Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 16, No. 1, 2025

A Review of Key Technologies for the Analysis and Classification of Social Media Hate Speech Content.

Yingjie Liu*, Chen Gong and Xingyue Ma

School of Information Network Security People's Public Security University of China, Beijng, China
*Corresponding author: 1697666388@qq.com

Abstract: With the explosive growth of social media content and the enhanced information-sharing capabilities of platforms, the proliferation of online hate speech has become a global governance challenge. Its dissemination patterns are rapidly evolving towards multimodality (the deep integration of text and images), further complicating content security management. On one hand, the subtlety and strong contextual dependence of hate speech significantly increase the difficulty of detection. On the other hand, emerging forms of dissemination, such as meme images, present dual challenges for classification tasks due to their inherent characteristics: a seemingly humorous facade, reliance on cultural context, and semantic conflicts between text and image. To address these issues, this paper focuses on two major technical approaches: unimodal text analysis and multimodal content classification. It provides a systematic review of the research progress in hate speech detection methods based on text and multimodal detection methods, while also analyzing their limitations. Furthermore, this paper consolidates the characteristics and applicable scenarios of current mainstream unimodal and multimodal hate speech datasets, offering reference directions for optimizing technical approaches and constructing datasets in subsequent research.

Keywords: Hate Speech Detection; Machine Learning; Deep Learning; Multimodal Analysis.

1. Introduction

With the explosive proliferation of social media platforms such as Twitter, Instagram, and YouTube, users generate billions of multimodal content items daily, encompassing text, images, audio, and video. Among this content, negative statements targeting specific individuals or groups can evolve into incendiary hate speech. Such content not only exacerbates social polarization and group fragmentation but may also trigger mental health crises such as depression and anxiety, the detrimental effects of which have been substantiated in numerous sociological studies [1]. Consequently, hate speech detection holds significant practical importance.

In the era of big data, manual processing and classification of vast amounts of textual data are both time-consuming and susceptible to human bias (Mullah & Zainon, 2021). As research utilizing Natural Language Processing (NLP) for Hate Speech Detection (HSD) has grown, traditional methods relying on keyword rules or shallow machine learning techniques (e.g., TF-IDF + SVM) have proven inadequate. This is because they struggle to capture the contextual semantic relationships embedded within the highly unstructured nature of social media text, which often employs subtle expressions of hostility such as metaphors and irony. The contextual dependency, polysemy, and dynamic nature of internet slang pose significant challenges for semantic parsing in text classification. Consequently, conventional rule-based hate speech detection approaches lack scalability and adaptability, making it difficult to handle the massive volume of user-generated content on social media platforms.

In contrast, machine learning and deep learning techniques have demonstrated promising results in the automatic identification of hate speech. Furthermore, shifts in attention patterns driven by technological revolutions have fueled the popularity of hybrid content formats like memes, which integrate images and text. Leveraging their viral propagation characteristics, these formats have become dominant modes of online expression. However, the exponential growth of content within these formats that supports radical speech has created an urgent demand for automated detection technologies. The image-text hybrid structure of memes presents dual detection challenges: first, the open-ended nature of image interpretation—where the diverse semantics and cultural specificity of visual symbols render the construction of comprehensive datasets infeasible; second, the challenge of integrating cross-modal information for classification—requiring the effective combination of image-derived data with textual information to ascertain the presence of hateful intent. Based on this, we analyze and summarize the application and challenges of machine learning and deep learning in both unimodal and multimodal hate speech detection. Our study includes the following aspects:

1. We summarize various state-of-the-art benchmark hate speech datasets employed for this task.
2. We provide a concise overview of recent relevant work pertaining to hate speech detection.
3. We conduct a systematic literature review across different data modalities, specifically focusing on text-based hate speech detection and multimodal hate speech detection.

2. Hate Speech Dataset

2.1. Text Hate Speech Dataset

While existing hate speech detection datasets offer comprehensive category coverage, significant variations exist in their quality and reliability. Some early datasets, such as Waseem's [2], suffer from ambiguous semantic boundaries, and Rozental et al.'s hierarchical dataset [3] is limited by insufficient sample representativeness. In contrast, datasets supported by academic conferences demonstrate stronger standardization. For instance, the HASOC benchmark [4], employing a graded task structure (subtasks A/B), provides a systematic evaluation framework. Although baseline models

achieve relatively low F1 scores (e.g., 0.52 for binary classification), it remains a crucial research platform. Similarly, the SemEval series has advanced the field, with Task 5 [5,6] focusing on multilingual hate speech detection concerning immigrants and women, and Task 6 [7] establishing the OLID dataset, which sets fine-grained classification standards and serves as a key reference for HASOC.

Current dataset development exhibits a bipolar trend: large-scale annotation projects enhance model generalization, while smaller, specialized datasets offer deeper insights in specific contexts. However, existing research exhibits notable limitations: first, systematic evaluation of non-Twitter platform data [8] is lacking; second, multilingual extension remains severely underdeveloped, and non-English datasets [9] are often discarded due to inconsistent annotation standards; third, the trade-off for data diversity may filter out valuable features. These shortcomings highlight three critical directions for constructing high-quality datasets: cross-platform validation, culturally sensitive multilingual annotation systems, and adversarial sample augmentation mechanisms.

2.2. Hate Memes Dataset

The construction of datasets for multimodal hateful meme detection centers on adversarial sample design, cultural context sensitivity, and multimodal fusion challenges. Prominent among these is the Hateful Memes Challenge Dataset released by Facebook AI Research [10], which compels models to integrate text-image features by deliberately isolating single-modality information, establishing it as a benchmark. SemEval-2020 Task 8 [11] advances model understanding of ironic contexts by distinguishing explicit hate, subtle sarcasm, and aggressive humor. The recently released MUTE-AACL22 [12] further focuses on identifying multimodal insinuation and puns, filling gaps in detecting non-explicit hate expressions through coupled designs involving visual metaphors (e.g., repurposing a dove image as a gun) and cross-cultural contexts. Additionally, MultiOFF [13] extends detection boundaries by addressing cross-cultural metaphors (e.g., juxtaposing religious symbols) and OCR adversarial text. Current dataset evolution trends include: emphasizing cultural specificity to address localized hate symbols, incorporating multimodal pre-trained models (e.g., CLIP, ViLT) to bridge the text-image semantic gap, and employing adversarial sample generation techniques to enhance model robustness. However, challenges remain, including metaphorical ambiguity, insufficient coverage of minority languages, and ethical review controversies.

Table 1. Data setstatistics

Dataset	Scale (Number of Posts)	Dataset Category
Waseem et al.	6,909 posts	Racism, gender discrimination, neither, both
OLID Semantic Evaluation Task 6	14,000 posts	A level: criminality, non-criminality; B level: targeted insult, non-targeted insult; C level: individual, group, other
HatEval Semantic Evaluation Task 5	English: 13,000 posts, Spanish: 6,600 posts	Subtask A: hate speech, non-hate speech; Subtask B: individual target, group target; Subtask C: aggressive, non-aggressive
HASOC: Identifying Hate Speech and Offensiveness in Indic Languages	HASOC English dataset 2020: 5,335 posts, HASOC English dataset 2019: 7,005 posts	Subtask A: hateful but non-offensive; Subtask B: hate speech, offensive speech, and profanity
ElSherief et al.	25,278 haters, 22,857 targets, 27,330 posts	Ancient language, class, disability, race, gender, nationality, religion, sexual orientation, hate motivation, confrontational non-hate samples
Hateful Memes Challenge Dataset	10,000 image-text pairs	Hateful memes, adversarial non-hateful examples
SemEval-2020 Task 8	10,000 labeled samples	Task A: emotion classification; Task B: humor type recognition; Task C: semantic strength quantification
MUTE-AACL22	4,158 multimodal memes	Hate motivation, non-hate motivation
MultiOFF	1,198 multimodal memes,	Hate motivation, non-hate motivation

3. Text Hate Speech Detection

3.1. Traditional Machine Learning for Detecting Hate Speech

The anonymity of social networks fosters hate speech, which not only exacerbates group conflicts, threatens public mental health, but also potentially exerts negative societal impacts [14]. With the explosive growth of social media data, detecting hate speech has become increasingly critical. Over the past few decades, hate speech and online bullying have been among the most extensively researched areas within Natural Language Processing [15]. Machine learning algorithms have made significant contributions to hate speech detection and social media content analysis [16]. In the realm of social media data analysis, these algorithms play a crucial role in identifying and classifying offensive comments [17].

Traditional machine learning methods primarily rely on supervised learning, dependent on manually annotated hate speech datasets. They employ classical algorithms such as Naive Bayes (NB) [18-20], Decision Trees (DT), Support Vector Machines (SVM), Linear Regression [21], and Logistic Regression (LR) [22], often combined with statistical feature engineering techniques like Term Frequency-Inverse Document Frequency (TF-IDF) and N-grams, to perform shallow semantic analysis of text. However, these approaches face dual challenges: firstly, the annotation process requires linguistics experts to individually assess hate intent, proving costly and susceptible to subjective bias; secondly, the models heavily rely on domain-specific

annotated data, leading to a sharp decline in generalization performance when encountering emerging online slang (e.g., abbreviations, puns) or cross-cultural linguistic variations. To address this limitation, researchers have shifted towards semi-supervised and unsupervised learning paradigms. For instance, Gitari et al. [23] employed a bootstrapping method to construct a lexicon from hate verbs, iteratively expanding it, and achieved optimal results by incorporating more features.

To overcome the performance limitations of single models, ensemble learning methods enhance detection robustness through algorithmic fusion [24, 25]. Ensemble techniques are developed to overcome the limitations of several individual machine learning algorithms while enhancing their strengths. Since each model possesses inherent weaknesses and none is perfect, ensemble methods attempt to combine the advantages of multiple models to achieve better performance than any single model. Statistically, combining two or more machine learning algorithms can significantly reduce variance and improve learning capacity. Bagging methodology, Random Forest (RF) , and boosting methods [26] are some examples of such ensemble techniques.

The revolutionary breakthrough of word embedding technology has introduced new dimensions to semantic understanding. Embeddings learn vector representations from discrete inputs, which are then utilized in downstream text mining tasks, enabling semantically related phrases to share similar vector representations. Over the years, numerous word embedding algorithms have been developed, including Global Vectors for Word Representation (GloVe), word2vec, and FastText [27]. The representations obtained through word embedding techniques are subsequently input into various classifiers.

3.2. Deep Learning Hate Speech Detection

Deep learning utilizes multi-layered neural network architectures to automatically extract deep semantic features from text, contrasting with traditional machine learning methods which often rely on handcrafted features (e.g., TF-IDF) or rule-based techniques (e.g., Word2Vec). Through end-to-end learning, deep learning models autonomously capture hierarchical information from lexical morphology to contextual relationships, significantly enhancing the accuracy and robustness of hate speech detection [28,29]. This capability grants them distinct advantages in sentiment analysis and hate content identification, leading to widespread application in data mining and text classification.

Recurrent Neural Networks (RNNs) are specifically designed for sequential data, overcoming the limitation of standard feedforward networks which process independent data points. RNNs retain information from previous inputs, making them suitable for tasks like sentiment analysis and hate speech detection, where word context and order are crucial.

Long Short-Term Memory networks (LSTMs), a type of RNN, effectively address the vanishing gradient problem, enabling them to capture long-range dependencies in sequences. This makes LSTMs particularly advantageous for longer texts in hate speech detection. They process text incrementally, updating a hidden state to capture contextual relationships. Attention mechanisms can be integrated to enhance performance by focusing on salient hate-indicative parts of the text.

Gated Recurrent Units (GRUs) are another RNN variant designed to mitigate the vanishing gradient problem and capture long-range dependencies, similar to LSTMs . GRUs also update a hidden state incrementally and effectively model sequential dependencies for context and sentiment understanding in hate speech detection.

Convolutional Neural Networks (CNNs), while prominent in computer vision, have been successfully applied to NLP tasks including sentiment analysis and hate speech detection. CNNs treat text as a 1D sequence, applying convolutional filters to extract local patterns and features, proving effective for capturing such structures in text data.

3.3. Transformer-based Models.

Since their introduction by Vaswani et al. [30] in "Attention is All You Need," Transformer-based models have revolutionized natural language processing (NLP). These models have become state-of-the-art for various NLP tasks, primarily due to their ability to effectively handle long-range dependencies and process text in parallel, offering superior efficiency and scalability compared to traditional RNNs and CNNs. The core innovation lies in the self-attention mechanism, which allows the model to dynamically weigh the relevance of all words in the input sequence when making predictions, thereby capturing comprehensive contextual information. Prominent Transformer architectures, including Bidirectional Encoder Representations from Transformers (BERT) [31], Generative Pre-trained Transformer (GPT), and A Robustly Optimized BERT Pretraining Approach (RoBERTa) [32], have demonstrated exceptional potential across NLP applications such as sentiment analysis and hate speech detection. Typically pre-trained on massive text corpora and then fine-tuned for specific downstream tasks, these models achieve high performance even with limited task-specific data. Their effectiveness in hate speech detection stems from deep contextual understanding, bidirectional context modeling, large-scale pre-training, transfer learning, and the powerful attention mechanism.

3.4. Hybrid Models.

Beyond the aforementioned models, hybrid approaches and graph-based methods have also emerged for hate speech detection. Hybrid models combine architectures to leverage complementary strengths; for instance, Mathew et al. [33] integrated CNNs with GRUs, while Khan et al. [56] employed a CNN-LSTM framework for sentiment analysis in Roman Urdu and English dialects, enabling the capture of both local patterns and long-term dependencies in text. Recently, Graph Neural Networks (GNNs) have gained prominence in NLP by representing text as graph structures (where nodes denote words/entities and edges represent relationships, constructed via techniques like dependency parsing or co-occurrence) to exploit inherent relational information, achieving state-of-the-art results. Sarracén et al. [34] proposed an unsupervised method using Graph Autoencoders (GAEs), representing texts as nodes encoded via Transformer and convolutional layers within a low-dimensional space, with reconstruction guiding hate speech detection across domains and languages. Convolutional Graph Neural Networks (CGNNs) specifically show promise in capturing structural and linguistic patterns, demonstrated effectively by Sarracén et al. through graph construction and embedding learning. Furthermore, their work introduced an unsupervised hybrid approach combining BERT's multi-headed self-attention for contextual word relations with graph-based word inference, significantly

enhancing detection efficiency. (Figure 1 illustrates the general hate speech detection pipeline.)

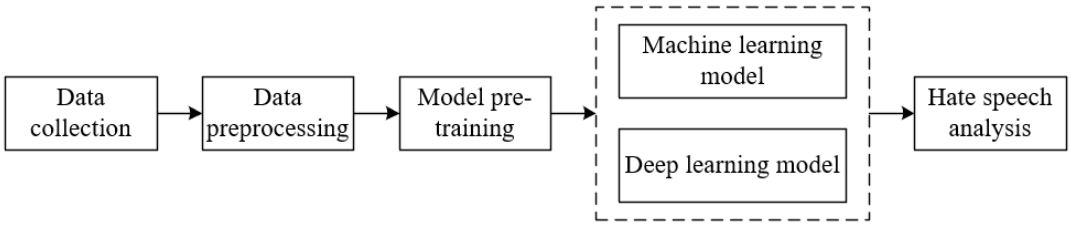

Figure 1. Text hate speech detection flow chart

4. Multimodal Hate Speech Detection

Multimodal analysis, integrating diverse data modalities (e.g., audio, visual, textual), has gained significant recent attention. Hateful meme classification exemplifies a key challenge in this domain [35]. Memes, as pervasive cultural vehicles on the internet (images, videos, posts), pose a significant societal threat due to their inherent complexity: meaning often emerges not from isolated text or visual components, which may appear benign individually, but from their contextual interplay. This multimodal fusion frequently employs irony, humor, or contextual association to generate novel—and often offensive—implicit meanings, facilitating the creation and spread of hate speech. Consequently, while detection of hateful, offensive, or aggressive content within unimodal data (text-only or image-only) is relatively mature, accurately identifying hateful memes necessitates the joint analysis and comprehension of both textual and visual modalities. Research focused on effectively combining these modalities for harmful content identification remains in its nascent stages.

References are cited in the text just by square brackets. (If square brackets are not available, slashes may be used instead, e.g.) Two or more references at a time may be put in one set of brackets. The references are to be numbered in the order in which they are cited in the text and are to be listed at the end of the contribution under a heading References, see our example below.

4.1. . General structure for meme classification

Meme classification constitutes a complex vision-language multimodal task. It differs significantly from other vision-language problems, such as image captioning, where the goal is to generate descriptive text for an image. In contrast, meme classification requires a decision based on text that is semantically related to the accompanying visual content. Consequently, cross-modal approaches integrating both visual and textual information are likely to yield superior performance for this task. While traditional vision-language methods rely on simplistic early or late fusion of separately learned unimodal features, models benefiting from multimodal pre-training demonstrate enhanced potential for meme classification. Based on a comprehensive literature review, we propose a generalized multimodal architecture for meme classification, illustrated in Figure 2. This architecture comprises two primary processing flows: a Language Processing Flow (LPF) and a Visual Processing Flow (VPF). A central Fusion and Pre-training (FPT) stage defines the strategies for merging these flows and incorporating pre-trained knowledge

Both NLP and computer vision possess rich histories in machine learning, which we categorize into first-generation (1G) and second-generation (2G) approaches. The advent of deep learning, catalyzed by breakthroughs like AlexNet [36] for vision (particularly CNNs) and BERT for NLP, marks the transition to these second-generation methods. Accordingly, we delineate both the LPF and VPF into their respective 1G and 2G implementations. Each generation will be elaborated upon in subsequent sections.

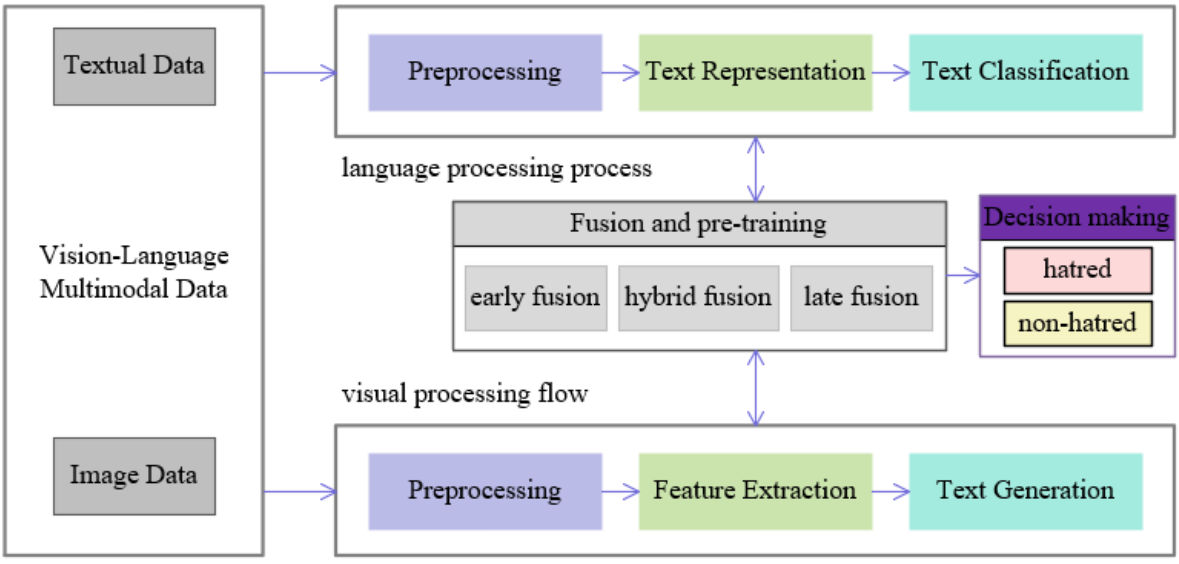

Figure 2. Meme Classification Flowchart

4.1.1. Language Processing Flow

Prior to the rise of deep learning (DL), natural language processing (NLP) pipelines predominantly relied on machine learning (ML) techniques involving sequential stages: preprocessing (stopword removal, tokenization, lemmatization), feature engineering (handcrafted/statistical features like Bag-of-Words, n-grams, TF-IDF, or static word embeddings from Word2Vec, GloVe, FastText), dimensionality reduction (e.g., PCA, LDA), and classification (e.g., SVM, Naive Bayes). While static embeddings captured general semantics better than BoW/n-grams, they lacked contextual awareness, failing to model polysemy, syntax, semantics, and coreference resolution effectively.

The advent of DL revolutionized NLP through its capacity for modeling complex nonlinear relationships. This "second-generation" shift introduced neural architectures like RNNs, CNNs, and transformative Transformer-based models (e.g., BERT, GPT-2, RoBERTa, ALBERT [37]). BERT's breakthrough, building upon semi-supervised learning, contextual representations (e.g., ELMo), transfer learning (ULMFiT [38]), and the Transformer architecture, marked a new era. Transformers utilize self-attention to capture intra-sequence relationships and enable parallelization, drastically reducing training time. The core architecture stacks encoder/decoder layers, each containing multi-head self-attention (capturing diverse representational subspaces) and position-wise feed-forward networks. Positional encodings inject order information, though they offer weaker sequential modeling than RNNs/LSTMs for precise positional dependencies. Crucially, models like BERT simplify application by learning rich linguistic knowledge via large-scale pretraining and adapting efficiently to downstream tasks through fine-tuning, eliminating extensive feature engineering.

4.1.2. visual processing flow

The success of AlexNet [36] shifted research focus from traditional computer vision methods—reliant on handcrafted features like LBP, SIFT, HOG, SURF, and BRIEF—towards deep learning. Traditional approaches faced significant limitations: feature engineering required task-specific redesign, classifiers lacked transferability, and the overall pipeline (preprocessing, feature extraction, feature selection, classification) necessitated inefficient optimization of each sub-step. Convolutional Neural Networks (CNNs) overcame these issues by automatically learning features from data and enabling efficient fine-tuning for related tasks via transfer learning.

AlexNet's breakthrough performance on ImageNet catalyzed a paradigm shift, marking the advent of the second-generation (2G) era for the Visual Processing Flow (VPF). Deep learning models, particularly CNNs, facilitate the integration of diverse, large-scale image and video datasets, learning robust features that generalize effectively. A CNN typically combines convolutional layers with activation functions (e.g., ReLU variants), downsampling layers (e.g., max pooling), fully convolutional layers, dense layers, and a final softmax layer. Post-AlexNet, research focused on enhancing CNN architectures. VGG [39] and GoogLeNet's Inception modules demonstrated benefits from increasing depth and width. ResNets [40] introduced residual learning blocks with identity shortcut connections, enabling training of networks hundreds or thousands of layers deep. DenseNet [41] further enhanced representational power through dense inter-layer connectivity. Significant progress was also made in object detection with numerous CNN-based methods (e.g., RCNN, Faster R-CNN, YOLO).

4.1.3. Fusion and Pre-training: Towards Multimodality.

Following, visual-language (VL) multimodal fusion strategies are categorized into three types: early fusion (combining modalities at the feature level immediately after extraction), late fusion (integrating decisions made independently by each modality), and hybrid fusion (merging outputs from both unimodal predictors and an early fusion component). Pre-training approaches are similarly distinguished: models combining unimodally pre-trained language and vision components via fusion are termed unimodally pre-trained multimodal models. Conversely, models pre-trained jointly on multimodal data are referred to as multimodally pre-trained models.

4.2. The Latest Advances in Meme Classification

Given the limited research specifically on multimodal meme classification, we draw inspiration from state-of-the-art approaches in related Vision-Language (VL) tasks. A foundational objective in VL exploration is understanding alignment relationships between multimodal feature spaces. Architectures combining CNNs and RNNs are commonly trained on aligned multimodal data to learn a joint embedding space, a standard approach for tasks like image captioning [75]. In contrast, Visual Question Answering (VQA) focuses on fusing VL modalities to determine the correct answer, necessitating precise correlation modeling between image and question representations rather than explicit alignment learning. Hateful meme detection similarly requires precise correlation modeling between image and text to uncover latent inter-modal relationships for accurate classification. Consequently, we leverage insights from VQA literature to inform the development of state-of-the-art meme classification models.

Early VQA research employed simple feature concatenation (early fusion), while later methods utilized bilinear pooling for multimodal feature learning. However, these approaches faced significant limitations: fusing features late in the model hindered effective VL alignment extraction, and representing CNN outputs as 1D vectors led to substantial degradation of spatial information [42]. Recent efforts have shifted towards cross-modal learning via multimodal pre-trained models like VisualBERT and UNITER [43], which have surpassed previous methods across multiple VL benchmarks.

5. Conclusion

This paper presents a comprehensive review of recent advancements in hate speech detection on social media. Although hate speech represents a relatively nascent research domain within computer science, it has long been a significant concern in the humanities and arts. Consequently, this review aims to provide researchers with a thorough understanding of current developments in the field. The analysis encompasses a broad spectrum of methods, including traditional machine learning, ensemble techniques, and deep learning approaches specifically applied to hate speech detection in social media content. By synthesizing insights from published literature,

this work serves as a valuable reference for researchers. It underscores the critical importance of ongoing research in this area, enabling the academic community to focus on advancing methodologies for hate speech detection.

Acknowledgements

The authors gratefully acknowledge the financial support from the Double First-Class Special Task for Security and Prevention Engineering at the Chinese People's Public Security University. (2023SYL08)

References

[1] Blaya C. Cyberhate: A review and content analysis of intervention strategies[J]. Aggression and violent behavior, 2019, 45: 163-172.

[2] Talat Z, Hovy D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter[C]//Proceedings of the NAACL student research workshop. 2016: 88-93.

[3] Rozental A, Biton D. Amobee at SemEval-2019 tasks 5 and 6: Multiple choice CNN over contextual embedding[J]. arxiv preprint arxiv:1904.08292, 2019.

[4] Mandl T, Modha S, Majumder P, et al. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages[C]//Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation. 2019: 14-17.

[5] Wang B, Ding H. YNU NLP at SemEval-2019 task 5: Attention and capsule ensemble for identifying hate speech[C]//Proceedings of the 13th International Workshop on Semantic Evaluation. 2019: 529-534.

[6] Yang X, Obadinma S, Zhao H, et al. SemEval-2020 task 5: Counterfactual recognition[J]. arxiv preprint arxiv:2008.00563, 2020.

[7] Zampieri M, Malmasi S, Nakov P, et al. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)[J]. arxiv preprint arxiv:1903.08983, 2019.

[8] Guest E, Vidgen B, Mittos A, et al. An expert annotated dataset for the detection of online misogyny[C]//Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume. 2021: 1336-1350. 2021

[9] Mulki H, Haddad H, Ali C B, et al. L-hsab: A levantine twitter dataset for hate speech and abusive language[C]//Proceedings of the third workshop on abusive language online. 2019: 111-118.

[10] Kiela D, Firooz H, Mohan A, et al. The hateful memes challenge: Detecting hate speech in multimodal memes[J]. Advances in neural information processing systems, 2020, 33: 2611-2624.

[11] Pàmies M, Öhman E, Kajava K, et al. LT@ Helsinki at SemEval-2020 Task 12: Multilingual or language-specific BERT?[J]. arxiv preprint arxiv:2008.00805, 2020.

[12] Hossain E, Sharif O, Hoque M M, et al. Deciphering hate: identifying hateful memes and their targets[J]. arxiv preprint arxiv:2403.10829, 2024.

[13] Suryawanshi S, Chakravarthi B R, Arcan M, et al. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text[C]//Proceedings of the second workshop on trolling, aggression and cyberbullying. 2020: 32-41.

[14] Wu J, Hong Q, Cao M, et al. A group consensus-based travel destination evaluation method with online reviews[J]. Applied Intelligence, 2022, 52(2): 1306-1324.

[15] Rodriguez A, Argueta C, Chen Y L. Automatic detection of hate speech on facebook using sentiment and emotion analysis[C]//2019 international conference on artificial intelligence in information and communication (ICAIIC). IEEE, 2019: 169-174.

[16] [19] Al-Garadi M A, Hussain M R, Khan N, et al. Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges[J]. IEEE Access, 2019, 7: 70701-70718.

[17] Weir G, Owoeye K, Oberacker A, et al. Cloud-based textual analysis as a basis for document classification[C]//2018 International Conference on High Performance Computing & Simulation (HPCS). IEEE, 2018: 672-676.

[18] Nurce E, Keci J, Derczynski L. Detecting abusive albanian[J]. arxiv preprint arxiv:2107.13592, 2021.

[19] Albadi,N.,M.Kurdi,andS.Mishra.Aretheyourbrothers?analysis anddetection ofreligioushatespeechinthearabictwittersphere.in2018IEEE/ACM InternationalConferenceonAdvancesinSocialNetworksAnalysisandMining (ASONAM).2018.IEEE.

[20] Davidson T, Warmsley D, Macy M, et al. Automated hate speech detection and the problem of offensive language[C]//Proceedings of the international AAAI conference on web and social media. 2017, 11(1): 512-515.

[21] Mollas I, Chrysopoulou Z, Karlos S, et al. ETHOS: a multi-label hate speech detection dataset[J]. Complex & Intelligent Systems, 2022, 8(6): 4663-4678.

[22] Wiegand M, Siegel M, Ruppenhofer J. Overview of the germeval 2018 shared task on the identification of offensive language[J]. 2018. 2018

[23] Gitari N D, Zu** Z, Damien H, et al. A lexicon-based approach for hate speech detection[J]. International Journal of Multimedia and Ubiquitous Engineering, 2015, 10(4): 215-230.

[24] Liao W, Zeng B, Yin X, et al. An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa[J]. Applied Intelligence, 2021, 51: 3522-3533. Appl Intell, 2021

[25] Fersini E, Rosso P, Anzovino M. Overview of the task on automatic misogyny identification at IberEval 2018[J]. Ibereval@ sepln, 2018, 2150: 214-228.

[26] Salminen J, Almerekhi H, Milenković M, et al. Anatomy of online hate: develo** a taxonomy and machine learning models for identifying and classifying hate in online news media[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2018, 12(1). 2018

[27] Salminen J, Almerekhi H, Milenković M, et al. Anatomy of online hate: develo** a taxonomy and machine learning models for identifying and classifying hate in online news media[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2018, 12(1)

[28] Pitsilis G K, Ramampiaro H, Langseth H. Detecting offensive language in tweets using deep learning[J]. arxiv preprint arxiv:1801.04433, 2018.

[29] Fu E, **ang J, **ong C. Deep Learning Techniques for Sentiment Analysis[J]. Highlights in Science, Engineering and Technology, 2022, 16: 1-7.

[30] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[31] Kennedy C J, Bacon G, Sahn A, et al. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application[J]. arxiv preprint arxiv:2009.10277, 2020.

[32] Pavlopoulos J, Sorensen J, Laugier L, et al. SemEval-2021 task 5: Toxic spans detection[C]//Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021). 2021: 59-69.

[33] Mathew B, Saha P, Yimam S M, et al. Hatexplain: A benchmark dataset for explainable hate speech detection[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(17): 14867-14875. 2021

[34] De la Peña Sarracén G L, Rosso P. Unsupervised embeddings with graph auto-encoders for multi-domain and multilingual hate speech detection[C]//Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022: 2196-2204. 2196～2204

[35] Gandhi A, Adhvaryu K, Khanduja V. Multimodal sentiment analysis: review, application domains and future directions[C]//2021 IEEE Pune section international conference (PuneCon). IEEE, 2021: 1-5.

[36] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[37] Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arxiv preprint arxiv:1909.11942, 2019.

[38] Howard J, Ruder S. Universal language model fine-tuning for text classification[J]. arxiv preprint arxiv:1801.06146, 2018.

[39] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arxiv preprint arxiv:1409.1556, 2014.

[40] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[41] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708. 4700～4708

[42] Gomez R, Gibert J, Gomez L, et al. Exploring hate speech detection in multimodal publications[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2020: 1470-1478.

[43] Chen Y C, Li L, Yu L, et al. Uniter: Universal image-text representation learning[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 104-120.