Hate Speech Detection: A Comprehensive Survey

Mariam Joan
University of Illinois, Urbana-Champaign
mjoan2@illinois.edu

December 8, 2024

Abstract

The purpose of this literature review is to explore advancements in natural language processing, either algorithmic and or deep learning techniques, that support identification of harmful or negatively impacting words, classified as hate speech, present within online text-based communication platforms such as within chats, and or comments that lead to adverse effects, either for an individual or group of individuals. We will focus on understanding a brief historical and global view of hate speech, its legislation, and highlight early to current US and English based research with technological developments of hate speech detection while understanding its challenges and where we are headed.

1 Introduction

The topic of hate speech within the academic community of natural language processing (NLP) is broad. When we look at what the effect of hate speech has on a particular community or persons, it is clear there must be regulation. We will explore how hate speech has been historically detected, along with current tactics, and look to significant research in this domain to understand where we are headed and what challenges remain. Hate speech can be destructive and have lasting adverse effects on the individual or groups in question. It is for this reason, why it is important to understand and advocate for the detection of hateful language, such that we might consider a scalable natural language processing solution spanning multiple domains, languages, and platforms that can be generalizable but targeted enough to mitigate its harmful effects.

2 Background

In this section, we will begin with a brief overview of approximately when hate speech came into the global conversation, even before the internet went public in April of 1993. After World War II, civil rights activity ignited to support preservation belonging to any kind of race or religion without fear of attack or genocide given recent radical Nazism. The United Nations became a strong voice for human rights and freedom of expression. The United Nations Human Rights Committee developed the Universal Declaration of Human Rights (UDHR) with Article 1 and 2, referencing how humans are equal without regard to religion, race, nationality, among others [13]. In 1966, the United Nations, crafted the International Covenant on Civil and Political Rights (ICCPR), which describes an ideal, yet critically important view that directly ties in with hate speech legislation. See a condensed excerpt from the ICCPR [12], Article 19 and 20 below.

- Article 19.1 Everyone shall have the right to hold opinions without interference.

- Article 19.2 Everyone shall have the right to freedom of expression, impart information and ideas, either orally, in writing or in print, through any media of choice.
- Article 20.2 Any advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence shall be prohibited by law.

There are many organizations within the United Nations, such as the Committee on the Elimination of Racial Discrimination, (CERD) who strongly advise legal enforcement of hate speech as it relates to racial discrimination. While it is clear there is a balance between allowing freedom of hateful expression, and determining if that expression inflicts harm on a human in an abusive way. There is also a component of understanding whether this hateful expression, is detrimental enough such that it should be regulated.

Is it true that words simply on a page or within an online forum can bring on harm to another human? This is a fundamental question that we must come back to again and again because there is simply no one answer. There are humans that are marginalized and or are being targeted for political, societal, economic, cultural, racial, ethnic reasons, and yet time and again we have seen hate towards others for such reasons can ignite wars, genocide, and the like. Does this mean that all hate speech can incur such disasters? Not necessarily, but hate itself has the ability to do so, and from what we have seen with regard to cases of cyberbullying, and loss of life because of bullying, we should know that truly no good thing comes from targeted hate speech, especially in our online and ultra connected society.

2.1 Legislation

The legislative response to hate speech varies. The United States Constitution is firm on freedom of speech, unlike other countries which include hate speech law. Although, as of the last decade, we have seen United States technology and social media companies advocate against speech that incites harm, discrimination, and or violence, toward a particular group of persons identified by their race, religion, or nationality, yet there are limits as to what these companies can do by law to mitigate or ban such behavior.

Social media platforms such as X (formerly Twitter), and Meta (formerly Facebook) have hate speech policies publicly available on their website. Meta's policy is noted as "Hate Speech" www.meta.com/policies/ while X's policy is termed as "Hateful Conduct" www.x.com/rule-and-policies/. The premise of both policies target primarily content that "attacks" humans based on race, ethnicity, gender, sexual orientation, and other labeled groups, while Meta defines these attributes as "protected characteristics".

3 Hate speech detection research

In this section, we will explore a number of important hate speech detection research papers to understand what natural language processing (NLP) techniques were used, what is used now and what advancements have these innovations made toward more accurate hate speech detection. We will learn that researchers consider other types of language, related to hate speech such as "abusive", and "offensive" and as we'll see, these modeling and detection methods can overlap so these studies are included. While it is difficult to pinpoint an exact date for when research in this domain began, we provide an example from the earliest period found after the internet went public, from 1997.

3.1 Smokey: Automatic Recognition of Hostile Messages (Spertus, E., 1997)

Spertus created software appropriately named "Smokey" that could correctly detect upwards of 64% of flames within a dataset out of 720 messages using a 47-element feature vector, and a Quinlan decision

tree generator [19]. The term ”flames” identified negative speech in text. Spertus utilized the programming language, LISP and a sentence parser from the Microsoft Research Natural Language Processing Group [6].

	okay	*maybe*	*flame*	*okay:flame*
original	574	88	58	9.9:1
removed	359	28	10	35.9:1
remaining	215	60	48	4.5:1

Table 2: Messages of each type and okay:flame ratio in original training set, removed messages, and remaining messages.

Figure 1: *Spertus, E., Smokey: Automatic Recognition of Hostile Messages, pg. 4*

The decision tree program utilized Rule classes such as Noun Phrase or Appositions. Spertus found many cases in which the word, ”You” was followed by a negative term such as ”bozos”, or where ”You” was used as second person or even alongside bad words such as “loser”, “idiot”, [19] etc. Additionally, shorter imperative statements, detected by the parser tended to be insulting such as with “Get Lost!!”, or “f** you” [19]. Alternatively, Smokey was also able to detect polite or more friendlier statements from positive words such as with, “kudos”, “great”, “super”, [19] etc.

```
If (Imperative-short (13)  > 0 ^
    Condescension-somewhat (21) <= 0 ^
    site-specific-insult (29) > 0)
   (Imperative-short (13) <= 1 ^
    Insult-recipient (25) > 0)
   (Insult-other (28) > 0 ^
    epithet (30) > 0)
   (Imperative-short (13) <= 0 ^
    Profanity-no villain (19) > 0)
   (Appositive-guys (1) > 0 ^
    site-specific-insult (29) > 0)
¤class flame

If (Appositive-NP (3) <= 0 ^
    Insult-villain (27) <= 0 ^
    site-specific-insult (29) <= 0 ^
    epithet (30) <= 0 ^
    exclamation-points (47) <= 2)
   (Appositive-NP (3) <= 0 ^
    Imperative-short (13) <= 0 ^
    site-specific-insult (29) <= 0 ^
    epithet (30) <= 0)
¤class ok
```

Figure 1: Ordered rules generated by C4.5. Numbers in parentheses are rule numbers

Figure 2: *Spertus, E., Smokey: Automatic Recognition of Hostile Messages, pg. 4*

The Quinlan decision tree was used like a classifier labeling data as *okay*, *maybe*, or *flame* then compared against a human labeled dataset in which case Smokey achieved 64% of flames correctly detected. Spertus had applied ordinary least squares regression which produced accurate coefficients, however, it did not

perform well on the test data [19]. The challenges Spertus discussed are relevant to what we see in today's hate speech detection, which is a model's difficulty in detecting sarcasm, and also challenges related to incorrect grammar, and or punctuation which we now have solutions for, e.g. character ngrams. Spertus looks to morphological analysis or logical parsing trees for sentences as possible solutions [19].

This paper lists a range of important research references to what could be attributed to hate speech from as early as 1977. For example, the International Journal of Verbal Aggression, an academic journal studying negative use of words, along with origin, etymology and meaning even before the internet was invented.

3.2 Detecting Hate Speech on the World Wide Web (Warner, Hirschberg, 2012)

Researchers Warner, and Hirschberg from Columbia University, found that a pattern of hate speech detection includes a "small set of high frequency stereotypical words" [20]. The authors partnered with Yahoo! and the American Jewish Congress to obtain data perceived as offensive including website content and news groups however, it was noted that although the websites appeared to be antisemitic they didn't contain derogatory terms but rather well seemingly crafted essays of antisemitic beliefs [20]. This is a key example of "implicit" hate speech language as we learned from Davidson, et al., because it is the implicit style that is harder to detect. The Yahoo data was difficult in terms of grammar and spelling, with wording such as, j@e@w@ and other misspellings.

There was a strong distinction they found with the language of stereotyping, having its own patterns, and themes, such as with anti-Hispanic speech referencing illegal borders entry, or anti-African American speech referencing unemployment. They believed in this so much that they trained a language model to support this categorization. The authors considered this as a word sense disambiguation task for and classified the website data by paragraph rather than sentence as to obtain more context within a ten word window [20].

Word sense disambiguation identifies the sense of content words in context. In this type of classifier model, we get the sense of a word depending on its context. For example, how long is the context, do the words need to be ordered or can we use a bag of words (BOW) approach, are POS tags included. The distributional hypothesis was a theory from Zellig Harris in 1954 which claimed that a word is likely to show its meaning by the word it's surrounded by.

In this research the authors believed that stereotype was the key to word sense. The data was human annotated with such categories as anti-woman, anti-muslim, anti-black, etc., and not anti-woman, not anti-muslim, and so on. They supported a Fleiss kappa inter-annotator agreement for 3 annotations which was ultimately used as their corpora and created a "gold" corpora which achieved a 59% precision and 68% recall as a baseline for the classifier [20].

Table 2: Classification Performance

	Accuracy	Precision	Recall	F1
Majority All Unigram	0.94	0.00	0.00	0.00
Majority Positive Unigram	0.94	0.67	0.07	0.12
Majority Full Classifier	0.94	0.45	0.08	0.14
Gold All Unigram	0.94	0.71	0.51	0.59
Gold Positive Unigram	0.94	0.68	0.60	0.63
Gold Full Classifier	0.93	0.67	0.36	0.47
Human Annotators	0.96	0.59	0.68	0.63

Table 3: Error Report

	False Negative	False Positive
Majority All Unigram	6.0%	0.1%
Majority Positive Unigram	5.6%	0.2%
Majority Full Classifier	5.5%	0.6%
Gold All Unigram	4.4%	1.8%
Gold Positive Unigram	3.6%	2.5%
Gold Full Classifier	5.7%	1.6%

Figure 3: *Warner, Hirschberg, Detecting Hate Speech on the World Wide Web, pg. 6*

Learnings from the authors include that hate speech can be considered as a classification problem, while bigram and trigram modeling did not support performance. With recall being lower, they suggested looking further into the tree parser, for possible patterns of phrases. The modeling used was a support vector machine (SVM) which took in the feature vectors, and the word sense, as 1 being antisemitic or 0 as not antisemitic where feature weights were log odds multiplied by the sense [20].

3.3 Automated Hate Speech Detection and the Problem of Offensive Language (Davidson, et al., 2017)

Davidson, et. al, from Cornell University and Qatar Computing Research Institute, spoke at the Eleventh International Association for the Advancement of Artificial Intelligence Conference on Web and Social Media in 2017 on the challenges of separating hate speech from offensive language using a multi-class classifier and humans to categorize data as either hate speech, offensive language or neither.

The authors utilized a lexicon from Hatebase https://hatebase.org/ to create a dataset from X (formerly Twitter) of approximately 25,000 tweets, of which were manually labeled as three categories, hate speech, offensive language or neither. However, with only 5% of tweets labeled as hate speech, and over 70% labeled as offensive language, the authors determined the lexicon was not robust enough and the human labeling needed to be reviewed [2].

The tweets were vectorized and stemmed to create bigram, unigram and trigram features weighted using TF-IDF and then tagged using Penn's Part-of-Speech (POS) tagging. For modeling, logistic regression was used with L1 regularization and tested against random forest, decision trees, naive Bayes, and linear SVMs [2]. It was determined that logistic regression with L2 regularization would prove better given previous research [2]. The model features included weighting based on tweet quality using Flesch-Kincaid Grade Level for fixed sentences, and tweet sentiment based from a sentiment lexicon in addition to frequency per tweet of mentions, retweets, hashtags, etc [2].

The authors found there are differences culturally, for example, with the word, "gay" or how African-American's use the term "n**a", or even "b*tch" when quoting rap lyrics [2]. The use of syntactic features highlighted ability to detect hate speech given verb and noun occur, e.g. "kill <minority group >" or a POS trigram the authors provided an example as <intensity ><user intent ><hate target >[2].

Modeling techniques used during this timeframe in NLP's history were constrained to ngram, linear regression and tree based. In this case, using both L1 and L2 regularization, achieved overall accuracy of 91% and recall and F1 at 90% [2]. The authors found that results of higher accuracy were more racial or homophobic in nature. In addition, there were problems with human labeling, incorrectly classifying offensive only when it related to sexism while hateful when racist or homophobic [2].

Overall, while the authors attempted to differentiate hate speech and offensive language, they found the latter as simply, "common place", as per human labelers [2]. Lexicons are not effective for hate speech but work better for offensive language while there are reliable patterns of words that can be used to highly predict hate speech for easy identification.

3.4 Understanding Abuse: A Typology of Abusive Language Detection Subtasks (Waseem, et al., 2017)

Given growth of hate speech within the natural language processing academic community, there have been multiple subtasks created to understand not just hate speech, and offensive language, but also "abusive language", [21] etc. The authors believe there is significance in understanding overlap in techniques used for detection. Subtasks are a term within a research community that allow multiple researchers to develop solutions for a particular problem such that outcomes can be shared and reused, and learned from.

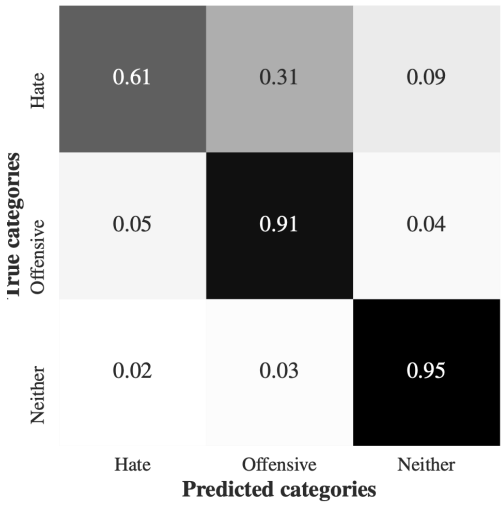


Figure 1: True versus predicted categories

Figure 4: *Davidson, et al., Automated Hate Speech Detection and the Problem of Offensive Language, pg. 3*

Regarding overlaps in understanding, cyberbullying and or trolling have been seen as a personal attack versus more discriminatory language directed at a group of people. There is also distinction with statements that are more "explicit or implicit" [21] where the ladder can be thought of as not outright racial slurs themselves but ambiguous or sarcastic language. This was the basis for the author's typology.

The research outcome recommended using a typology to categorize the type of hate speech, such as directed or generalized, and implicit or explicit. The authors found the best features were POS tagging, coreference resolution, character ngrams, and word embeddings. For each component in the typology the authors describe how different features can provide better model outcomes. For example, directed abuse, POS sequences, and coreference resolution are useful in detecting patterns of phrases [21]. Coreference resolution can support identification of expressions, such as syntactically that can connect back to the same entity [21]. In conclusion, Waseem, et al., believe their typology can provide a framework that "synthesizes the different subtasks in abusive language detection" [21].

3.5 A Survey on Hate Speech Detection Using Natural Language Processing (Schmidt, et al., 2017)

This research paper from Schmidt and Weigand, out of Saarland, Germany is the most cited for research for hate speech detection. The authors discuss automation of hate speech detection and features such as bag of words (BOW), in addition to use of character n-grams given the wide usage of characters or symbols in place of letters such as with, e.g. "ki11" [18]. The researchers also apply clustering when faced with data sparsity, using Latent Dirichlet Allocation (LDA) which is a Bayesian probabilistic model. The LDA model supports "a topic distribution indicating to what degree a word belongs to each topic" [18].

Word embeddings are introduced in this paper and it was highlighted that in hate speech detection, it's not necessarily the words themselves but just as importantly the sentence that needs to be identified because it's the sentence content that determines the intent or topic. It was noted that paragraph embeddings show more effectiveness than word embeddings alone as seen from work of researchers Le and Mikolov, 2014 [18].

Authors highlight sentiment analysis and hate speech detection being closely related. Sentiment analysis is a very broad topic in the natural language processing community. Lexical resources are important according to the authors, including lexical weights to the degree of hate speech from a process of adaptive learning [18]. Lexical features are widely used according to authors as a starting point such as POS tagging but, it doesn't improve model per say nor use of n-grams [18].

Dependency parsing was noted for example to identify groups of minority persons such as Jewish or American Indian focusing on the relationship as described in the paper as identifying the minority group with a particular animal or derogatory object, e.g. *nsubj*, *pobj*, [18] etc. The researchers state that it's the dependency of the syntactic relationship that can help detect hate speech [18]. The authors discuss knowledge-based features such as ConceptNet [8] developed in 2004 and modeling techniques included Bayesian Logistic Regression [18].

Overall, the author's research highlights many facets of the components of modeling to detect hate speech including consideration of other forms of medium such as images, or video associated with the hate speech content. In addition, the authors suggest, if possible, to obtain user data, such as with the number of times the user posted hate speech, their follower counts, and even knowing the user's gender as possibly significant. Related, they also considered the time in which the hate speech posts were sent, in connection with geographic related crimes. This topic is considered as crime prediction and while not relevant for hate speech detection per say, it's important to note the correlation.

3.6 How Hate Speech Varies by Target Identity: A Computational Analysis (Yoder, et al., 2022)

Presented at the 26th Conference on Computational Natural Language Learning (CoNLL) in 2022, researchers from Carnegie Mellon University highlight that models trained on different "targeted identities" [23] do not generalize well, thus supporting the question that hate speech detection targets differ. The authors reference the typology from Waseem et al. of implicit and explicit in addition to context and historic properties such as with chat history and or societal observations, respectively, calling hate speech a "phenomenon" [23].

Identities are referenced throughout the research which equate to groups of persons that are targeted for a particular reason such as race, gender, religion, etc. From model outcomes they determined that demographic and or stereotype references are more prevalent and varied within hate speech than reference to societal hierarchies [23]. The authors describe hierarchies as containing power groups although this distinction was not considerable with model performance [23]. Ultimately, the authors believe that each targeted group and their identity will require different feature tactics, stressing that it is the identity of the target group that is the emphasis when building a model.

Hate speech was also described similarly as other researchers while they referenced the "linguistic properties of (hate speech) incitement" [9][22] which would be a great topic to research. The data used in this study included from the Hate Speech Dataset Catalogue hatespeechdata.com which catalogs datasets used in hate speech research. The data was resampled on a "ratio of 30/70 of hate to non-hate" [23] to reduce variance including annotations to better support identification of identities such as transgender, in addition to multiple group identities such as Asian transgender, etc.

A model was trained on a corpus of each type of identity, and "evaluated on corpura targeting other identities" [23]. They referred to this as "cross-identity generalization" [23], where data was sampled and assembled to support multiple identities for training. The authors used a DistilBERT [17] model, trained on 5 epochs, on 10% of the train set showing outcomes of model averages while using a logistic classifier with unigram features and L2 regularization, as a baseline [23].

It was shown that the model did not generalize well across identity classes, such as, Women, Blacks, Asians, and Jews in addition to more general identities as race and ethnicity, gender, and or religion [23].

Dataset	Domain	Original size
Civil Comments (Borkan et al., 2019)	News comments	1999516
Social Bias Inference Corpus (Sap et al., 2020)	Reddit, Twitter, Gab, Stormfront	44781
Kennedy et al. (2020)	YouTube, Twitter, Reddit	39565
HateXplain (Mathew et al., 2021)	Twitter, Gab	20148
Contextual Abuse Dataset (Vidgen et al., 2021)	Reddit	27494
ElSherief et al. (2021)	Twitter	19650
Salminen et al. (2018)	YouTube, Facebook	3222

Table 1: Overview of datasets used in this study. Original size is the number of instances before resampling for experiments. The last 3 datasets are only used in the experiment removing hate toward dominant social groups (section 6.2).

Figure 5: *Yoder, et al., How Hate Speech Varies by Target Identity: A Computational Analysis, pg. 4*

The F1 score was at best 70% and at worst 40%. Utilizing Principle Component Analysis (PCA) to visualize the data, the authors showed that different identities did occupy a separate part of the grid, for example with religion, gender, and race, thus highlighting a potential pattern of demographics [23]. The authors hypothesize that if "demographic categories are particularly discriminative, hate speech classification performance will drop sharply when attempting to generalize across categories" [23].

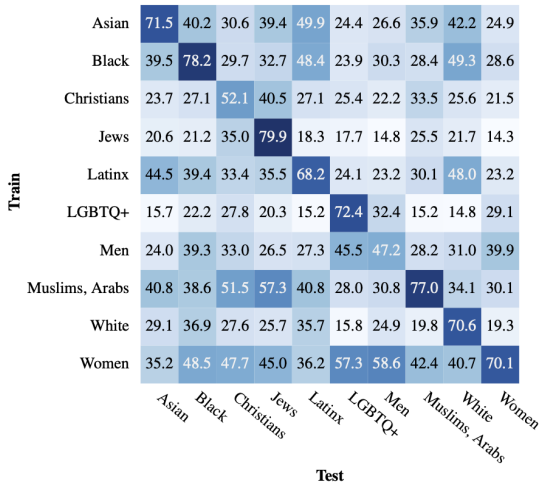


Table 3: Hate speech classification performance (F1 score) across identity-specific corpora

Figure 6: *Yoder, et al., How Hate Speech Varies by Target Identity: A Computational Analysis, pg. 4*

The authors describe conducting multiple studies, where one in which targets concepts such as marginalized, dominant, and other identities also trained on the DistilBERT [17] model and this time, generalization performed better, over demographic identities. Sparse Additive Generative Model (SAGE) [3] is a model that learns using a frequency distribution and smoothing on the corpus, to provide "representative words" [23] supporting hate speech detection. Even with the addition of this feature, generalization across identities did not perform well. The authors correlate this to the importance of dataset selection and identifying skews, toward a particular group will have a detrimental effect on model performance.

3.7 Hate Speech Classifiers Learn Normative Social Stereotypes (Davani, et al., 2022)

Researchers from the University of Southern California, discuss in this research, how social stereotypes can be useful in hate speech classification detection and fairness. It is imperative for the natural language processing research community to understand the possibilities of bias within textual datasets and what data is being overly represented. This directly affects hate speech research studies given over-representation of a particular hate speech targeted group. As one example, the authors claim model outputs, trained from social media platforms, will interpret phrases differently from non-social mediums and may contain biases [1]. It is this bias detection that is ultimately just as important as the hate speech detection in these cases.

As the foundation of a model, datasets either from text or human annotations, can show bias simply from the "mapping of language to numeric representations affected by stereo-typical co-occurrences" [1]. Interestingly, the authors believe that in order to truly understand the language of these biases, it is important to look to "social psychological theories of prejudice, and stereotypes" [1] of which they reference and utilize the Stereotype Content Model (SCM) [5].

Unlike other research papers, this research is heavily focused on the annotation process. The authors preselect 857 US based human participants as annotators, from a specific set of criteria, e.g. male, female, political affiliation, age, and racial diversity, with quality assessment provided, and included a warmth and competence requirement, derived from the SCM model [1].

The first section of this research delves into participant level analysis and findings, from item disagreement using Fleiss, to participant item-level disagreement using a ratio, and group level disagreements as an average. They utilized a number of tools to support understanding annotation tendencies, and overall participant analysis, such as the probabilistic Rasch model, and Item Response Theory (IRT) [1].

This in depth study is a breadth of fresh air amidst a pool of research studies that did not exhibit such rigor, in highlighting the importance of labeled data, in an effort to reduce bias, and the cost of blindly using human annotators without any level of intervention into the types of participant annotators, and their level of responses and annotations.

The authors identified (3) variables for hate labels, along with using the Rasch model tendency detection, and group level disagreement, to support findings [1]. The authors used semantics and tokenization to support what they refer to as "normative social stereotypes" [1], with a pre-trained language model using GloVe which is an unsupervised learning algorithm to obtain word vectors. The authors describe their methodology as obtaining a similarity of each "social group token with entirety of words in dictionaries of warmth and competence in latent vector space" [1] which then "maps each word to a t-dimensional vector trained on word co-occurrences in a corpus of English Wikipedia articles" [1] of which were averaged using a "cosine similarity of the social group token and numeric presentation of the words of the two dictionaries" [1].

In Study 3, the authors focused on understanding whether social stereotypes influence hate speech classifier's predication bias toward those groups" [1]. For this they used three different classification models: 1) BERT, 2) RoBERTa, and 3) Support Vector Machine (SVM) with Term Frequency-Inverse Document Frequency (TF-IDF). Both models 1 and 2 used transformers from huggingface.com. The models trained on a random sample of 80% of the train data, composed of social posts from the Gab Hate Corpus (GHC) where they are able to detect false positives and false negatives of which they obtained ratios [1].

The model F1 scores on average were within the range of 35 to 48 percent. The authors focused their false positives and negatives to detection of social group tokens which are for example, *white*, or *non-binary* [1]. In the entire study they used 8 different social groups. They found BERT false positive ratio to be on average 58% where the classifier was unable to distinguish between hateful and non-hateful speech, for example labeled with *bisexuals*, *homosexuals* [1] but performed better for posts with labels including, *Latino* or *Buddhist* [1]. The BERT and SVM "are more likely to misclassify instances as not containing hate speech when texts mention stereotypically incompetent social groups" [1].

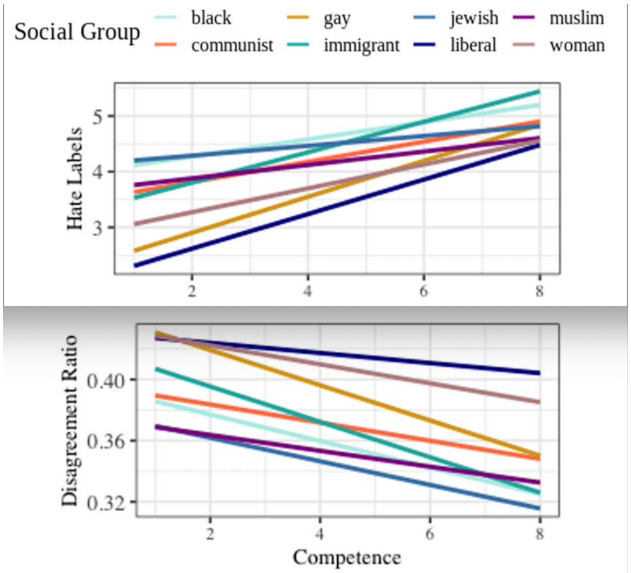


Figure 2: The relationship between the stereotypical competence of social groups and (1) the number of hate labels annotators detected, (2) their tendency to detect hate speech, and (3) their ratio of disagreement with other participants (top to bottom).

Figure 7: *Davani, et al., Hate Speech Classifiers Learn Normative Social Stereotypes, pg. 5*

One of the most significant aspects in this research point to whether or not the human annotators labeled the post given their social group bias. The authors concluded if the annotator level was novice, they tended to want to protect the group if seen as "warm and competent" [1] and if the novice annotator perceived the social group as competent then they were more likely to detect hate speech correctly [1].

In addition, judgements of social groups lacking competence "elicit passive harm" [1]. The authors conclude it is still yet a difficult task to measure bias within hate speech datasets, and or annotations given reference to marginalized groups, and unequal representation of these groups among non-marginalized. All in all, it's critical to have unbiased annotators to further prevent prediction bias in hate speech classification models.

3.8 HateCOT: An Explanation-Enhanced Dataset for Generalizable Offensive Speech Detection via Large Language Models (Ngheim, Daume, 2024)

Both Ngheim, and Daume, researchers from the University of Maryland in connection with Microsoft Research, attempt to make use of large language models (LLMs) to generalize offensive language using a dataset named Hate-related Chain-of-Thought or HateCOT where Chain-of-Thought is used as a prompting technique. HateCOT is a 52,000 object collection of 8 curated datasets from other researchers such as HateCheck (2021) [16], HateXplain (2021) [10], and Latent_Hate (2021) [4].

This dataset includes text, a corresponding hate speech label, and GPT3.5-Turbo explanation. The purpose of using HateCOT is to pre-train a small language model (SLM) on a targeted group. The authors state there are existing pretrained hate speech labeled transformer models being used in the research community, such as HateBERT, and fBERT, with varying levels of performance [14]. The authors cite cyberbullying, sexist, racist,

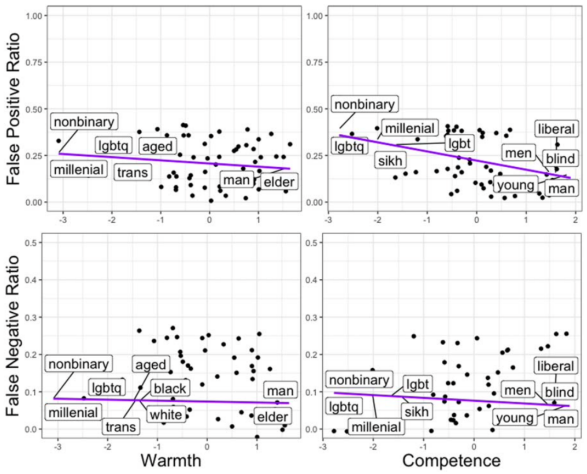


Figure 6: Social groups' higher stereotypical competence and warmth is associated with lower false positive and negative predictions in hate speech detection.

Figure 8: *Davani, et al., Hate Speech Classifiers Learn Normative Social Stereotypes, pg. 10*

and hate as semantically similar but distinct concepts [14]. Using Llama 2 Chat-FH or Llama 7B, a decoder only model, and low rank adaptation (LoRA) [7] techniques the initial study is performed on a dataset of simply neutral versus non-neutral labels of which Llama 7B achieves 64% F1 score with one explanation provided, while achieves only slightly higher when up to 3 explanations are provided, whereas anything after 4, degrades in performance. The LoRA technique takes in model weights, and "injects trainable rank decomposition matrices into each layer of the transformer, reducing trainable parameters for a task" [7].

Figure 2: Performance resutls of LLMs on test sets in various settings.

Dataset	Best Model @K=256	F1 @K=256	F1 Base + Full	F1 %	Data Size @K=256	Data Size Full	Data %
HateCheck	LLAMA 13B	0.95	0.99	96%	512	1,864	27%
HateXplain	LLAMA 13B	0.64	0.72	89%	768	12,088	6%
Latent_Hate	LLAMA 13B	0.66	0.64	103%	768	11,460	7%
Implicit_Hate	COT-T5-XL	0.56	0.38	147%	1,536	2,707	57%

Figure 9: *Davani, et al., HateCOT: An Explanation-Enhanced Dataset for Generalizable Offensive Speech Detection via Large Language Models, pg. 7*

Further experimentation includes testing using other models such as Llama 13B, OPT-IML (1.3B parameters) encoder only, Flan-T5-L (780M parameters) encoder-decoder, and COT-T5-XL (1.8M parameters) variant of the previous. The significant point here is to note model size, considering parameters. A zero shot classification baseline study outcome showed that it was larger models that performed better in generating explanations only fine tuning with HateCOT while pre training on HateCOT produced favorable results across all models [14] while COT-style prompting favors larger models. The authors validate in this research, that in context learning (ICL) can be used to boost model performance. ICL means further context was provided to

the prompt [14]. Please see Appendix A for visual of F1 comparison of model output.

4 Discussion

In order to move hate speech detection research forward, we have to continue asking the right questions, and partnering with companies and or government agencies to work toward a scalable solution that mitigates the societal harm of hate speech, on targeted social groups and or individuals. We should consider a primary goal as, building language models, that can detect hate speech toward groups and or individuals within text online, and at scale. Please see discussion questions below:

- What are proven modeling techniques to detect different variations of hate speech and how can academic researchers combine efforts to maximize results?
- Given historical to current research, what are critical features required to support hate speech detection and what is required for high-quality model output?
- Algorithmically and or within neural modeling, what are the latest proven techniques for deep learning and hate speech detection with generalization? Which has proven greater accuracy in the present and are traditional classification models still useful in certain cases?
- What does hate speech detection deep learning architecture look like and what architecture has proven most successful?
- What is critical for natural lanaguage processing subtasks to consider, in order to prevent prediction bias or bias within human labeled datasets?

5 Challenges

There were many challenges stated by the researchers in each of the studies above, with some as a pattern, such as with biases in human labeling, while others were more of a technological constraint, such as with a model's inability to generalize hate speech within targeted identity groups, as Yoder, et al., termed as. There are still fairly fundamental needs that modeling hasn't been able to answer. For example, understanding implicit hate speech language, which we read about in Waseem, et al. When speaking about implicit, it refers to not necessarily direct or specific hateful phrases, but, rather language that is hidden as indirect sarcasm or contempt or belittling.

Early before transformer architecture was introduced by Vaswani, et al., in 2017, the academic natural language processing research community were utilizing techniques such as POS tagging, dependency parsing, word sense disambiguation, embeddings, regressive modeling with TF-IDF, etc., to support different forms of hate speech classification. The challenges here remained in the type of tagging or labeling, required to detect the distinct targeted groups, and the model's inability to generalize across all targeted groups.

This relates to the challenges of human annotated data. In the Davidson, et al., research, it was found that annotators were labeling some content as offensive only when referring to sexism, and hateful for homophobic or racist. While the innovative work from Davani, et. al we learned an entirely new way to interact and integrate with annotators which included a rigorous selection process, and quality check that supported multiple layers of interpreting model output outcome such as with warmth and competence.

6 A path forward and proposal

Looking forward, there are many hate speech detection research topics that can be taken further, such as with understanding the various pretrained encoder decoder large language models (LLMs), as seen with Ngheim and Daume's HateCOT research and their ability to answer the problem of generalization of targeted hate speech groups or individuals. Hate speech detection has gone through many technical iterations, more generally, beginning with POS tagging, tree parsing, embeddings, and regressive modeling, and now adding neural nets.

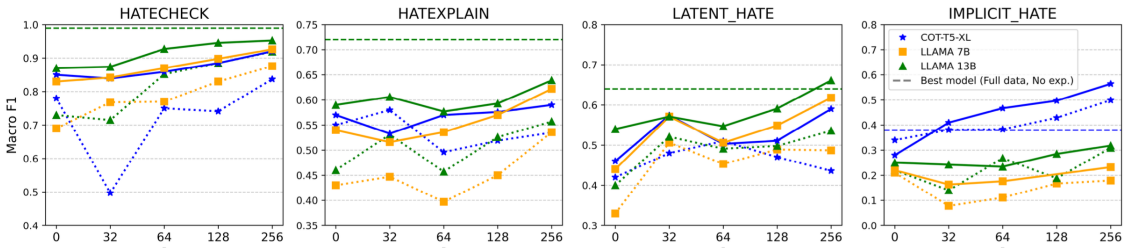
Researchers have expressed hate speech as a language, with varying opinions on what classifies hate speech. We saw that researchers considered studying hate speech along side abusive and or offensive language, and yet offensive language was found to be "common place" [1]. Additionally, we also have to consider explicit and or implicit hate speech language and the difficulties of a model detecting implicit hate speech.

Hate speech effects targeted cultural, racial, ethnic, social, and many other groups, which in many cases are minority labeled groups, or individuals. The breadth of targeted groups points to the importance of cross-functional research as such as with sociology or psychology in understanding what is it in the language that "incites" hate [22] and what biases or stereotypes, exist in society or culture, that ignite this hateful language toward others. What is the psychology of an individual who generates hate speech online?

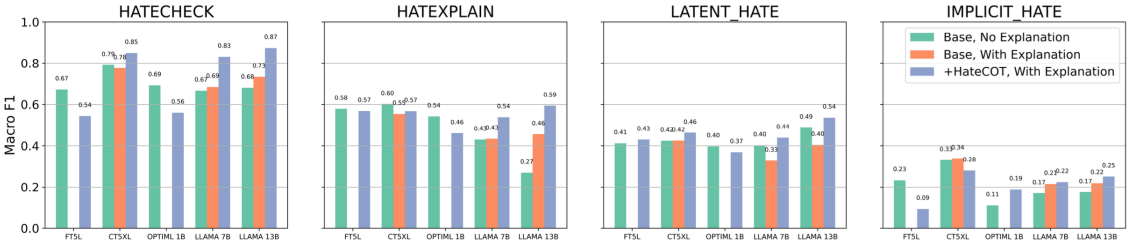
As a path forward, it would be recommended for academic researchers to consider the harmful effects that hate speech online has within society and on the group and or individuals. We need to work together with global companies and or government agencies to ensure safeguards are enabled online to mitigate harm from hate speech. Online can refer to social media platforms, chat rooms, comments, email and or many other textual based online environments.

How can we build a scalable language model, that generalizes well, understands targeted hate speech groups, accurately detects hateful language toward others, and is intelligent enough to prevent against bias, all while dramatically decreasing the negative effects of hate speech online? Let's consider this as our goal and path forward.

A Appendix



(b) Macro F1 scores for models in zero-shot setting with explanation after K-shot in-domain finetuning at various values of K. Dashed line represents finetuned base models, solid line represents models pre-trained on *HateCOT*, then in-domain finetuned. For each dataset, the horizontal dashed line represents the base version of the best performing model at K=256 which is finetuned using the entire training data *without* any rationale for comparison, denoted as *Best model (Full data, No. exp)*.



(a) Macro F1 scores of LLMs in zero-shot setting using 3 configurations. *Base* refers to out-of-the-box models, *+HateCOT* denotes their pretrained counterpart on our dataset. *FT5L*: Flan-T5-L, *CT5XL*: COT-T5-XL. Results for Base Flan-T5-L and OPTIML models for *With Explanation* settings omitted to reflect their inability to generate explanation according to the prompt.

Figure 10: *Davani, et al., HateCOT: An Explanation-Enhanced Dataset for Generalizable Offensive Speech Detection via Large Language Models, pg. 7*

References

[1] Aida Mostafazadeh Davani et al. "Hate Speech Classifiers Learn Normative Social Stereotypes". In: *Transactions of the Association for Computational Linguistics* 11 (2023), pp. 300–319. DOI: 10.1162/tacl_a_00550. URL: https://aclanthology.org/2023.tacl-1.18.

[2] Thomas Davidson et al. "Automated Hate Speech Detection and the Problem of Offensive Language". In: *CoRR* abs/1703.04009 (2017). arXiv: 1703.04009. URL: http://arxiv.org/abs/1703.04009.

[3] Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. "Sparse additive generative models of text". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Bellevue, Washington, USA: Omnipress, 2011, pp. 1041–1048. ISBN: 9781450306195.

[4] Mai ElSherief et al. "Latent Hatred: A Benchmark for Understanding Implicit Hate Speech". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens et al. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 345–363. DOI: 10.18653/v1/2021.emnlp-main.29. URL: https://aclanthology.org/2021.emnlp-main.29.

[5] Susan T. Fiske et al. "A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition". In: *Journal of Personality and Social Psychology* 82.6 (2002), pp. 878–902. DOI: 10.1037/0022-3514.82.6.878. URL: https://doi.org/10.1037/0022-3514.82.6.878.

[6] Microsoft Research Natural Language Processing Group. Redmond, WA. URL: https://www.microsoft.com/en-us/research/group/natural-language-processing/.

[7] Edward J. Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models". In: *CoRR* abs/2106.09685 (2021). arXiv: 2106.09685. URL: https://arxiv.org/abs/2106.09685.

[8] Hugo Liu and Push Singh. "ConceptNet — A Practical Commonsense Reasoning Tool-Kit". In: *BT Technology Journal* 22 (2004), pp. 211–226. URL: https://api.semanticscholar.org/CorpusID:266028051.

[9] Alexandria Marsters. "When Hate Speech Leads to Hateful Actions: A Corpus and Discourse Analytic Approach to Linguistic Threat Assessment of Hate Speech". In: *Georgetown University-Graduate School of Arts & Sciences* (2019). URL: https://api.semanticscholar.org/CorpusID:210494076.

[10] Binny Mathew et al. "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection". In: *CoRR* abs/2012.10289 (2020). arXiv: 2012.10289. URL: https://arxiv.org/abs/2012.10289.

[11] Toby Mendel. "Does International Law Provide for Consistent Rules on Hate Speech?" In: *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. Ed. by Michael Herz and PeterEditors Molnar. Cambridge University Press, 2012, pp. 417–429.

[12] "United Nations". *"International Covenant on Civil and Political Rights"*. 1966. URL: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights.

[13] "United Nations". *"Universal Declaration of Human Rights"*. 1966. URL: https://www.un.org/sites/un2.un.org/files/2021/03/udhr.pdf.

[14] Huy Nghiem and Hal Daumé Iii. "HateCOT: An Explanation-Enhanced Dataset for Generalizable Offensive Speech Detection via Large Language Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 5938–5956. DOI: 10.18653/v1/2024.findings-emnlp.343. URL: https://aclanthology.org/2024.findings-emnlp.343.

[15] John T. Nockleby. "Hate Speech in Context: The Case of Verbal Threats". In: *Buffalo Law Review* 42 (1994), p. 653. URL: https://api.semanticscholar.org/CorpusID:221963969.

[16] Paul Röttger et al. "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)*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 41–58. DOI: 10.18653/v1/2021.acl-long.4. URL: https://aclanthology.org/2021.acl-long.4.

[17] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: 2020. arXiv: 1910.01108 [cs.CL]. URL: https://arxiv.org/abs/1910.01108.

[18] Anna Schmidt and Michael Wiegand. "A Survey on Hate Speech Detection using Natural Language Processing". In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Ed. by Lun-Wei Ku and Cheng-Te Li. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1–10. DOI: 10.18653/v1/W17-1101. URL: https://aclanthology.org/W17-1101.

[19] Ellen Spertus. “Smokey: automatic recognition of hostile messages”. In: *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*. AAAI’97/IAAI’97. Providence, Rhode Island: AAAI Press, 1997, pp. 1058–1065. ISBN: 0262510952.

[20] William Warner and Julia Hirschberg. “Detecting Hate Speech on the World Wide Web”. In: *Proceedings of the Second Workshop on Language in Social Media*. Ed. by Sara Owsley Sood, Meenakshi Nagarajan, and Michael Gamon. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 19–26. URL: https://aclanthology.org/W12-2103.

[21] Zeerak Waseem et al. “Understanding Abuse: A Typology of Abusive Language Detection Subtasks”. In: *Proceedings of the First Workshop on Abusive Language Online*. Ed. by Zeerak Waseem et al. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 78–84. DOI: 10.18653/v1/W17-3012. URL: https://aclanthology.org/W17-3012.

[22] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. “Detection of Abusive Language: the Problem of Biased Datasets”. 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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 602–608. DOI: 10.18653/v1/N19-1060. URL: https://aclanthology.org/N19-1060.

[23] Michael Yoder et al. “How Hate Speech Varies by Target Identity: A Computational Analysis”. In: *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*. Ed. by Antske Fokkens and Vivek Srikumar. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 27–39. DOI: 10.18653/v1/2022.conll-1.3. URL: https://aclanthology.org/2022.conll-1.3.