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

An In-Depth Review of Anomaly Detection Techniques in Social Networks Using Machine Learning Techniques

Sarfaraz Alam[1], Mohammad Faisal[2]

Submitted: 05/08/2024 **Revised:** 18/09/2024 **Accepted:** 26/09/2024

Abstract: The proliferation of social media platforms has facilitated the emergence of harmful online phenomena, including hate speech, cyberbullying, and automated bot activity, thereby undermining the safety of digital ecosystems. In response, researchers are increasingly leveraging deep learning and machine learning methodologies to develop automated detection and mitigation systems. This review synthesizes recent advancements across several key domains: sentiment analysis, cyberbullying prevention, hate speech identification, and social bot detection, with a particular focus on the evolution of ML/DL architectures such as convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and graph convolutional networks (GCNs). Furthermore, it critically examines persistent challenges in cyberbullying intervention systems, underscoring the necessity of integrating psychological and socio-cultural insights. The discussion extends to potential strategies for enhancing personal agency through improved support mechanisms and digital literacy education. Overall, the analysis substantiates the superior efficacy of deep learning approaches compared to traditional machine learning techniques, with the ultimate objective of informing the creation of scalable solutions to counteract detrimental online behaviors.

Keywords: *Social Network Sites (SNS), NLP, RNNs, Long Short-Term Memory (LSTM), GPT, BERT, CNNs*

1. Introduction:

The rise of internet usage, social media, and digital technology has led to a vast amount of data from various sources, making big data analysis challenging. Techniques like text mining, web mining, data mining, and machine learning are used to process large datasets for informed decision-making. However, the complexity of pattern recognition on microblogging platforms like Facebook, Instagram, Snapchat, YouTube, Twitter, and WhatsApp makes this task particularly challenging [2]. Fig.1 shows how social media users' opinions serve various purposes, including product analysis, political predictions, news distribution, marketing, crime prediction, and terrorist tracking. Natural Language Processing (NLP) technology enhances user preferences and public sentiment analysis from open discussion platforms.

[5]. Social media and online discussion forums have transformed communication, providing platforms for open dialogue and information exchange. However, they have also become channels for misinformation, hate speech, and cyberbullying, posing serious risks to society. Effectively identifying and mitigating harmful content is essential for fostering a respectful digital space. Modha et al. [6] investigated hate speech on Instagram between 2020 and 2021, categorizing online content as either aggressive or non-aggressive. Meanwhile, Kaur et al. [7] proposed a multi-faceted approach to detecting harmful content, incorporating activity-based, user-based, context-based, and network-based factors. Platforms such as Twitter facilitate discussions on trending topics through various forms of user engagement [8]. Unfortunately, bad actors exploit these platforms by deploying automated accounts—often referred to as social bots or sybil accounts—to manipulate public opinion, spread rumors, disseminate false information, promote dangerous products, defame individuals, fabricate fake followers, and engage in social phishing and spamming. These accounts also

[1,2] *Department of Computer Application, Integral University, Lucknow, India.*

[1] *sarfaraz@iul.ac.in,* [2] *mdfaisal@iul.ac.in*

participate in cyberattacks through tactics like profile cloning and coordinated manipulation [7][8]. This study takes a micro-sociological approach to cyberbullying, emphasizing the psychological dynamics that influence such behaviors and highlighting the importance of understanding individual agency in online interactions. The exponential growth of social media has introduced significant security challenges, including fraudulent accounts, financial scams, misinformation, propaganda, spam bots, cyberattacks, fake profiles of public figures and organizations, and the widespread dissemination of false information—particularly during crises such as the COVID-19 pandemic.

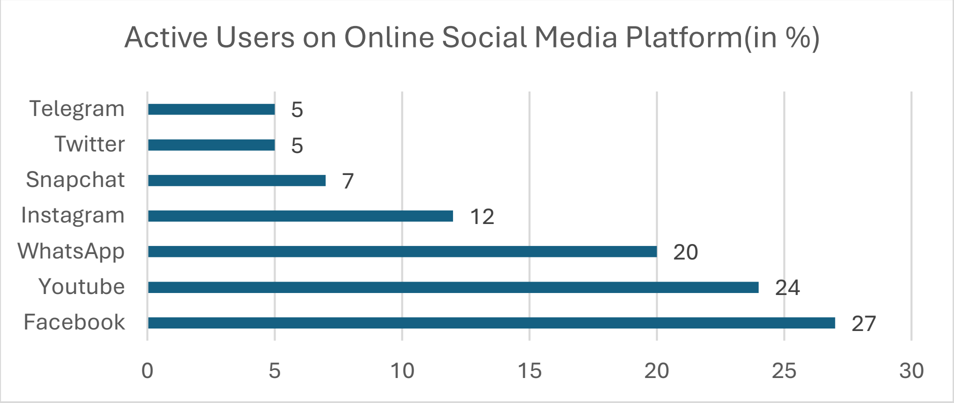

Fig. 1: Active Users on Online Social Media Platform

2. Types of Anomalies on Social Networking sites:

Unusual or suspicious acts that diverge from typical user behavior are referred to as anomalies on social networking sites (SNS). These irregularities may be a sign of malevolent activity, disinformation, privacy violations, or security issues. The primary categories of abnormalities on SNS are as follows:

2.1 Hate Speech:

2.1 Hate Speech: Hate speech is defined as any text, image, audio, or video that denigrates, threatens, or discriminates against individuals or groups based on traits such as race or ethnicity, religion, gender or gender identity, sexual orientation, disability, nationality, or political affiliation.

2.1.1 Types of Hate Speech

- **Direct Attacks:** Overt taunts, threats, or derogatory remarks directed at a specific person or group.
- **Incitement to Violence:** Promoting physical violence against a group (e.g., threats of lynching, calls for extermination).
- **Hate Symbols & Memes:** Spreading hate via memes, coded language, or imagery (such as racist caricatures or Nazi insignia).
- **Dog Whistles & Coded Language:** Spreading hate while evading notice by using subtle or oblique allusions.

2.2 Cyberbullying: Cyberbullying is the practice of harassing, intimidating, embarrassing, or threatening others via digital platforms (such as social media, messaging applications, forums, or gaming groups). Cyberbullying, in contrast to conventional bullying, may happen around-the-clock, reach a larger audience, and frequently stay anonymous.

2.2.1 Types of Cyberbullying

- **Harassment:** Delivering disrespectful, threatening, or damaging texts on a regular basis.
- **Flaming:** posting something that is provocative, hostile, or inflammatory in an attempt to start arguments.
- **Trolling:** publishing offensive or provocative content with the intention of offending other people.
- **Impersonation (Catfishing):** fabricating profiles in order to mislead, control, or harm someone's reputation.

2.3 Fake Account: A profile on a social networking site that has been made using inaccurate or misleading information is called a fake account. These accounts may be utilized for a number of dishonest, malevolent, or fraudulent purposes**.**

2.3.1 Types of Fake Accounts

- **Bot Accounts:** Automated accounts designed to influence interaction by like, commenting, or sharing material.
- **Impersonation Accounts:** Fake accounts that pose as other people in order to trick others, such as politicians, celebrities, or everyday users.
- **Scam Accounts:** Designed to trick visitors with investment schemes, phishing, or phony freebies.
- **Catfish Accounts:** False identities are used in online connections to emotionally or financially mislead others.

3. Initial procedures for identifying Anomalies on Social Networking Sites

3.1 Data collection

Users may suffer as a result of the frequent use of damaging or abusive language on social media sites. The data format is crucial for precise analysis since the technique of data collection and analysis depends on the medium used to distribute the material. Public data from microblogging networks may be accessed using APIs offered by sites such as Sina-Weibo and Twitter. Facebook and Sina-Weibo use the Facebook Graph and Tencent APIs for streaming data, while Twitter provides a REST-API for obtaining static data. Additionally, articles and other pertinent material are gathered from the websites of these APIs.

3.2 Pre-processing of data

Natural Language Processing (NLP) involves data pre-processing, which includes normalization, tokenization, stop word removal, and text cleaning. Non-essential elements like links, punctuation, hashtags, and numeric characters are often removed. However, eliminating these may not always improve text clarity. Both data cleaning and pre-processing stages have been analyzed.

Tokenization is a method that breaks down text into individual words or phrases to understand its context, analyze word sequences, remove stop words, and normalize text by reducing variability and aligning it with a predefined standard.

3.3 Datasets to recognize hate speech

We investigated in detail the datasets used by the research community to build and evaluate their models.

The datasets pertaining to anomalies on SNS are detailed in Table 1.

References	Dataset	Dataset Description	Language
A. Rodriguez et al., 2022	Unstructured Information from Facebook Comments on social media	The remarks fall into one of three categories: positive, neutral, or negative.	English
H. Wu, et al., 2022	Comments on social media	The comments come into one of three categories: positive, neutral, or negative.	English
S. Khan, et al., 2022	Comments on social media like Twitter	Dataset 1 (Balanced): This dataset contains an equal distribution of four content categories—spam, normal, hateful, and abusive—ensuring that each category has a proportional representation. Dataset 2 (Unbalanced): In this dataset, the distribution of categories is uneven, with	English

		content classified into four groups—spam, normal, hostile, and abusive—where some categories may have significantly more instances than others.	
L.-L. Tao, et al., 2022	Internet reviews of travel websites	There is a total of 3486, 3846, 4549, 1960, and 2575 internet reviews for five distinct hotels, which were categorized into five review datasets.	English
R.M. Cruz, et al., 2022	Thomas Davidson, Zeerak Waseem	24,783 classified instances in three categories—"hate," "offensive," and "non-offensive"—make up the first dataset. In another, 16,907 incidents have been categorized as "Racism," "Sexism," and "Neither."	English
F.R. Nascimento, et al., 2022	Social Media comments	Information collected from Twitter over a number of months is included four datasets in the whole dataset.	English
M. Luo, et al., 2022	Data articles from Newspapers and News Channels	Data has been gathered from news outlets.	English
A. Kumar, et al., 2022	Laptop Dataset, Restaurant Dataset	1. Reviews − 1326 990 615 comments 2. Reviews − 4131 1535 927 comments	English
H.S. Alatawi, et al., 2021	Social media comment-Twitter	The three datasets used in this study are: the Twitter White Supremacy Dataset, the Stormfront Dataset, and the Balanced and Combined Datasets.	English
A. Zhao, et al., 2021	Social Media comments	The remarks fall into one of three categories: neutral, negative, or positive.	English
D.R. Beddiar, et al., 2021	Ask Fm	10,000 comments were used as a dataset to detect cyberbullying.	English
S. Alsafari, et al, 2021	Comments from Social media platform	hateful or offensive comments from social media	English
S. Kaur, et al., 2021	The dataset includes toxic comment classification, Wikipedia talk labels, Twitter text collection, internet argument dataset, hate speech dataset, and insult detection dataset.	Classified as Identity Hate, Obscene, Threatening, Toxic, Severe Toxic, Insulting, and Not Insulting.	English

M. Sharma, et al, 2021	Identification Dataset for Offensive Languages	This dataset, sourced from SemEval-2019 Task 6, categorizes social media foul language into neutral, offensive, and non-offensive categories.	English
J. Kocon, et al., 2021	Talk Labels on Wikipedia	There have been 100,000 comments, categorized into three groups: friendly, neutral, and hostile.	English

Table1: Dataset associated with the various Anomalies on SNS

4 An analysis of different Machine Learning Models for Anomaly Detection

Anomaly detection in machine learning can be classified into three categories based on how training samples are labeled: **supervised learning methods, semi-supervised learning methods, and unsupervised learning methods** [23].

4.1 Supervised learning

Recent supervised learning algorithms are used to categorize text for detecting abusive material. They examine various elements like comment content, user profiles, behavior, and social graph structure. The training data is labeled by experts or crowdsourcing systems, but the efficacy of these methods heavily relies on the amount of labeled data used.

4.2 Semi-Supervised learning

Semi-supervised machine learning methods are used to create models that combine data with and without labels. Xiang et al. proposed a method for detecting hate speech in Twitter corpus, reducing manual annotation effort. They used bootstrapping to identify unlabeled data and Latent Dirichlet Allocation (LDA) to identify offensive tweets by extracting variables from linguistic regularities in profane language.

4.3 Unsupervised learning

Algorithms for unsupervised learning determine potential clustering patterns in data. Consequently, data does not need to be labeled. It involves developing the ability to distinguish between input data that has been labeled and data that has not. An unconventional technique called Growing Hierarchical Self-Organizing Maps (SOMs) was put out by Capua et al. and is capable of efficiently clustering texts, including bully traces. [25,26] used a machine learning algorithm-based clustering technique to examine the sentiment of Twitter tweets.

4.4 Machine learning models for hate speech recognition

Social networks' anonymity encourages hate speech, which is a global issue, and allows people to cover up their illegal online activity. Given the increasing amount of social media data, identifying hate speech is essential since it may have detrimental effects on society [17]. The following discussion covers the latest machine learning methods for identifying hate speech.

4.4.1 Classical Machine Learning methods: Hate speech detectors use shallow detection methods, which categorize texts using shallow classifiers. These algorithms are trained on a tagged dataset, producing a model that distinguishes between hate and non-hate speech. TF-IDF and Ngrams are feature representation techniques. Hate speech and sentiment analysis are traditionally done using supervised machine learning techniques like Naive Bayes [27], Decision Tree, Support Vector Machine [29], Linear Regression [28], and Logistic Regression [30].

4.4.2 Ensemble approach: Ensemble techniques combine the strengths of multiple machine learning algorithms to improve performance. No model is flawless since every model has drawbacks. Collective

methods, such as bagging, Random Forest, and boosting [31], strive to reduce variance while enhancing learning capability. These techniques, along with other methods, contribute to statistical analysis.

4.4.3 Word-embeddings based methods: Word embedding is a technique that uses dispersed representations to learn vectorized representations, which are then used in text mining operations. Techniques like FastText, Glove, and word2vec [32] have been developed over time, providing representations for different classifiers.

4.4 Deep learning model for hate speech detection

Deep learning adds a multi-layer structure to neural networks, improving accuracy and performance. Unlike traditional machine learning methods that require explicit feature selection, deep learning models learn and extract information independently. Research in data mining and text classification has utilized deep learning algorithms to anticipate and classify hate speech messages [33]. An overview of deep learning models used for hate speech detection is provided.

4.4.1 Recurrent neural networks (RNNs): Recurrent Neural Networks (RNNs) are artificial neural networks that analyze sequential or time-series data, retaining past information for future predictions, making them useful for tasks like hate speech detection.

Long Short-Term Memory (LSTM), a Recurrent Neural Network, effectively detects anomalies on social networking sites by retaining context and capturing word dependencies. Its performance improves with attention mechanisms, highlighting hate speech-indicative text elements.

4.4.2 Convolution neural networks (CNNs): Convolutional Neural Networks (CNNs), commonly used in computer vision for image classification and object detection, have been adapted for Natural Language Processing applications like sentiment analysis and hate speech detection. CNNs use vector representation and one-dimensional convolution to detect local patterns.

4.5 Transformer based models

Transformer-based models like Bidirectional Encoder Representations from Transformers (BERT)[35], Generative Pre-trained Transformer (GPT), and Robustly Optimized BERT Pretraining Approach (RoBERTa)[32] are effective in Natural Language Processing tasks like sentiment analysis, machine translation, and question-answering. They understand context, use bidirectional learning, transfer learning, and attention mechanisms, making them effective in sentiment analysis and hate speech detection.

The general framework for detecting anomalies on SNS is depicted in Fig. 2.

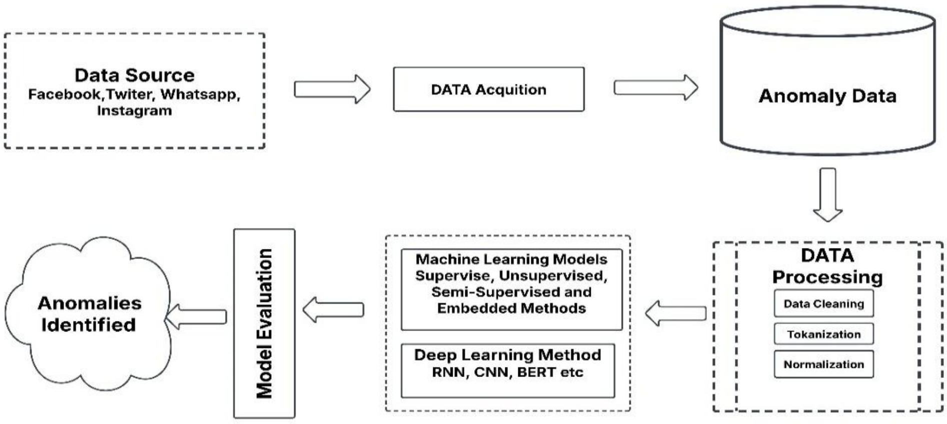
Data Source
Facebook,Twiter, Whatsapp, Instagram
DATA Acquition
Anomaly Data
DATA Processing
Data Cleaning
Tokanization
Normalization
Machine Learning Models Supervise, Unsupervised, Semi-Supervised and Embedded Methods
Deep Learning Method RNN, CNN, BERT etc
Model Evaluation
Anomalies Identified

Fig. 2: General Approach for Anomaly Detection

The current attempts to use the aforementioned techniques for anomalies identification are compiled in Table 2.

Reference	Dataset Size	Discription of Datasets	Models
Albadi, et al, 2018	11,874	Whether offensive or not, directed or undirected, towards an individual, group, or entity	NB BiLSTM BERT
Curry, et al, 2021	4185	Conversational AI is capable of detecting abuse by analyzing its explicit and implicit directedness, target groups, and severity across various forms of discrimination.	SVM, BERT, MLP, Random Fores
Caselli, T., et al, 2020	14,100	Offers tagged annotations for content that is rude and insulting.	BERT
Pamungkas, et al, 2020	1320	Twitter swear words are divided into classifications that are abusive and non-abusive.	linear support classifier (LSVC), LR, RF
Fanton, M., et al., 2021	5003	categorizes content according to gender, ethnicity, religion, sexual orientation, handicap, and country of origin into two groups: hateful and non-hateful.	GPT-2
Zampieri, M., et al., 2019	14100	The task structure is divided into branches A, B, C, which determine if content is offensive, target or untargeted, and categorize the target as an individual, group, or other.	SVM, BiLSTM, CNN
Qian, J., et al., 2019	33	Classifies content into two categories: hate or not.	SVM,CNN, RNN
G.L. De la Pena, 2022	14100	Classified as either offensive or not offensive.	LSTM with GNN
Mollas, I., et al.,2020	998	Binary as Hate/ Not.	LR, SVMs, RF and CNN
Salminen, J., et al, 2018	5143	Binary classification categorizes hate and non-hate content, while multinomial classification groups content into 21 groups	Logistic Regression, Decision Tree, Random Forest

Table 2 Review of the models proposed for anomalies on SNS

5 Findings and discussion

This survey reviews recent advancements in detection of various anomalies on SNS, analyzing publicly available datasets. Four key challenges identified include noisy and imbalanced data, highly skewed distributions in multi-label and multi-class settings, sparse feature vector representation in machine learning and deep learning models, and the noisy nature of the data. The analysis emphasizes the importance of expert annotation in dataset preparation, particularly in identifying hate speech. It also highlights the skewed nature of online hate speech data, highlighting the need for further research to develop effective methods for annotating new posts and online communications, distinguishing between toxic/abusive language, hate speech, and offensive language.

Next, we gave a summary of the various models and datasets utilized in sentiment analysis and hate speech identification. According to our analysis of the papers, certain methods label offending texts as hate speech. Despite the introduction of deep learning models, Fig. 3 shows that machine learning models continue to be more prevalent in research investigations.

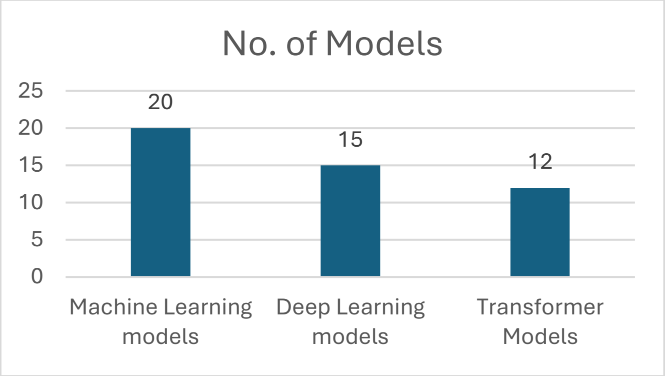

Fig 3: Number of Models for Anomalies Detection on SNS

6 Challenges and issues of Anomaly detection on SNS

Anomaly detection in social networks is complex due to the dynamic nature of these networks, high dimensionality of data, privacy concerns, adversarial attacks, noisy networks, and incomplete data. Defining anomalies is challenging due to evolving user behavior, privacy concerns, fake profiles, spam, and lack of labeled datasets and standard evaluation metrics. Scalability and real-time processing are critical challenges, and black-box AI models lack interpretability, making benchmarking and improving detection models a significant issue.

7 Conclusion

This article provides an overview of current methods in anomalies detection, analyzing various methodologies and datasets. Conventional machine learning approaches like SVM, Decision Tree, NB, and LR were analyzed, while deep learning models like CNNs, RNNs, and Transformer-based architectures like BERT and GPT have shown remarkable abilities in detecting anomalies like hate speech, fake account detection and cyberbullying in online content.

8 Acknowledgement

This work is acknowledged under Integral University manuscript No IU/R&D/2023-MCN0002417.

9 References

[1] Hande, R. Priyadharshini, B.R. Chakravarthi, Kan CMD: Kannada Code Mixed dataset for sentiment analysis and offensive language detection, in: Proceedings of the Third Workshop on Computational Modeling of People’s Opinions, Personality, and Emotion’s in Social Media. 2020.

[2] J. Cao, et al., A risky large group emergency decision-making method based on topic sentiment analysis, Expert Systems with Applications 195 (2022), 116527.

[3] P.K. Roy, et al., A framework for hate speech detection using deep convolutional neural network, IEEE Access 8 (2020) 204951–204962.

[4] H. Liu, et al., A fuzzy approach to text classification with two-stage training for ambiguous instances, IEEE Transactions on Computational Social Systems 6 (2) (2019) 227–240.

[5] F.M. Plaza-Del-Arco, et al., A multi-task learning approach to hate speech detection leveraging sentiment analysis, IEEE Access 9 (2021) 112478–112489.

[6] S. Modha, et al., Detecting and visualizing hate speech in social media: A cyber watchdog for

surveillance, Expert Systems with Applications 161 (2020), 113725.

[7] S. Kaur, S. Singh, S. Kaushal, Abusive content detection in online user-generated data: a survey, Procedia Computer Science 189 (2021) 274–281.

[8] Alothali E, Hayawi K, Alashwal H (2020) Characteristics of similar-context trending hashtags in Twitter: a case study. In: International Conference on Web Services. 2020. Springer

[9] M. Luo, X. Mu, Entity sentiment analysis in the news: A case study based on negative sentiment smoothing model (nssm), International Journal of Information Management Data Insights 2 (1) (2022), 100060.

[10] A. Rodriguez, Y.-L. Chen, C. Argueta, FADOHS: framework for detection and integration of unstructured data of hate speech on facebook using sentiment and emotion analysis, IEEE Access 10 (2022) 22400–22419.

[11] S. Khan, et al., HCovBi-caps: hate speech detection using convolutional and Bidirectional gated recurrent unit with Capsule network, IEEE Access 10 (2022) 7881–7894.

[12] H. Wu, et al., Phrase dependency relational graph attention network for Aspectbased Sentiment Analysis, Knowledge-Based Systems 236 (2022), 107736.

[13] A. Kumar, et al., BILEAT: a highly generalized and robust approach for unified aspect-based sentiment analysis: BILEAT, Applied Intelligence 52 (12) (2022) 14025–14040.

[14] R.M. Cruz, W.V. de Sousa, G.D. Cavalcanti, Selecting and combining complementary feature representations and classifiers for hate speech detection, Online Social Networks and Media 28 (2022), 100194.

[15] L.L. Tao, T.-H. You, A multi-criteria decision-making model for hotel selection by online reviews: Considering the traveller types and the interdependencies among criteria, Applied Intelligence 52 (11) (2022) 12436–12456.

[16] F.R. Nascimento, G.D. Cavalcanti, M. Da Costa-Abreu, Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on M. Subramanian et al. Alexandria Engineering Journal 80 (2023) 110–121 120 social media using ensemble learning, Expert Systems with Applications 201 (2022), 117032.

[17] H.S. Alatawi, A.M. Alhothali, K.M. Moria, Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT, IEEE Access 9 (2021) 106363–106374.

[18] Zhao, Y. Yu, Knowledge-enabled BERT for aspect-based sentiment analysis, Knowledge-Based Systems 227 (2021), 107220.

[19] D.R. Beddiar, M.S. Jahan, M. Oussalah, Data expansion using back translation and paraphrasing for hate speech detection, Online Social Networks and Media 24 (2021), 100153.

[20] S. Alsafari, S. Sadaoui, Semi-supervised self-training of hate and offensive speech from social media, Applied Artificial Intelligence 35 (15) (2021) 1621–1645.

[21] S. Kaur, S. Singh, S. Kaushal, Abusive content detection in online user-generated data: a survey,

[22] Procedia Computer Science 189 (2021) 274–281.

[23] M. Sharma, I. Kandasamy, V. Kandasamy, Deep learning for predicting neutralities in offensive language identification dataset, Expert Systems with Applications 185 (2021), 115458.

[24] J. Kocon, et al., Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach, Information Processing & Management 58 (5) (2021), 102643

[25] E. Fu, J. Xiang, C. Xiong, Deep Learning Techniques for Sentiment Analysis, Highlights in Science, Engineering and Technology 16 (2022) 1–7.

[26] Xiang, B. and L. Zhou. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. in Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers). 2014.

[27] F.E. Ayo, et al., A probabilistic clustering model for hate speech classification in twitter, Expert Systems with Applications 173 (2021), 114762.

[28] S.S. Jacob, R. Vijayakumar, Sentimental analysis over twitter data using clustering based machine learning algorithm, Journal of Ambient Intelligence and Humanized Computing (2021) 1–12.

[29] Nurce, E., J. Keci, and L. Derczynski, Detecting abusive Albanian. arXiv preprint arXiv:2107.13592, 2021.

[30] Suryawanshi, S., et al. Multimodal meme dataset (MultiOFF) for identifying offensive

content in image and text. in Proceedings of the second workshop on trolling, aggression and cyberbullying. 2020.

[31] A. Jiang, et al., SWSR: A Chinese dataset and lexicon for online sexism detection, Online Social Networks and Media 27 (2022), 100182.

[32] Wiegand, M., M. Siegel, and J. Ruppenhofer, Overview of the germeval 2018 shared task on the identification

[33] of offensive language. 2018.

[34] Mollas, I., et al., Ethos: an online hate speech detection dataset. arXiv preprint arXiv:2006.08328, 2020

[35] Salminen, J., et al. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. in Proceedings of the International AAAI Conference

[36] on Web and Social Media. 2018.

[37] E. Fu, J. Xiang, C. Xiong, Deep Learning Techniques for Sentiment Analysis, Highlights in Science, Engineering and Technology 16 (2022) 1–7.

[38] Albadi, N., M. Kurdi, and S. Mishra. Are they our brothers? analysis and detection of religious hate speech

[39] in the arabic twittersphere. in 2018 IEEE/ACM International Conference on Advances in Social Networks

[40] Analysis and Mining (ASONAM). 2018. IEEE.

[41] Curry, A.C., G. Abercrombie, and V. Rieser, ConvAbuse: Data, analysis, and benchmarks for nuanced abuse

[42] detection in conversational AI. arXiv preprint arXiv:2109.09483, 2021

[43] Caselli, T., et al. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. in Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020.

[44] Pamungkas, E.W., V. Basile, and V. Patti. Do you really want to hurt me? predicting abusive swearing in social media. in Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020.

[45] Fanton, M., et al., Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. arXiv preprint arXiv:2107.08720, 2021.

[46] Zampieri, M., et al., Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666, 2019.

[47] Qian, J., et al., A benchmark dataset for learning to intervene in online hate speech. arXiv preprint arXiv:1909.04251, 2019.