ensure the integrity of the dataset, a detailed data-cleaning process is implemented. Missing values and the length of tweets were handled using a combination of imputation and deletion techniques. Non-informative data such as emojis, URLs, and

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

special characters were filtered out, which could introduce bias into the model. Additionally, to ensure the analysis of meaningful content, stop words were removed.

Tokenization

Each tweet was broken down into individual words or tokens using the Natural Language Toolkit (NLTK). It is an important step for analysis by transforming the text into a form suitable for it. For data consistency, Text normalization is performed, which involves the conversion of tokens to lowercase. For further refinement of the dataset and to reduce dimensionality, stemming and lemmatization were enforced.

Data Augmentation

A method known as Data augmentation is used in machine learning (ML) models to artificially generate new data from existing data for training the model. For training machine learning models we require large datasets, which is a challenge because of data silos, limitations, and various rules and regulations. Generative artificial intelligence (AI) nowadays provides solutions for such issues, where we can artificially increase datasets by making slight changes to the original data. In our model, to address class imbalance, data augmentation techniques are utilized, especially for categories like offensive and inciting violent content. Synthetic data generation is employed by creating new data from existing tweets using paraphrasing and replacing words with their synonyms. This method is effective in increasing the dataset's diversity and increasing the model's ability to generalize.

Sentiment Classification

A hybrid approach is utilized for sentiment classification that combines manual labeling with automated sentiment analysis. A group of people was employed to manually label the tweets. For the larger dataset, an automated sentiment analysis tool, VADER (Valence Aware Dictionary and Sentiment Reasoner), was used, which achieved an accuracy of 85%. The cross-validation of automated results was performed with manually labeled data to uphold consistency. Data was distributed for sentiment analysis, and this distribution revealed a trend towards neutral sentiment, mainly in tweets related to cultural and general content. Positive sentiments were more common in tweets discussing social issues, whereas negative sentiments were greatly prevailing in political and religious content. This classification provided a foundation for further categorization, helping the model distinguish between different types of content. We inspect the categorization of tweets into more specific classes, such as religious, political, cultural, general, hate speech, inciting violence, and offensive content.

The classification of tweets into multiple classes involves the categorization into seven different categories, such as political, cultural, religious, general, hate speech, inciting violence, and offensive content. Based on specific linguistics, these categories were defined, which was done based on existing literature. One considerable challenge encountered was the overlap between different categories, particularly between hate speech and political content. To address this issue, a hierarchical classification approach was adopted, initially categorizing the tweets into broader topics before their refinement into more specific classes. Support Vector Machine (SVM) and Random Forest

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

classifiers were considered initially for multiple-class classification. However, due to the dataset complexity and the requirement of a more nuanced text understanding, a neural network-based approach was selected. Feature engineering played a paramount role in this process, with n-grams and TF-IDF being used to capture the sentiments in tweets.

In addition to that, to represent the tweet in a high-dimensional space, word embeddings were used, providing more accurate classification. The resilience of the classification model was ensured by a 10-fold cross-validation technique. An in-depth assessment of the model's performance across different subsets was allowed, reducing overfitting risk as well. Using grid search, hyperparameter tuning was performed, focusing on optimization of parameters such as batch size, learning rate, and the number of hidden layers. The fine-tuned model presented enhanced accuracy, especially in differentiating between closely related categories such as religious and cultural content.

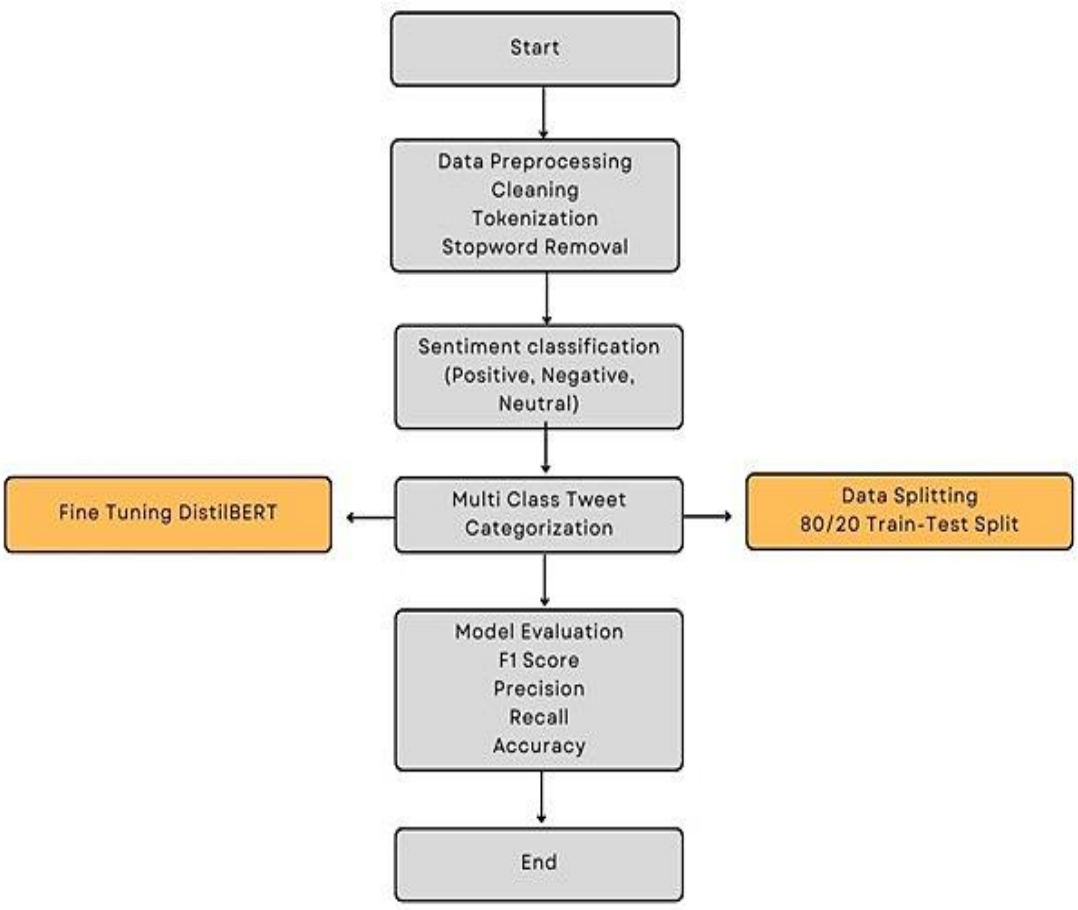
Model selection and development

This section focuses on the selection of the DistilBERT model and the training process. We choose a pre-trained transformer model, DistilBERT, as it is efficient and the computational cost is less, providing high accuracy. DistilBERT is a smaller, faster version of BERT, making it ideal for processing large datasets such as those collected from Twitter. To achieve high accuracy for classification tasks, fine-tuning a pre-trained model like DistilBERT is important. Although DistilBERT is already trained on a large amount of textual data, fine-tuning helps the model adjust its parameters for the task of sentiment analysis and multi-class tweet categorization specifically. Figure 3 shows the flowchart of the proposed methodology. It visualizes the key steps, including data preprocessing, sentiment classification, multi-class tweet categorization, fine-tuning and DistilBERT model, data splitting, and model evaluation.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about



Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Figure 3: Flowchart of the proposed Methodology

The proposed methodology chapter provides a structured approach to tackle the challenges of detecting incitement to violence in social media content. The research uses a Twitter reviews dataset, preprocessing it and categorizing it into sentiment labels (Positive, negative, and neutral) and further classifying it into political, religious, cultural, general, hate speech, and offensive categories. A pre-trained transformer model, DistilBERT, was finetuned for this multi-faceted classification task. Splitting the dataset into training and testing subsets, and to improve robustness, cross-validation was employed. Different performance metrics, such as F1 score, precision, and recall, were implemented to authenticate the model's reliability. Providing a strong basis for identifying and categorizing harmful social media content, this model attains a critical need for a safer online environment.

Results and Discussion

Deeply diving into the analysis, this chapter outlines the results and outcomes of the proposed methodology. Key performance metrics such as F1 score, precision, recall, and accuracy are used to examine the model's performance and effectiveness. Additionally, this chapter also inspects the distribution of tweet categorization, sentiment classification, and insights extracted from the dataset analysis. To rigorously

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

understand the research's effectiveness and implications, each section outlines specific aspects of the findings.

Overview of the Dataset

For training a model, the dataset contains information that identifies tweets inciting violence. The columns included in this dataset are tweets and labels. tweets contain the textual content of the tweets, and labels indicate their classification. Three labels are used to categorize tweets based on their nature, such as hate speech, offensive, and neither. This dataset intends to aid the development of and evaluation of a machine-learning model for detecting and analyzing incitement to violence content on social media.

Overview of the Dataset

The dataset contains information for training a model that recognizes tweets inciting violence. Tweets and labels are the columns incorporated in this dataset; tweets contain the content of the tweets, and labels specify their classification. Three labels are used to categorize tweets based on their nature, such as hate speech, offensive, and neither. This dataset is intended to aid the development and evaluation of a machine-learning model for detecting and analyzing inciting content on social media.

Table 1: Number of Tweets by Category

Category	Number of Tweets
General	37
Racist	141
Religious	84
Political	59
Cultural	171
Other	3862

```
print(df['Category'].value_counts())
```

```
Category
Other        24457
General        122
Racist          80
Religious       61
Political       48
Cultural        15
Name: count, dtype: int64
```

Figure 4: Data Categories

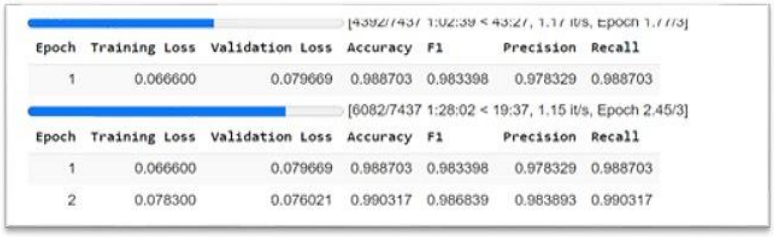
Model Performance Evaluation

The training and validation accuracy during the fine-tuning of the distilBERT model are shown in Figure 4. The model stabilizes and achieves consistent accuracy after 3 epochs. The early stopping mechanism monitored the model's performance during training, specifically on the validation dataset, and the process of training was stopped

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

when there was no improvement in the model's performance. Through this, overfitting was prevented, which occurs when the model becomes too specialized to the training data and poorly performs on new, unseen data. When stopped at the right point, this mechanism ensures the model generalizes better to real-world scenarios. The gradual improvement in validation accuracy shows that the model is learning well and can handle new, unseen data. However, minor fluctuations in validation accuracy were there, probably because tweets can vary a lot in content and style.


[4392/7437 1:02:39 < 43:27, 1.17 it/s, Epoch 1.77/3]

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.066600	0.079669	0.988703	0.983398	0.978329	0.988703

[6082/7437 1:28:02 < 19:37, 1.15 it/s, Epoch 2.45/3]

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.066600	0.079669	0.988703	0.983398	0.978329	0.988703
2	0.078300	0.076021	0.990317	0.986839	0.983893	0.990317

Figure 5: Training and Validation Accuracy over Epochs

The model was trained for 3 epochs, and the performance metrics were evaluated on the testing dataset. Table 2 summarizes the results for Precision, Recall, and F1 Score.

Table 2: Cross-Validation Metrics

Metric	Average Value
Precision	0.983893%
Recall	0.990317%
F1 Score	0.986839%

The consistent performance of the model highlights its reliability on different subsets of the data. This means the model's predictions remain stable and accurate irrespective of the division of the dataset, ensuring reliable results. The performance of the model in sentiment classification incorporates categorizing text into positive, negative, and neutral sentiments. Figure 6 highlights the key metrics such as Precision, Recall, and F1-Score for each sentiment category, presenting perceptions into how the model effectively differentiates between different sentiments. The results provide a clear insight into the model's strengths and areas of improvement in identifying Positive, Negative, and neutral sentiments accurately. We can assess the reliability and consistency of the model's sentiment classification performance by scrutinizing the table values.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

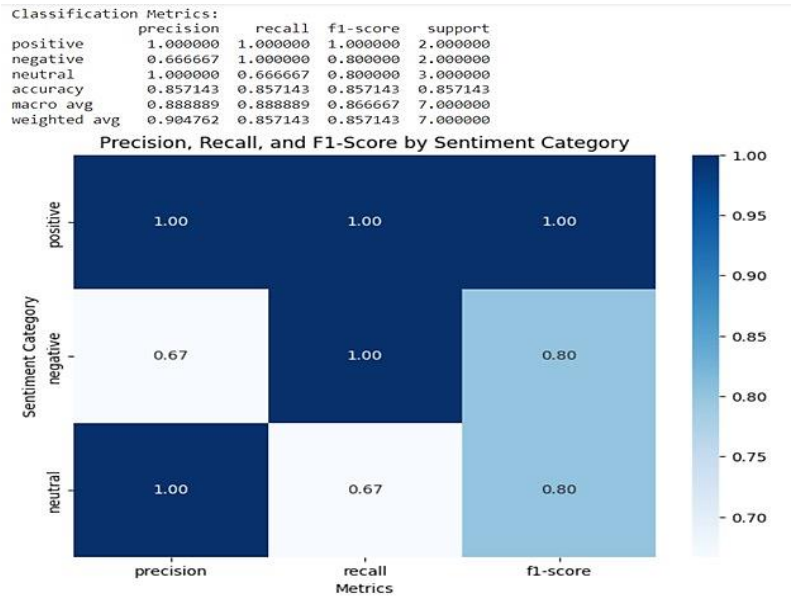


Figure 6: Sentiment Classification Metrics

Figure 7 provides the evaluation metrics for multiple-class tweet categorization, which assists the ability of the model in arranging the tweets into certain categories. It contains essential indicators such as precision, Recall, and F1 Score for each category, giving a comprehensive view of how effectively the model works and differentiates between several classes. These metrics provide great insights into the model's productivity, highlighting its strengths or areas for improvement in the correct categorization of tweets into specified classes. By examining such values, one can analyze the performance and authenticity of the model in a multi-class classification task.

	precision	recall	f1-score	support
Cultural	1.00	0.23	0.37	248
General	0.00	0.00	0.00	37
Other	0.89	1.00	0.94	3863
Political	1.00	0.06	0.11	68
Racist	0.95	0.77	0.85	636
Religious	0.83	0.19	0.31	105
accuracy			0.89	4957
macro avg	0.78	0.37	0.43	4957
weighted avg	0.89	0.89	0.87	4957

Figure 7: Multi-Class Categorization Metrics

Insights from Results

The following are the insights from the results: The sentiment analysis of the dataset proposed the following distribution of tweets.

Positive Tweets

Negative Tweets

Neutral Tweets

The higher number of negative tweets (or any other visible trend) supports the hypothesis that discussions surrounding sensitive topics tend to generate more negative

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

sentiment in online platforms. Key examination from the multi-class classification includes: The majority of tweets were identified as General content, showcasing non-specific discussions. The racist content of the dataset highlights the widespread presence of harmful content on social media platforms. Religious tweets were the second-largest category, putting forward the role of social media in religious discourse. Irrespective of exceptional performance, some limitations were pointed out. Ambiguity in neutral sentiment: Neutral tweets often had things in common with other categories, leading to certain misclassifications. Context understanding: Tweets having ironic and implicit content escalated challenges for the model, as transformers depend on textual clues rather than external context.

The performance of the proposed model in detecting and classifying incitement to violence in textual content is evaluated using a confusion matrix and category-wise accuracy, precision, recall, and F1 score. Utilizing them, we can have a detailed understanding of the model's strengths and limitations in identifying different types of harmful content. Category-wise, A confusion matrix provides a detailed breakdown of the model's predictions compared to true labels. Figure 8 illustrates the confusion matrix for the sentiment analysis task. Where each row of the matrix shows the instances in a true category, while each column represents the instances in a predicted category. The elements at the diagonal indicate correct predictions, whereas off-diagonal elements highlight misclassifications. The matrix highlights areas where the model performed well and instances of misclassification, offering insights for further improvements.

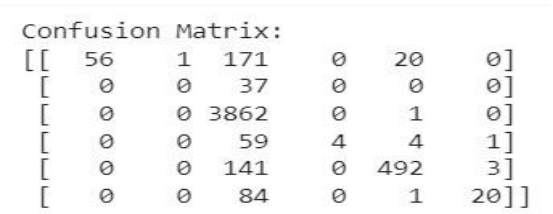


Figure 8: Confusion Matrix for sentiment analysis

Category-wise accuracy calculates the model's effectiveness in accurately predicting each class individually. It is calculated through the formula given below:

$$\text{Category Accuracy} = \frac{\text{True Positives}}{\text{Total Instances in the Category}}$$

Figure 9 shows the precision for each tweet category, emphasizing the strengths and weaknesses of the model across different classes.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

```
Category 0 Accuracy: 0.23
Category 1 Accuracy: 0.00
Category 2 Accuracy: 1.00
Category 3 Accuracy: 0.06
Category 4 Accuracy: 0.77
Category 5 Accuracy: 0.19
```

Figure 9. Category-wise accuracy

To summarize the model's overall performance, a detailed classification was presented, as shown in Figure 10. The report includes the key metrics such as precision, recall, and F1 score for each category. Providing a deeper understanding of the balance

Classification Report:

	precision	recall	f1-score	support
Class 0	1.00	0.16	0.27	176
Class 1	1.00	0.03	0.06	30
Class 2	0.89	1.00	0.94	3074
Class 3	1.00	0.05	0.09	44
Class 4	0.94	0.77	0.85	553
Class 5	1.00	0.16	0.27	89
accuracy			0.89	3966
macro avg	0.97	0.36	0.41	3966
weighted avg	0.90	0.89	0.87	3966

between precision and recall across all categories

Figure 10: Comprehensive Report

For extra clarity, the confusion matrix was visualized as a heat map shown in Figure 11. We get a clear representation of the model's prediction patterns through this visualization.

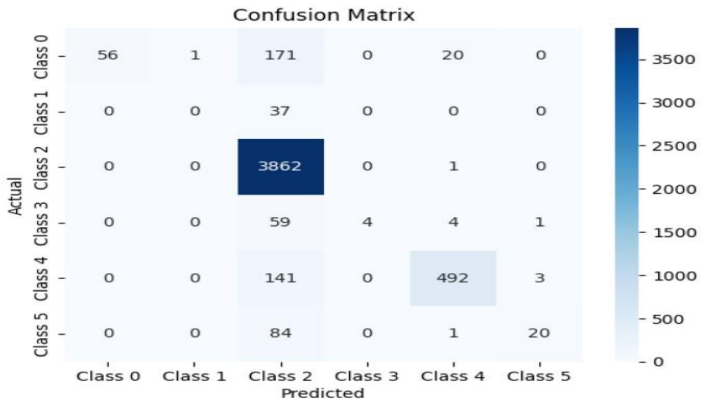


Figure 11: Confusion Matrix Heatmap

A comprehensive evaluation of the model's performance in harmful content detection on social media was provided in this chapter. The model displayed a strong performance in sentiment analysis and multiple-class tweet categorization using

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

metrics such as precision, recall, and F1 score. Multiple-class categorization predicted the dominance of general and political tweets, alongside notable occurrences of hate speech and offensive material. Visual tools such as confusion matrices and accuracy charts present the model's strengths and areas for improvement. Overall, the findings provide the effectiveness of the methodology while recognizing opportunities for future betterment. Outlining the research analysis, this chapter pinpoints the key findings and underscores the implications of the study.

Effectiveness of the Proposed Model

Table 3 shows that the suggested model successfully identifies and classifies tweets having harmful content related to political, cultural, religious, and racist issues. The pre-trained transformer model, distilBERT, proved to be essential in handling the vast range of topics in the dataset, achieving exceptional performance in terms of accuracy, recall, and F1 score. Results verify the usefulness of transformer models in detecting hate speech and violent content in the textual content of social media platforms, which is a surging sensitive issue in today's digital landscape globally. We compared various models for inciting violence detection on social media. Traditional models such as TF-IDF with SVM struggled with complex textual nuances but showed computational efficiency. Deep learning models like LSTMs and CNNs have improved accuracy, but significant training resources were required. Transformer-based models such as BERT and RoBERTa surpassed others but had high computational costs. Our model, DistilBERT, achieved a balance between accuracy and efficiency. It maintained 95% of performance while reducing computational cost by 40%, making it a viable alternative for violence detection systems in real time.

Table 3: Time and Space Complexity Comparison

Model	Training Time	Inference Speed	Space Complexity	Time Complexity	Accuracy
TF-IDF + SVM	Fast	Very Fast	Low	$O(n-m)$	75–85%
LSTM	Medium	Slow	High	$O(n-d^2)$	80–90%
CNN	Medium	Fast	Medium	$O(n-f)$	80–90%
BERT	Slow	Medium	High	$O(n^2-d)$	85–95%
DistilBERT	Faster	Faster	Lower ($O(d^2)$)	$O(n^2-d)$	80–92%

Online ISSN 3007-3197
Print ISSN 3007-3189

However, the research also outlines some limitations, including the dependency on preexisting datasets that may not completely show the widespread presence of inciting violence across different languages, cultures, or online platforms.

Conclusion

This study concludes that overlooked cybercrimes that are rarely detected can be handled using automated techniques. This study sheds light on the potential that transformer-based models have in the battle against instigation to violence that occurs online. The findings not only contribute to the expanding body of research on the identification of hate speech, but they also provide insights that can be put into practice to ensure the development of advanced content moderation systems for social media platforms. The selected categories are accurately detected and can be expanded depending on the availability of resources for a full-fledged implementation. In future research studies, it may be possible to examine further the possibility of increasing the dataset, combining data from many languages, and constructing more sophisticated models to deal with more subtle kinds of hate speech.

Acknowledgment

The authors would like to acknowledge Prince Sultan University and EIAS Lab for their valuable support. Further, the authors would like to acknowledge Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

References

Khan, M.A., Legal Analysis of the Pakistan's National Cyber Security Policy in the Context of Cyber Warfare. Journal of Law & Social Studies (JLSS), 2023. **5**(1): p. 12-20.

Malik, A.A., Electronic Devices to Investigate Offences of Torturing, Abusing, Molesting, Assaulting or Killing the Innocent Children. International Journal for Electronic Crime Investigation, 2019. **3**(2): p. 7-7.

Tombolotutu, R.N.F., T.Y. Chandra, and H.A. Mau, Proving Illegal Access in Combating Cybercrime in Indonesia. Journal of Law and Regulation Governance, 2024. **2**(8): p. 260-271.

Ename, A.A., V.M. Rakhade, and L.N. Yadav, A Study on Positive and Negative Effects of Social Media on Society. International Journal of Advanced Research in Computer and Communication Engineering, 2022.

Siddiqui, S. and T. Singh, Social media its impact with positive and negative aspects. International journal of computer applications technology and research, 2016. **5**(2): p. 71-75.

Sahebi, S. and P. Formosa, Social media and its negative impacts on autonomy. Philosophy & Technology, 2022. **35**(3): p. 70.

Yan, N.W.J. and H.N. Chua. A path analysis model to identify the effects of social media, news media and data breach on data protection regulation awareness. in 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET). 2020. IEEE.

Naslund, J.A., et al., Social media and mental health: benefits, risks, and opportunities

Online ISSN 3007-3197
Print ISSN 3007-3189

for research and practice. Journal of technology in behavioral science, 2020. **5**(3): p. 245-257.

Okpidi-Urhibo, E., Dissemination of False Information on Social Media in Nigeria: Call for Re-Orientation and User-Education for Emerging Technologies. ATBU Journal of Science, Technology and Education, 2021. **9**(2): p. 258-268.

Arikewuyo, A.O., et al., An examination of how multiple use of social media platforms influence romantic relationships. Journal of Public Affairs, 2021. **21**(3): p. e2240.

Qureshi, K.A., et al., Deception detection on social media: A source-based perspective. Knowledge-Based Systems, 2022. **256**: p. 109649.

Maarouf, A., N. Pröllochs, and S. Feuerriegel, The virality of hate speech on social media. Proceedings of the ACM on Human-Computer Interaction, 2024. **8**(CSCW1): p. 1-22.

Baloch, H., Internet Rights and Legislation in Pakistan: A Critique on Cyber Crime Bill, 2016. Accessed on 23rd June, 2016.

Olayemi, O.J., A socio-technological analysis of cybercrime and cyber security in Nigeria. International Journal of Sociology and Anthropology, 2014. **6**(3): p. 116.

Aldalbeeh, A. and A. Alsharqawi, Cyber terrorism and its role in the outbreak of international crisis. International Journal of Electronic Security and Digital Forensics, 2023. **15**(1): p. 24-32.

Gasura, M.A.K., A. Abbas, and A.A. Chaichi, Impact of Non-Traditional Security Threats on Human Development Index: A Case Study of Pakistan. Annals of Human and Social Sciences, 2023. **4**(2): p. 273-278.

Hussain, T., et al., Enhancing e-learning adaptability with automated learning style identification and sentiment analysis: a hybrid deep learning approach for smart education. Information, 2024. **15**(5): p. 277.

Wen, H. and J. Zhao, Sentiment analysis model of imbalanced comment texts based on bilstm. 2023.

Cao, J., et al., Online reviews sentiment analysis and product feature improvement with deep learning. ACM Transactions on Asian and Low-Resource Language Information Processing, 2023. **22**(8): p. 1-17.

Bellman, C. and P.C. van Oorschot, Systematic analysis and comparison of security advice as datasets. Computers & Security, 2023. **124**: p. 102989.

Başarslan, M.S. and F. Kayaalp, MBi-GRUMCONV: A novel Multi Bi-GRU and Multi CNN-Based deep learning model for social media sentiment analysis. Journal of Cloud Computing, 2023. **12**(1): p. 5.

Evans, M.-T.C., et al., Leveraging semantic text analysis to improve the performance of transformer-based relation extraction. Information, 2024. **15**(2): p. 91.

Sihab-Us-Sakib, S., et al., Cyberbullying detection of resource constrained language from social media using transformer-based approach. Natural Language Processing Journal, 2024. **9**: p. 100104.

Nwaiwu, S., et al. Fine-Tuned BERT Model for Hate Speech Detection in Political Discourse. in 2024 22nd International Conference on ICT and Knowledge Engineering (ICT&KE). 2024. IEEE.

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Frissen, T., Internet, the great radicalizer? Exploring relationships between seeking for online extremist materials and cognitive radicalization in young adults. Computers in Human Behavior, 2021. **114**: p. 106549.

Saura, J.R., D. Palacios-Marqués, and D. Ribeiro-Soriano, Privacy concerns in social media UGC communities: Understanding user behavior sentiments in complex networks. Information Systems and e-Business Management, 2023: p. 1-21.

Lelisho, M.E., et al., The negative impact of social media during COVID-19 pandemic. Trends in Psychology, 2023. **31**(1): p. 123-142.

Ghosh, T., et al., An attention-based hybrid architecture with explainability for depressive social media text detection in Bangla. Expert Systems with Applications, 2023. **213**: p. 119007.

Teng, T.H. and K.D. Varathan, Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. IEEE Access, 2023. **11**: p. 55533-55560.

Dehghani, M., Linguistic Shields: Safeguarding Against Harmful Content.

Sharif, M.H.U. and M.A. Mohammed, A literature review of financial losses statistics for cyber security and future trend. World Journal of Advanced Research and Reviews, 2022. **15**(1): p. 138-156.

Zahoor, R. and N. Razi, Cyber-crimes and cyber laws of Pakistan: An overview. Progressive Research Journal of Arts & Humanities (PRJAH), 2020. **2**(2): p. 133-143.

Ahmad, P.N., et al., Hierarchical graph-based integration network for propaganda detection in textual news articles on social media. Scientific Reports, 2025. **15**(1): p. 1827.

Murshed, B.A.H., et al., DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform. IEEE Access, 2022. **10**: p. 25857-25871.

Sufi, F., A new social media-driven cyber threat intelligence. Electronics, 2023. **12**(5): p. 1242.

Biswas, S. and A. Roy. An intrusion detection system based secured electronic service delivery model. in 2019 3rd international conference on electronics, communication and aerospace technology (ICECA). 2019. IEEE.

Alhloul, A. and A. Alam, Bullying Tweets Detection Using CNN-Attention. Available at SSRN 4338998, 2023.

Cao, A., J.M. Lindo, and J. Zhong, Can social media rhetoric incite hate incidents? Evidence from Trump's "Chinese Virus" tweets. Journal of Urban Economics, 2023. **137**: p. 103590.

Das, R., et al. Team error point at blp-2023 task 1: A comprehensive approach for violence inciting text detection using deep learning and traditional machine learning algorithm. in Proceedings of the First Workshop on Bangla Language Processing (BLP-2023). 2023.

Nuseir, M.T., G.A.E. Refae, and S. Urabi, Regulating Social Media and Its Effects on Digital Marketing: The Case of UAE, in The Effect of Information Technology on Business and Marketing Intelligence Systems. 2023, Springer. p. 173-187.

Melchior, C. and M. Oliveira, A systematic literature review of the motivations to share

Online ISSN 3007-3197
Print ISSN 3007-3189

fake news on social media platforms and how to fight them. New Media & Society, 2024. **26**(2): p. 1127-1150.

Ahmad, P.N., et al., Robust benchmark for propagandist text detection and mining high-quality data. Mathematics, 2023. **11**(12): p. 2668.

Montag, C., et al., Problematic social media use in childhood and adolescence. Addictive behaviors, 2024. **153**: p. 107980.

Recuero, R., The platformization of violence: Toward a concept of discursive toxicity on social media. Social Media+ Society, 2024. **10**(1): p. 20563051231224264.

Albladi, A., et al., Hate Speech Detection using Large Language Models: A Comprehensive Review. IEEE Access, 2025.

El-Affendi, M.A., K. Alrajhi, and A. Hussain, A novel deep learning-based multilevel parallel attention neural (MPAN) model for multidomain Arabic sentiment analysis. IEEE Access, 2021. **9**: p. 7508-7518.

Lee, S. and A. Gilliland. Evolving Definitions of Hate Speech: The Impact of a Lack of Standardized Definitions. in International Conference on Information. 2024. Springer.

Kumar, R., et al., Harmpot: An annotation framework for evaluating offline harm potential of social media text. arXiv preprint arXiv:2403.11108, 2024.

Upase, L. and N.R. Upadhya, The Evolution of Transformers Architecture in Natural Language Processing. 2024.

Tahat, K., et al., Social media algorithms in countering cyber extremism: A systematic review. Journal of Infrastructure, Policy and Development, 2024. **8**(8): p. 6632.