A Survey on Combating Hate Speech by Detection and Prevention

Anonymous ACL submission

Abstract

With the rise in social media (SM) platforms that offer easy access, community formation, and online debate, the issue of hate speech has risen rapidly. The hate detection, and countering it becomes a growing challenge to society, researchers, companies, and policymakers. Hate speech is in the form of text or multimodal such as memes, GIFs, audio, or video. The scientific study of hate speech from a computer science view has gained attention in recent years. Mostly it is considered a supervised task where the annotated corpora and shared resources play a big role. To combat it, SM, employing modern AI tools is getting attention. This survey comprehensively examines the work done to combat hate in the English language so far. This structures the state-of-the-art methodologies employed for unimodal identification, studies conducted in multimodal hate identification, the role of Explainable AI, prevention of hate speech through style transfer, and counter-narrative generation for the English language. The efficacy and limitations are also discussed. Compared with the earlier surveys this paper concisely gives a well-organized presentation of the methods to combat hate.

1 Introduction

The recent exponential growth of the internet, technology, and social media has revolutionized communication but also provides a platform to disseminate hateful content. United Nations strategy and plan of action on hate speech describes hate speech as any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language concerning a person or a group based on who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or identity factor [1]. Hate speech is used as a broad umbrella term for numerous user-created content intended to disparage, or dehumanize, any individual or any group based on some characteristics such as race, color, gender, nationality, ethnicity, etc (Nockleby, 2000). With the advancement of natural language processing (NLP) technology, substantial research has been conducted on automatic textual hate speech detection in recent years. There are large-scale publicly available datasets collected from various social medial platforms and tagged into subvariants of hate such as aggression ((Kumar et al., 2018),(Bhattacharya et al., 2020) Hate((Toraman et al., 2022)(Davidson et al., 2017), (Mathew et al., 2021), (Mollas et al., 2020)), Offensive ((Davidson et al., 2017), (Zampieri et al., 2019), (Rosenthal et al., 2021)), Abusive (Nobata et al., 2016),(Curry et al., 2021), (Caselli et al., 2020),(Founta et al., 2018)), Harassment((Golbeck et al., 2017)) Toxic ((Wulczyn et al., 2017),(Sarker et al., 2023b), (Bhat et al., 2021),(Georgakopoulos et al., 2018)), Cyberbullying (Dadvar et al., 2012) (Dinakar et al., 2012), Racism (Waseem and Hovy, 2016)(Kwok and Wang, 2013), Sexism (Waseem and Hovy, 2016),Flame (Spertus et al., 1997) Misogynistic (Fersini et al., 2022). Facebook reports 510K comments/minute and X reports 350 tweets per minute [2]. Recent research has focused on developing automatic systems to detect hate speech on social media platforms. These typically employ semantic content analysis techniques built on Natural Language Processing (NLP) and Machine Learning (ML) methods such as Support vector machine (SVM), Logistic regression (LR), Convolution neural network (CNN), Long short-term memory (LSTM), Gated recurrent unit (GRU), Bidirectional encoder representations from the transformer (BERT), etc. The task typically involves classifying a comment into non-hate or hateful and measured in terms of

[1]https://www.un.org/en/hate-speech/un-strategy-and-plan-of-action-on-hate-speech

[2]https://bernardmarr.com/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

performance metrics. Hate speech is disseminated via multimodal data such as memes (text superimposed within images), audio, and video. Most of the work revolves around meme identification with the concept of early fusion and late fusion. With the emergence of LLM, the employment of a vision transformer-based approach for identification has risen. However, little research focus is on audio or video identification. Recently model has been aided with the explanability information for better classification. The multimodal approach focuses on leveraging the multi modalities features. (Gomez et al., 2020), (Suryawanshi et al., 2020), (Kirk et al., 2022) covers the task. Recently the identification of hate span, transforming/ rephrasing the offensive into non-offensive has been in attention to counter the hate speech. The remainder of this paper is structured as follows. Section 2 reviews the dataset organized for the different subtasks of hate identification, Section 3 describes the features related to hate speech detection in the uni-modal identification. Section 4 introduces the tasks done to solve multimodal hate prediction, Section 5 presents the study to counter hate speech; and finally, Section 6 is about the implementation and role of explainable AI, Section 7 reports some challenges and Section 8 concludes this work and discusses future work.

2 Corpus Details

This section covers the dataset collection process, the annotator's role, available datasets, and challenges associated.

2.1 Data set collection and preparations

Most of the work done in hate speech detection for unimodal and multimodal relies on the labeled data. These corpus are mainly crawled through Twitter (Wijesiriwardene et al., 2020) (Jha and Mamidi, 2017) (Fersini et al., 2018), Facebook (Kumar et al., 2018), Reddit (Mollas et al., 2020), Gab (Kennedy et al., 2018), Wikipedia comments (Wulczyn et al., 2017) (Pavlopoulos et al., 2020). Nearly all the user-generated content has been crawled using a keywords-based approach (Waseem and Hovy, 2016) (De Gibert et al., 2018) with words being in negative polarity. Most of the datasets are topical focus (Kumar et al., 2018) (Founta et al., 2018) i.e the specific topics and abusive phenomena addressed The preprocessing is performed depending on the data quality and structure. This typically involves filtering and normalizing textual inputs, such as tokenization, stopword removal, misspelling correction, noise reduction, stemming, and lemmatization.

2.2 Annotations

The preprocessed data requires a manual review of the post/meme/audio to tag it into further granularity of hate. The data annotation is a relevant source of variability. There are various annotation frameworks (Founta et al., 2018)(Bhattacharya et al., 2020) (Zampieri et al., 2019). Typically the annotations are performed by hiring the experts, amateur/non-experts, or on crowdsourcing platforms such as Crowdflower (Davidson et al., 2017), and Amazon Mechanical Turk (Zampieri et al., 2019) (Founta et al., 2018). The annotators are generally pre-informed about the task. As per the annotation scheme, there are three main strategies. The first is a binary scheme: two mutually-exclusive values, (typically yes/no) to mark the presence or absence of a given phenomenon. The second is a non-binary scheme: more than two mutually exclusive values. The third strategy features multi-level annotation, with finer-grained schemes accounting for different phenomena. The quality of annotated data is measured by the Inter Annotators agreement score. Most of the authors did not give much about the annotation process and only provided an Inter-annotator agreement score. Fleiss Kappa (Zampieri et al., 2019), Cohen Kappa (Golbeck et al., 2017) (De Gibert et al., 2018), Kripendrof's Kappa (Kumar et al., 2018) (Bhattacharya et al., 2020)

2.3 Available datasets

The most common binary annotated corpus is (Golbeck et al., 2017) (De Gibert et al., 2018), (Bretschneider and Peters, 2016), (Ghosh et al., 2022) (Gao and Huang, 2017). To have better coverage of the hate variants, the binary task shifted to the single layer 3-class (Davidson et al., 2017)(Waseem and Hovy, 2016)(Toraman et al., 2022)(Mathew et al., 2021) or multi-class (Founta et al., 2018)(Waseem, 2016)annotated data. There is an issue of imbalance in the number of hate and non-hate instances. (Davidson et al., 2017), (De Gibert et al., 2018), (Curry et al., 2021) constitutes 6%-20% hateful instances. (Kurrek et al., 2020) (Mathew et al., 2021) (Pavlopoulos et al., 2021) constitute 50%-60% of the abusive instances. (Röttger et al., 2021), (Borkan et al., 2019) consists of around 70%-80% hate instances. The single layer tagging is shifted towards the creation of hierarchical annotation schema (Zampieri et al., 2019) (Basile et al., 2019) (Mandl et al., 2021) covering the targets associated with hate in the subsequent layers. The targets are trans people (Röttger et al., 2021), religion (Mathew et al., 2021)(Kennedy et al., 2020), misogyny (Fersini et al., 2018). The tagging is also done at the multinomial level (Mollas et al., 2020), on a scale (Wulczyn et al., 2017), multi-task multi-tagging (Vidgen et al., 2020). The recent shift in marking the toxic span is also gaining pace. (Pavlopoulos et al., 2021) proposed annotated data for toxic span. (Sarker et al., 2023a) released ≈ 20K comments marked for toxic phrases. The shift in data creation from unimodal to multimodal is slow. The creation of multimodal data has recently gained pace with the datasets like (Kiela et al., 2020), (Gomez et al., 2020), (Suryawanshi et al., 2020), (Fersini et al., 2022), (Ramamoorthy et al., 2022), (Aprosio et al., 2020), (Yang et al., 2019) (Tekiroglu et al., 2022) (Qian et al., 2019a). To combat the hate and generate counter-narrative statements some datasets (Chung et al., 2019)(Fanton et al., 2021)(Qian et al., 2019a)

3 Unimodal: Textual Identification

The distinguishing approach in the classification tasks is the usage of features. This section covers the various approaches utilized to compute the features, and various methods employed to improve the performance of the classifier. The encoded features are generally applied to the machine learning or deep neural network to get the probability distribution of the classes.

3.1 Simple surface features

Traditional Machine learning algorithms utilize surface features such as word n-gram, and char n-grams features (Nobata et al., 2016) (Waseem and Hovy, 2016) (Zhang et al., 2018) (Chen et al., 2012) (Xu et al., 2012). These features were feded into Support vector machine (SVM) (Kapil and Ekbal, 2020)(Zhang et al., 2018), logistic Regression (LR) (Waseem and Hovy, 2016)(Qian et al., 2018), Random Forest (RF)(Davidson et al., 2017). The other linguistic features such as Part-of-Speech (PoS) tag unigrams, bigrams, and trigrams, weighted by their TF-IDF and removing any candidates with a document frequency lower than 5; number of syllables; Flesch-Kincaid Grade Level and Flesch Reading Ease scores that to measure the 'readability' of a document (Zhang et al., 2018), (Davidson et al., 2017). (Gambäck and Sikdar, 2017), (Kapil and Ekbal, 2020) (Gambäck and Sikdar, 2017) used CNN with n-gram approach. Character n-grams provide the model to capture the obfuscation such as fck, kll, a$$hole. It is found to be more predictive than token n-grams (Mehdad and Tetreault, 2016)

3.2 Word Embeddings

With the passage of time, distributed word representations (based on neural networks) also referred to as word embeddings are developed. These are Word2vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2017), GloVE (Pennington et al., 2014). Word Embedding is based on distributed assumptions and mapped words into a high-dimension feature space and maintains the semantic information. For each target sentence S = w_1, w_2, , w_N, each token w_i is substituted into a real-valued vector x_i using word embedding, where $x_i \in \mathrm{R}^d$ is the word vector, d is the dimensions of word vectors. These word embeddings were used with CNN (Badjatiya et al., 2017)(Kapil and Ekbal, 2020), LSTM (Zhou et al., 2021a) (Pitsilis et al., 2018), GRU (Zhang et al., 2018), (Zhou et al., 2021a)

3.3 Transformer based approaches

The inclusion of transformer-encoder-based features outperformed the traditional machine learning and deep neural network techniques. It is leveraging the concept of multi-head self-attention (Vaswani et al., 2017). These models have emerged as the preferred approach for a variety of NLP tasks, owing to their capacity for effectively handling long-range dependencies while processing text in parallel. This parallel processing makes them more efficient and scalable than standard RNNs and CNNs. Transformer-based models' basic notion is their attention mechanism, which allows the model to focus on relevant areas of the input text while making predictions. More than 60% of the models submitted in the shared task (Bhattacharya et al., 2020)(Mandl et al., 2021) were based on transformer encoder embeddings (Curry et al., 2021)(Basile et al., 2019) (Fersini et al., 2022). Specifically BERT is used by (Mozafari et al., 2020a)(Zhou et al., 2021a)(Mathew et al., 2021).

3.4 Lexical resources

To make use of the general assumption that hateful posts contain negative words, these words can be used as the feature. There are many publicly available hate-related lexicons. The domain-specific lexicons is created by (Davidson et al., 2017) of size 179, (Bassignana et al., 2018) created HurtLeX, a multilingual lexicon of <100,000 hate words in 53 languages,(Olteanu et al., 2018) created 163 hate words, (Qian et al., 2019b) collected 2105 lexicons, and (Wiegand et al., 2018) proposed 1651 words. (Gitari et al., 2015) created a lexicon using subjectivity and syntactic features related to hate speech. (Xiang et al., 2012), (Nobata et al., 2016) (Burnap and Williams, 2016), (Burnap and Williams, 2015) employed lexicon lists, recently BERT based methods (Koufakou et al., 2020) leveraged from the lexicon. The recent development in encoding has seen lesser creation of lexicon to capture standardized vocabulary and semantic information.

3.5 Knowledge enriched features

The creation of a large number of annotated data poses a great challenge. It is therefore a wise idea to transfer this knowledge via Multi-task learning (MTL), transfer learning, zero-shot learning, few-shot learning etc. Given m learning tasks

$$\{T_i\}_{i=1}^{m} \tag{1}$$

where all the tasks or subset of them are related, multi-task learning aims to help improve the learning of a model for classification task T_i by using the knowledge in some or all of the m tasks. (Kapil and Ekbal, 2020) experimented CNN-based MTL on five hate datasets. (Ghosh et al., 2023a) transformer-based multi-task network, to address (a) aggression identification, (b) misogynistic aggression identification, (c) identifying hate-offensive and non-hate-offensive content, (d) identifying hate, profane, and offensive posts, (e) type of offense. The other form of MTL were employed such as Fuzzy based (Liu et al., 2019), multi-task multi-lingual (Mishra et al., 2021). The empirical analysis showed the approaches following MTL outperformed the other classifier with the (Maity et al., 2023) analyzing the efficacy of MTL over Single task learning (STL). Transfer learning: Transfer learning aims to transfer the learned knowledge in one domain or application to another domain for which no data exists. (Mozafari et al., 2020a) fine-tuning BERT-based transfer learning, and (Yuan et al., 2023) explored deep transfer learning by projecting multiple datasets in a common space. (Qian et al., 2021) proposed Variational Representation Learning (VRL) along with a memory module based on LB-SOINN (Load-Balancing Self-Organizing Incremental Neural Network) to lifelong data learning without forgetting the previously learned knowledge, There are some other learning such as Few-shot learning (FSL) i.e generally as n-shot learning, a category of artificial intelligence that also includes one-shot learning (in which there is only one labeled example of each class to be learned) and zero-shot learning (in which there are no labeled examples at all). Several work involved the usage of these learning (Mozafari et al., 2022), (Awal et al., 2023), (Pamungkas et al., 2021)

3.6 Relation with Sentiment analysis and emotion

Hate speech data is closely related to sentiment and emotion analysis, as understanding the underlying negative sentiments and intense emotions is crucial for accurate detection and effective intervention. (Gitari et al., 2015) (Dinakar et al., 2012) followed the approach where a classifier dedicated to detecting negative polarity is applied prior to the classifier specifically checking for evidence of hate speech. (Van Hee et al., 2015) uses sentiment lexicon to identify the number of positive, negative, and neutral words in a comment text. The BERT-based models have also leveraged the sentiment and emotion data in the training. (Min et al., 2023) validate the correlations between hate speech and certain negative emotion states and propose an emotion-correlated hate speech detector. (Rajamanickam et al., 2020) advantage of the affective features to gain auxiliary knowledge through a Hard-sharing double encoder model and gated double encoder based on BILSTM. (Zhou et al., 2021a) use multiple feature extraction units to share multi-task parameters to better share sentiment knowledge, and then gated attention is used to fuse features for hate speech detection. (Kapil and Ekbal, 2021) proposed CNN-based MTL sharing sentiment analysis data. (Kapil and Ekbal, 2022) (Ghosh et al., 2023a) make use of sentiment and emotion recognition data in the BERT-based MTL.

3.7 Augmentation

As the neural networks are data-specific the performance of the model can be enhanced by increasing the training data by augmentation and solving the problem of data scarcity and data imbalance. Most researchers have employed pre-trained transformers to generate synthetic posts. (Wullach et al., 2021) utilized GPT LLM (BERT, RoBERTa, ALBERT) for generating synthetic data (Ilan and Vilenchik, 2022) applied data augmentation using real unlabelled data, selected from the online platform. Unlike other data augmentation approaches that generate synthetic data, HARALD (Hate Augmentation with ReAL Data) generates a continuous stream of relevant real data authored by multiple authors with diverse stylistic, grammatical, and semantic forms. (Hartvigsen et al., 2022) created machine-generated datasets TOXIGEN by developing a demonstration-based prompting framework and an adversarial classifier-in-the-loop decoding method to generate subtly toxic and benign text with a massive trained language model. (Kim et al., 2023) proposed TOXIGEN-CONPROMPT, a pre-training strategy to leverage machine-generated data via contrastive learning. (Cao and Lee, 2020) deep generative reinforcement learning adversarial generated based data augmentation to enhance the performance by 5%.

3.8 Impliciteness

The detection method mainly works well for hate expressed explicitly. One of the challenging aspects is to detect hate expressed in an implicit manner (Kumar et al., 2018)(Kim et al., 2022) (Hartvigsen et al., 2022). Previous research has mostly addressed overt or explicit hate speech, in an accurate way neglecting the more prevalent type of coded or indirect language.(ElSherief et al., 2021) proposed benchmark corpus. In (Wiegand et al., 2021), Wiegand discusses the challenges of learning implicit abuse in existing datasets and suggests improvements to their design. (Qian et al., 2019b)deciphered hate symbols using a sequence-to-sequence model using Urban Dictionary. (Ocampo et al., 2023a) generated adversarial implicit hate messages leveraging auto-regressive models. (Ghosh et al., 2023b) explicitly incorporates user- and conversational context to detect implicit hate (Wiegand et al., 2023) proposed new data set generated from GPT-3 to identify euphemistic abuse. (Cooper et al., 2023) designed Hate speech detection models inoculated against real-world homoglyphs. (Ocampo et al., 2023b) investigate implicit and explicit embedding representations. (Kim et al., 2022) leveraged contrastive learning to learn implicit posts.

4 Multi-modal

The early works of multimodal hate identification involve the usage of meta tweet features aided to the main tweet (Founta et al., 2018), (Qian et al., 2018). (Pitsilis et al., 2018) proposed an ensemble of recurrent neural network(RNN) classifiers, incorporating various features associated with user-related information, such as users' tendency towards racism or sexism. (Founta et al., 2019) (Chatzakou et al., 2017) utilizes a wide variety of metadata such as tweet-based, user-based, and network-based features. The properties of bullies and aggressors were studied. (Rajadesingan et al., 2015) derived 10 features grouped into text-based features, emotion-based features, familiarity-based features, contrast-based features, and complexity-based features (Waseem and Hovy, 2016) leveraged the gender and demographic information, (Unsvåg and Gambäck, 2018) investigates the potential effects of users' features such as gender, network (number of followers and friends), activity (number of statuses and favorites), and profile information (geo-enabled, default profile, default image, and number of public lists). (Chaudhry and Lease, 2022) investigate profiling users by their past utterances as an informative prior. But in the current scenario, social media has also seen an upsurge in memes, GIFs, audio, and video to propagate hate. However, most of the data are available for multimodal meme identification. Memes – that have recently emerged as popular engagement tools and which, in their usual form, are image macros shared through social media platforms mainly for amusement – are also being increasingly used to spread hate and/or instigate social unrest, and therefore seem to be a new form of expression of hate speech on online platforms (Fersini et al., 2022)(Suryawanshi et al., 2020). Some of these multimodal publications are only hate speech because of the combination of the text with a certain image (Kiela et al., 2020). Multimodal hate speech detection integrates various data types, such as text, images, audio, and video, to enhance the accuracy and robustness of identifying hate speech. The next part covers the feature extractor and usage of a multimodal pre trained transformer.

4.1 Feature Extraction

The text superimposed is generally extracted through Optical character recognition (OCR). Unimodal feature extraction: The textual feature is extracted by using pre-trained word embedding (Mikolov et al., 2013)(Pennington et al., 2014) through LSTM ((Gomez et al., 2020), (Botelho et al., 2021), (Aman et al., 2021) RF ((Gomez et al., 2020) CNN (Suryawanshi et al., 2020). The transformer encoder BERT ((Sabat et al., 2019), (Kiela et al., 2020), (Hossain et al., 2022), (Prasad et al., 2021)), to get encoded text representations. Several pre-trained CNN architectures have been used. These are Imagenet used by (Gomez et al., 2020) (Sabat et al., 2019)(Hossain et al., 2022) Xception (Botelho et al., 2021) VGG 16 (Suryawanshi et al., 2020) (Aman et al., 2021) (Lee et al., 2021) ResNET (Ma et al., 2022) (Zhang et al., 2023a). Early multimodal identification work generally involves merging the unimodal features through fusion. To have better representations unimodal features were fused based on concatenation (Kumar et al., 2021) (Kiela et al., 2020) (Hossain et al., 2022)(Kumar and Nandakumar, 2022). The fusion based on summation (Kumar et al., 2021),(Zhou et al., 2021b). The transformer architecture serves as the foundation for today's cutting-edge vision language learning models. There are two main approaches: Single-stream models/ early fusion, such as VisualBERT (Kiela et al., 2020), UNITER (Zhang and Wang, 2022) (Lippe et al., 2020), OSCAR (Lippe et al., 2020) (Kiela et al., 2020), uses a single transformer to process the image and language input at the same time. Dual-stream models/ late fusion, such as LXMERT (Lippe et al., 2020), CLIP (Kumar and Nandakumar, 2022), DeVLBERT, VilBERT (Lee et al., 2021), rely on separate transformers for vision and language, which are then combined towards the end of the model. New approaches leveraging the multi-modal techniques to enhance the performance have been proposed.

4.2 Context aware information

(Zhou et al., 2021b) utilizes image captioning process (Xu et al., 2022) proposed MET-Meme rich in metaphors . (Cao et al., 2022) proposed PromptHATE to prompt pre-trained language models (PLMs) for multimodal classification. (Shang et al., 2021)developed GNN based KnowMeme to enrich from human commonsense knowledge. (Hossain et al., 2024) developed context-aware framework, (Pramanick et al., 2021) proposed MOMENTA that leverages local and global perspectives to detect memes. (Botelho et al., 2021) decipher implicit hate (Yang et al., 2022) uses cross-domain knowledge transfer (Chhabra and Vishwakarma, 2023) leverages knowledge distillation architecture

4.3 Audio and Video detection

(Rana and Jha, 2022) proposed new Video hate detection data and combined the auditory features representing emotion and the semantic features to detect hateful content. (Das et al., 2023) curate 43 hours of videos from BitChute and manually annotate them as hate or non-hate, along with the frame spans which could explain the labelling decision. They showed that models having multiple modalities surpasses the performance obtained by uni-modal variants. (Gupta et al., 2023) explore the context for hate detection for video pages by using like description, transcript, and vsual input) (Ibañez et al., 2021) develop a hate speech classifier from online short-form TikTok videos (Bhesra et al.) collected audio based hate speech data, (Prasad et al., 2023) video frame features in the multimodal identification.

5 Dehatify

This section mainly deals with the advancement in the style transfer and counter-narrative response. Preventing hate speech through style transfer entails rephrasing toxic information in neutral or positive language, and using advanced NLP techniques to change the tone while preserving content. In NLP, style transfer involves adding certain stylistic attributes to text while maintaining its basic structure and meaning. It follows the concept of encoder and decoder. The model is trained using unsupervised (no parallel data) or in a supervised manner (parallel data).

5.1 Span Prediction

Span prediction refers to identifying the start and end positions of a relevant text segment within a larger document. the inclusion of shared task (Pavlopoulos et al., 2021)To ease the moderators, this part will predict the toxic span. There were 36 system submission, winner employing BERT with CRF. The results were computed using character-based F1. (Ranasinghe and Zampieri,

2021) present MUDES, a multilingual system to detect offensive spans in texts. It features pre-trained models, a Python API for developers, and a user-friendly web-based interface. (Pouran Ben Veyseh et al., 2022) proposed multi-task setting for toxic span prediction, and (Nouri, 2022) developed data augmentation with dual training for Offensive Span Detection

5.2 Style transfer

(Mangal and Jindal) filtered out hate words based on a lexicon. The void is predicted by using Google with the CBOW model. The second approach uses back translation to lose the original style but preserves content, it is then regenerated using desired styles. (Santos et al., 2018) trained GRU-based encoder-decoder using non-parallel data. The framework combines collaborative classifier, attention, and cycle consistency loss. (Ahmad et al., 2022) proposed a decoding technique following lexical constraints over the zero-shot style transfer method. (Masud et al., 2022) curated a parallel corpus of hate texts and their counterpart. A model NACL, a hate speech normalization operating in three stages: identifying the hate posts identifying the toxic span, and then rephrasing it to non-hate. (Tran et al., 2020) designed a retrieve, generate, and edit unsupervised style transfer pipeline. The part of Speech (POS) tag sequences is identified followed by the generation of suitable candidates and corrected by the edit module. (Atwell et al., 2022) released a parallel corpus of comments with its style-transferred counterparts. The proposed model leverages discourse framework and parsing to preserve content.

5.3 Counter Narratives

The counter-narrative data is prepared with the intervention of humans. These data will be trained and the output is to generate counter narration with respect to the post. (Bonaldi et al., 2022)presented generated dialogue data aided by the intervention of human expert annotators to automatize counter-narrative writing. (Hong et al., 2024) proposed constrained generation of counter speech by incorporating two conversation outcomes in the text generation by prompt with instructions, prompt and select, LLM finetune, and LLM Transformer reinforcement learning. (Tekiroglu et al., 2020) employed Generative pre-trained transformer (GPT)-2 to generate silver counter-narratives followed by expert validation/post-editing. (Chung et al., 2019) described the creation of the first large-scale multilingual hate speech/counter-narrative pairs by experts. (Fanton et al., 2021) presented a HITL framework for data collection based on an author-reviewer paradigm. (Chung et al., 2021) presented a knowledge-bound counter-narrative incorporating external knowledge retrieved through extracted and generated keyphrases. The process of dehatification needs to be more researched into with the sota methods.

6 Model Implementation and Explainable AI

6.1 Model parameters and Evaluation metric

The experiments were performed using a 5-fold cross-validation (Zampieri et al., 2019)(Ghosh et al., 2022) (Kapil and Ekbal, 2020) approach. The 4-fold training set is split into 15% validation and 85% training while the last fold is treated as the test set to evaluate the model. Most of the deep learning models were implemented using Keras (Zhang et al., 2018) (Pitsilis et al., 2018) with Tensorflow as the backend. Evaluation of the performance of hate speech (and also other related content) detection typically adopts the classic Precision, Recall, and F1 metrics. Precision measures the percentage of true positives among the set of hate speech messages identified by a system. The model employing precision (Badjatiya et al., 2017) (Dinakar et al., 2012)(Wiegand et al., 2018), recall (Burnap and Williams, 2015)(Gitari et al., 2015)(Waseem and Hovy, 2016) The model performance for unimodal is measured by F1 (harmonic mean of precision, and recall) (Kapil and Ekbal, 2020)(Waseem and Hovy, 2016)(Zhang et al., 2018)(Badjatiya et al., 2017). Most of the multimodal models employ AUC-ROC (Kumar et al., 2021)(Kiela et al., 2020) (Shome and Kar, 2021) as its metric. The F1 score also used (Hossain et al., 2022)(Aman et al., 2021)(Lee et al., 2021) The quantitative metrics generally used in the generative task are consistency preservation (Santos et al., 2018), perplexity (Santos et al., 2018) (Masud et al., 2022), BLEU (Bilingual Evaluation Understanding) (Ahmad et al., 2022) (Masud et al., 2022)(Tran et al., 2020)(Atwell et al., 2022), ROGUE (Tran et al., 2020), METEOR (Tran et al., 2020) The novelty of generated text is also measured using relevance, and effectiveness (Hong et al., 2024)(Bonaldi et al., 2022)

6.2 Mitigating Bias

Annotator bias refers to the systematic errors or tendencies introduced by individuals who label or annotate data used in machine learning and other data-driven applications. (Wich et al., 2021)(Al Kuwatly et al., 2020) This bias can affect the quality, reliability, and generalizability of the annotated data, leading to skewed or misleading results in models trained on such data. (Waseem, 2016) concluded that annotator bias can stem from various sources, including personal biases, unclear tagging details, task complexity, social bias, etc. Several bias mitigation methods are proposed to make the model more efficient. (Cheng et al., 2021) proposed debiasing strategy based on Reinforcement learning (RL), (Sahoo et al., 2022)extraction of social bias data, (Zhang et al., 2023b) introduced two mitigation approaches such as multi-task intervention, and data-specific intervention. (Mun et al., 2023)(Elsafoury et al., 2022) investigated countering of stereotypical bias,(Badjatiya et al., 2019) (Maity et al., 2019) mitigated internal stereotypical bias through knowledge representations, (Davidson et al., 2019)studied racial bias (Xia et al., 2020) proposed demoting racial bias by adversarial training, (Mozafari et al., 2020b) mitigated racial bias (Ahmed et al., 2022) tackled racial bias using geometric learning, (Halevy et al., 2021) mitigating racial bias using ensemble and (Shah et al., 2021) studied reducing target group bias.

6.3 Explanable AI

The performance of the model can be enhanced by making the model learn the human rationale of the input in an explainable form. (Lin et al., 2024) explainable approach through reasoning (**?**) introduced knowledge informed encoder-decoder to generate implications of biased text, (Clarke et al., 2023) introduced rule By example, an exemplar-based contrastive learning framework to explainable hate speech detection. (Yang et al., 2023) introduced the framework HARE, harnessing the reasoning capabilities of LLMs.

7 Challenges

Degradation of datasets, non-uniform definitions of hate, non-disclosure of the annotation guidelines, annotators' bias, time-consuming annotation, mental illness, etc. The mental health of hate victims has also been studied.

7.1 Effect on Mental Health

Cyberbullying and other subhate can be a detrimental cause in mental health. The computational approach has not solved it, rather a string of surveys based on questionnaires, and responses, the degree of scale of depression is studied. (Bucur et al., 2021) analyzed the mental depression states related to postings (Saha et al., 2019) psychological effects of hateful speech related to depression. (Wachs et al., 2022) relationship between online hate speech victimization and adolescents's mental well-being. The questionnaires were administered to assess OHSV, depressive symptoms, and resilience. (Torres et al., 2020) analyzed the effect of social, verbal, physical, and cyberbullying victimizations on academic performances.

8 Conclusion and Future Work

In this survey, we provided a critical assessment of how the automatic identification of hate speech in text has advanced over the last several years. Other realms of hate speech that we examined included cyberbullying, abusive language, discrimination, sexism, extremism, and radicalization. The work done in the unimodal text identification, multimodal hate identification, style transfer, counter narrative generation, discussion on mental health is done. The future work should more focus on fine grained hate detection, more mathematical efficient fusion approach, adding more explanability, and via continuous learning paradigm.

Limitations

Hate speech detection is a very vast domain covering multiple languages. This survey covers only the research done so far for the English language. The number of open repositories is very few, and the inconsistent guidelines and differences in annotator expertise further complicate the reliability of the data, impacting the effectiveness and accuracy of detection models. The data in most cases very difficult to share because of privacy issues. Most of the work completed is not deployed and if deployed released by very few. The multimodal audio and video identification are in the very preliminary stage.

References

Zishan Ahmad, Vinnakota Sai Sujeeth, and Asif Ekbal. 2022. Zero-shot hate to non-hate text conversion using lexical constraints. *IEEE Transactions on Computational Social Systems*.

Zo Ahmed, Bertie Vidgen, and Scott A Hale. 2022. Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning. *EPJ Data Science*, 11(1):8.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the fourth workshop on online abuse and harms*, pages 184–190.

Aayush Aman, Gopal Krishna, Tushar Anand, and Anubhaw Lal. 2021. Identification of offensive content in memes. In *Data Science and Security: Proceedings of IDSCS 2021*, pages 438–445. Springer.

Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2020. Creating a multimodal dataset of images and text to study abusive language. *arXiv preprint arXiv:2005.02235*.

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. Appdia: A discourse-aware transformer-based style transfer model for offensive social media conversations. *arXiv preprint arXiv:2209.08207*.

Md Rabiul Awal, Roy Ka-Wei Lee, Eshaan Tanwar, Tanmay Garg, and Tanmoy Chakraborty. 2023. Model-agnostic meta-learning for multilingual hate speech detection. *IEEE Transactions on Computational Social Systems*.

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The world wide web conference*, pages 49–59.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Elisa Bassignana, Valerio Basile, Viviana Patti, et al. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *CEUR Workshop proceedings*, volume 2253, pages 1–6. CEUR-WS.

Meghana Moorthy Bhat, Saghar Hosseini, Ahmed Hassan, Paul Bennett, and Weisheng Li. 2021. Say 'yes'to positivity: Detecting toxic language in workplace communications. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2017–2029.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168.

Kirtilekha Bhesra, Shivam Ashok Shukla, and Akshay Agarwal. Audio vs. text: Identify a powerful modality for effective hate speech detection. In *The Second Tiny Papers Track at ICLR 2024*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Austin Botelho, Bertie Vidgen, and Scott A Hale. 2021. Deciphering implicit hate: Evaluating automated detection algorithms for multimodal hate. *arXiv preprint arXiv:2106.05903*.

Uwe Bretschneider and Ralf Peters. 2016. Detecting cyberbullying in online communities.

Ana-Maria Bucur, Marcos Zampieri, and Liviu P Dinu. 2021. An exploratory analysis of the relation between offensive language and mental health. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3600–3606.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242.

Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5:1–15.

Rui Cao and Roy Ka-Wei Lee. 2020. Hategan: Adversarial generative-based data augmentation for hate speech detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6327–6338.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. 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*, pages 6193–6202.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22.

Prateek Chaudhry and Matthew Lease. 2022. You are what you tweet: Profiling users by past tweets to improve hate speech detection. In *International Conference on Information*, pages 195–203. Springer.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 international conference on privacy, security, risk and trust and 2012 international confernece on social computing*, pages 71–80. IEEE.

Lu Cheng, Ahmadreza Mosallanezhad, Yasin N Silva, Deborah L Hall, and Huan Liu. 2021. Mitigating bias in session-based cyberbullying detection: A non-compromising approach. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, volume 1.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. Multimodal hate speech detection via multi-scale visual kernels and knowledge distillation architecture. *Engineering Applications of Artificial Intelligence*, 126:106991.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. Conan–counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. *arXiv preprint arXiv:1910.03270*.

Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. *arXiv preprint arXiv:2106.11783*.

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. 2023. Rule by example: Harnessing logical rules for explainable hate speech detection. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Portia Cooper, Mihai Surdeanu, and Eduardo Blanco. 2023. Hiding in plain sight: Tweets with hate speech masked by homoglyphs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2922–2929.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. Convabuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational ai. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403.

Maral Dadvar, Franciska MG de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*, pages 23–25. Universiteit Gent.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1014–1023.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):1–30.

Fatma Elsafoury, Steven R Wilson, Stamos Katsigiannis, and Naeem Ramzan. 2022. Sos: Systematic offensive stereotyping bias in word embeddings.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. *arXiv preprint arXiv:2107.08720*.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549.

Elisabetta Fersini, Paolo Rosso, Maria Anzovino, et al. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science*, pages 105–114. ACM.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*.

Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, and Vassilis P Plagianakos. 2018. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th hellenic conference on artificial intelligence*, pages 1–6.

Soumitra Ghosh, Asif Ekbal, Pushpak Bhattacharyya, Tista Saha, Alka Kumar, and Shikha Srivastava. 2022. Sehc: A benchmark setup to identify online hate speech in english. *IEEE Transactions on Computational Social Systems*, 10(2):760–770.

Soumitra Ghosh, Amit Priyankar, Asif Ekbal, and Pushpak Bhattacharyya. 2023a. A transformer-based multi-task framework for joint detection of aggression and hate on social media data. *Natural Language Engineering*, 29(6):1495–1515.

Sreyan Ghosh, Manan Suri, Purva Chiniya, Utkarsh Tyagi, Sonal Kumar, and Dinesh Manocha. 2023b. Cosyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6159–6173.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Shrey Gupta, Pratyush Priyadarshi, and Manish Gupta. 2023. Hateful comment detection and hate target type prediction for video comments. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3923–3927.

Matan Halevy, Camille Harris, Amy Bruckman, Diyi Yang, and Ayanna Howard. 2021. Mitigating racial biases in toxic language detection with an equity-based ensemble framework. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–11.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.

Lingzi Hong, Pengcheng Luo, Eduardo Blanco, and Xiaoying Song. 2024. Outcome-constrained large language models for countering hate speech. *arXiv preprint arXiv:2403.17146*.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, M Ali Akber Dewan, Nazmul Siddique, and Md Azad Hossain. 2022. Identification of multilingual offense and troll from social media memes using weighted ensemble of multimodal features. *Journal of King Saud University-Computer and Information Sciences*, 34(9):6605–6623.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah M Preum. 2024. Align before attend: Aligning visual and textual features for multimodal hateful content detection. *arXiv preprint arXiv:2402.09738*.

Michael Ibañez, Ranz Sapinit, Lloyd Antonie Reyes, Mohammed Hussien, Joseph Marvin Imperial, and Ramon Rodriguez. 2021. Audio-based hate speech classification from online short-form videos. In *2021 International Conference on Asian Language Processing (IALP)*, pages 72–77. IEEE.

Tal Ilan and Dan Vilenchik. 2022. Harald: Augmenting hate speech data sets with real data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2241–2248.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16.

Prashant Kapil and Asif Ekbal. 2020. A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, 210:106458.

Prashant Kapil and Asif Ekbal. 2021. Leveraging multi-domain, heterogeneous data using deep multitask learning for hate speech detection. *arXiv preprint arXiv:2103.12412*.

Prashant Kapil and Asif Ekbal. 2022. Transformer based ensemble learning to hate speech detection leveraging sentiment and emotion knowledge sharing. In *CS & IT Conference Proceedings*, volume 12. CS & IT Conference Proceedings.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv. July*, 18.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679.

Youngwook Kim, Shinwoo Park, Youngsoo Namgoong, and Yo-Sub Han. 2023. Conprompt: Pre-training a language model with machine-generated data for implicit hate speech detection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Hannah Kirk, Bertie Vidgen, Paul Röttger, Tristan Thrush, and Scott Hale. 2022. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, Viviana Patti, et al. 2020. Hurtbert: Incorporating lexical features with bert for the detection of abusive language. In *Proceedings of the fourth workshop on online abuse and harms*, pages 34–43. Association for Computational Linguistics.

Deepak Kumar, Nalin Kumar, and Subhankar Mishra. 2021. Quarc: Quaternion multi-modal fusion architecture for hate speech classification. In *2021 IEEE international conference on big data and smart computing (BigComp)*, pages 346–349. IEEE.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 1–11.

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 1621–1622.

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5138–5147.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 2359–2370.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*.

Han Liu, Pete Burnap, Wafa Alorainy, and Matthew L Williams. 2019. Fuzzy multi-task learning for hate speech type identification. In *The world wide web conference*, pages 3006–3012.

Zhiyu Ma, Shaowen Yao, Liwen Wu, Song Gao, and Yunqi Zhang. 2022. Hateful memes detection based on multi-task learning. *Mathematics*, 10(23):4525.

Krishanu Maity, Gokulapriyan Balaji, and Sriparna Saha. 2023. Towards analyzing the efficacy of multi-task learning in hate speech detection. In *International Conference on Neural Information Processing*, pages 317–328. Springer.

Krishanu Maity, Nilabja Ghosh, Raghav Jain, Sriparna Saha, and Pushpak Bhattacharyya. 2019. Stereo-hate: Towards identifying stereotypical bias and target group in hate speech detection. *Natural Language Engineering*, 1:00.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schaefer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, et al. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. *arXiv e-prints*, pages arXiv–2112.

Arpan Mangal and Deepanshu Jindal. Style transfer: Saving the world from abusive speech.

Sarah Masud, Manjot Bedi, Mohammad Aflah Khan, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Proactively reducing the hate intensity of online posts via hate speech normalization. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3524–3534.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pages 299–303.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Changrong Min, Hongfei Lin, Ximing Li, He Zhao, Junyu Lu, Liang Yang, and Bo Xu. 2023. Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective. *Information Fusion*, 96:214–223.

Sudhanshu Mishra, Shivangi Prasad, and Shubhanshu Mishra. 2021. Exploring multi-task multi-lingual learning of transformer models for hate speech and offensive speech identification in social media. *SN Computer Science*, 2:1–19.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset. *arXiv e-prints*, pages arXiv–2006.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020a. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 928–940. Springer.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020b. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2022. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896.

Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9759–9777.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

John T Nockleby. 2000. Hate speech. *Encyclopedia of the American constitution*, 3:1277–79.

Nasim Nouri. 2022. Data augmentation with dual training for offensive span detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2569–2575.

Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. 2023a. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2758–2772.

Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. 2023b. Unmasking the hidden meaning: Bridging implicit and explicit hate speech embedding representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6626–6637.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, 58(4):102544.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pages 59–69.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742.

Amir Pouran Ben Veyseh, Ning Xu, Quan Tran, Varun Manjunatha, Franck Dernoncourt, and Thien Huu Nguyen. 2022. Transfer learning and prediction consistency for detecting offensive spans of text. *Findings of the Association for Computational Linguistics: ACL 2022*.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455.

Nishchal Prasad, Sriparna Saha, and Pushpak Bhattacharyya. 2021. A multimodal classification of noisy hate speech using character level embedding and attention. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Nishchal Prasad, Sriparna Saha, and Pushpak Bhattacharyya. 2023. Multimodal hate speech detection from videos and texts. Technical report, EasyChair.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019a. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 118–123.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2019b. Learning to decipher hate symbols. In *Proceedings of NAACL-HLT*, pages 3006–3015.

Jing Qian, Hong Wang, Mai ElSherief, and Xifeng Yan. 2021. Lifelong learning of hate speech classification on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2304–2314.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 97–106.

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint modelling of emotion and abusive language detection. *arXiv preprint arXiv:2005.14028*.

Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava DaS, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*.

Aneri Rana and Sonali Jha. 2022. Emotion based hate speech detection using multimodal learning. *arXiv preprint arXiv:2202.06218*.

Tharindu Ranasinghe and Marcos Zampieri. 2021. Mudes: Multilingual detection of offensive spans. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 144–152.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. Solid: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58.

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *arXiv preprint arXiv:1910.02334*.

Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM conference on web science*, pages 255–264.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting unintended social bias in toxic language datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 132–143.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. *arXiv preprint arXiv:1805.07685*.

Jaydeb Sarker, Sayma Sultana, Steven R Wilson, and Amiangshu Bosu. 2023a. Toxispanse: An explainable toxicity detection in code review comments. In *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–12. IEEE.

Jaydeb Sarker, Asif Kamal Turzo, Ming Dong, and Amiangshu Bosu. 2023b. Automated identification of toxic code reviews using toxicr. *ACM Transactions on Software Engineering and Methodology*, 32(5):1–32.

Darsh J Shah, Sinong Wang, Han Fang, Hao Ma, and Luke Zettlemoyer. 2021. Reducing target group bias in hate speech detectors. *arXiv e-prints*, pages arXiv–2112.

Lanyu Shang, Christina Youn, Yuheng Zha, Yang Zhang, and Dong Wang. 2021. Knowmeme: A knowledge-enriched graph neural network solution to offensive meme detection. In *2021 IEEE 17th International Conference on eScience (eScience)*, pages 186–195. IEEE.

Debaditya Shome and Tejaswini Kar. 2021. Conoffense: Multi-modal multitask contrastive learning for offensive content identification. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4524–4529. IEEE.

Ellen Spertus et al. 1997. Smokey: Automatic recognition of hostile messages. In *Aaai/iaai*, pages 1058–1065.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41.

Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. *arXiv preprint arXiv:2204.01440*.

Serra Sinem Tekiroglu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. *arXiv preprint arXiv:2004.04216*.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225.

Christopher E Torres, Stewart J D'Alessio, and Lisa Stolzenberg. 2020. The effect of social, verbal, physical, and cyberbullying victimization on academic performance. *Victims & Offenders*, 15(1):1–21.

Minh Tran, Yipeng Zhang, and Mohammad Soleymani. 2020. Towards a friendly online community: An unsupervised style transfer framework for profanity redaction. *arXiv preprint arXiv:2011.00403*.

Elise Fehn Unsvåg and Björn Gambäck. 2018. The effects of user features on twitter hate speech detection. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 75–85.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the international conference recent advances in natural language processing*, pages 672–680.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting east asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172.

Sebastian Wachs, Manuel Gámez-Guadix, and Michelle F Wright. 2022. Online hate speech victimization and depressive symptoms among adolescents: The protective role of resilience. *Cyberpsychology, Behavior, and Social Networking*, 25(7):416–423.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Maximilian Wich, Christian Widmer, Gerhard Hagerer, and Georg Groh. 2021. Investigating annotator bias in abusive language datasets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1515–1525.

Michael Wiegand, Jana Kampfmeier, Elisabeth Eder, and Josef Ruppenhofer. 2023. Euphemistic abuse–a new dataset and classification experiments for implicitly abusive language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16280–16297.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language–what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056.

Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L Shalin, Krishnaprasad Thirunarayan, Amit Sheth, and I Budak Arpinar. 2020. Alone: A dataset for toxic behavior among adolescents on twitter. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, pages 427–439. Springer.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4699–4705.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2887–2899.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666.

Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. 2022. Multimodal hate speech detection via cross-domain knowledge transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4505–4514.

Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the third workshop on abusive language online*, pages 11–18.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. Hare: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505.

Lanqin Yuan, Tianyu Wang, Gabriela Ferraro, Hanna Suominen, and Marian-Andrei Rizoiu. 2023. Transfer learning for hate speech detection in social media. *Journal of Computational Social Science*, 6(2):1081–1101.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420.

Jing Zhang and Yujin Wang. 2022. Srcb at semeval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 585–596.

Linhao Zhang, Li Jin, Xian Sun, Guangluan Xu, Zequn Zhang, Xiaoyu Li, Nayu Liu, Qing Liu, and Shiyao Yan. 2023a. Tot: topology-aware optimal transport for multimodal hate detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4884–4892.

Zhehao Zhang, Jiaao Chen, and Diyi Yang. 2023b. Mitigating biases in hate speech detection from a causal perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6610–6625.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 745–760. Springer.

Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin. 2021a. Hate speech detection based on sentiment knowledge sharing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 7158–7166.

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021b. Multimodal learning for hateful memes detection. In *2021 IEEE International conference on multimedia & expo workshops (ICMEW)*, pages 1–6. IEEE.