Investigating Machine Learning Models in Hate Speech Detection

Mubashir Shaikh[1], Nihal Momin[2]
[1,2]*Student, SIES College of Commerce and Economics, Mumbai*
doi.org/10.64643/IJIRTV12I12-202315-459

Abstract*—Hate speech detection is one of the harder problems in natural language processing (NLP). Machine learning (ML) models tend to do well on average but still get a lot of borderline cases wrong. In this paper, we try to understand why that happens by doing an exploratory data analysis (EDA) on the Davidson dataset, which is a popular hate speech dataset. We create a new feature called a severity score that turns the original three class labels into a continuous number, and we also measure how much annotators disagreed with each other when labelling the tweets. We find that disagreement and severity are only weakly linked (Spearman $\rho = 0.29$), which means content being moderately harmful does not automatically mean annotators will disagree about it. We then look more closely at which tweets caused the most disagreement and find that the issue is not really the words used, it is more about the sentence structure. Ambiguous tweets tend to use indirect, roundabout phrasing, while tweets that everyone agreed on tend to be more direct and straightforward. This points to a real limitation of standard ML models that only look at word counts, since they completely miss these structural differences.

***Index Terms*—annotator disagreement, Davidson dataset, machine learning limitations, natural language processing, part-of-speech analysis, semantic ambiguity hate speech detection, severity scoring.**

TABLE I *List of Abbreviations*

Abbreviation	Full Form
NLP	Natural Language Processing
ML	Machine Learning
EDA	Exploratory Data Analysis
TF-IDF	Term Frequency-Inverse Document Frequency
POS	Part-of-Speech
LOWESS	Locally Weighted Scatterplot Smoothing
LR	Logistic Regression
HSC	Hate Speech Classification

I. INTRODUCTION

Social media platforms have made it easier than ever for people to share opinions online, but this has also led to a rise in harmful content like hate speech and offensive language. Automatically detecting this kind of content using machine learning has become an active research area, with approaches spanning n-gram features, word embeddings, and deep neural classifiers [7]. Early foundational studies established criteria for identifying racist content on Twitter and showed that annotator demographics significantly shape labelling decisions [20].

Offensive language is not always obvious, it may be wrapped in sarcasm, coded phrasing, or context-dependent meaning [14]. Hate speech detection is especially difficult because it is inherently subjective and context-dependent, properties that standard NLP benchmarks are not designed to capture [6]. Research has also shown that annotator insensitivity to dialect can introduce racial bias, with tweets in African American English being mislabelled as offensive far more often than equivalent standard English tweets [11]. When even human annotators disagree on whether something is hate speech, models trained on their labels naturally inherit that confusion [13]. Not much research has examined exactly when and why that disagreement happens, and whether it is tied to how harmful the content is, or to something else entirely.

II. LITERATURE REVIEW

Saleh et al. (2023) propose two approaches to detect hate speech, a custom HSW2V+BiLSTM model (93% F1) and fine-tuned BERT (96% F1). The gap was that no prior work had combined domain-specific embeddings with BiLSTM or applied BERT to binary

hate speech classification, demonstrating that domain-specific training is effective even with smaller datasets [1].

Parihar et al. (2021) review ML and deep learning methods including SVM, LSTM, CNN, and BERT for detecting hate speech across multiple languages. They highlight key challenges such as data imbalance and poor annotation agreement and call for future research on multilingual and domain-specific models [2].

Garg et al. (2023) survey unintended biases in toxic speech detection, categorizing them by source (sampling, lexical, annotation) and target (racial, gender, psychographic) with mitigation strategies. They develop a bias taxonomy and recommend continuous pipeline-wide monitoring rather than one-time fixes [3].

Biere (2018) applies a CNN to classify tweets into hate speech, offensive language, or neither, achieving 91% accuracy on the Davidson dataset. Despite this, the model still misclassified 80% of actual hate tweets due to class imbalance, demonstrating that three-class separation remains the core challenge [4].

Florio et al. (2020) show that hate speech detection models degrade over time as social media language evolves. Incremental retraining helps, but stale data hurts BERT specifically, no prior work had examined this time-based degradation [5].

Fortuna et al. (2022) argue that common NLP benchmarks and inter-annotator agreement metrics are poorly suited for hate speech detection, since hate speech is inherently subjective and context dependent. The paper calls for ethics-centred, context-aware evaluation frameworks [6].

Schmidt and Wiegand (2017) survey NLP approaches to hate speech detection, covering n-grams, word embeddings, sentiment features, and SVM classifiers. They identify the lack of a standard benchmark dataset and multilingual coverage as the key barrier to comparability across studies [7].

Chaudhary et al. (2021) presents a unified framework classifying NLP-based hate speech methods into reactive (post-detection) and proactive (preemptive) strategies. They argue that treating hate speech purely as a classification problem is insufficient, and that future work must adopt a broader, prevention-focused approach covering the full content moderation pipeline [8].

Yin and Zubiaga (2021) review obstacles to cross-dataset generalisation in hate speech detection. Models trained on one dataset perform poorly on another due to annotation style and data distribution differences; domain-adaptive training is recommended as the key remedy [9].

Mathew et al. (2021) introduce HateXplain, annotated with classification labels, target communities, and human rationales. Models scoring well on accuracy frequently fail on explainability metrics, and training with rationales reduces unintended bias, showing accuracy alone is insufficient [10].

Sap et al. (2019) show that annotator insensitivity to dialect introduces racial bias: models mislabelled 46% of non-offensive African American English tweets as offensive vs. only 9% in standard English, directly relevant to why identity-based terms drive disagreement in our dataset [11].

Dehghan and Yanikoglu (2025) show that even detailed annotation guidelines cannot eliminate interpretive disagreement, supporting our finding that disagreement in the Davidson dataset reflects genuine semantic ambiguity rather than annotator error [12].

Rottger et al. (2022) contrast descriptive annotation (personal subjectivity, $\kappa = 0.20$) with prescriptive annotation (strict guidelines, $\kappa = 0.78$), explaining why even a prescriptively annotated dataset like Davidson still yields disagreement on borderline cases [13].

ElSherief et al. (2021) introduce an implicit hate speech corpus, showing that most models trained on explicit language fail on hate conveyed through sarcasm and insinuation, directly aligned with our finding that ambiguous tweets rely on indirect structures [14].

Rawat et al. (2024) review hate speech detection methods published between 2018 and 2023, spanning ML, deep learning, ensemble, GNN, and GCN approaches. They identify class imbalance as a persistent unsolved challenge across all categories. This directly supports our observation about the Davidson dataset's imbalance between hate speech and offensive language labels and its downstream effect on model reliability [15].

Jahan and Oussalah (2023) review BERT, GPT, and LSTM across multilingual hate speech datasets, noting

that costly human annotation creates a persistent bottleneck, contextualising the TF-IDF and bag-of-words limitations this paper examines [16].

Trivedi et al. (2025) finds that even fine-tuned ELECTRA (F1 = 0.8980 on 1.2M samples) still struggles with sarcasm, coded language, and label noise, reinforcing that the ambiguous cases confusing human annotators remain unsolved at scale [17].

Uma et al. (2021) survey disagreement-aware training strategies including soft labels and ensemble methods, directly supporting our recommendation to train on annotator distributions rather than majority-vote labels [18].

Antypas and Camacho-Collados (2023) conduct a large-scale cross-dataset evaluation showing that hate speech models inherently learn dataset-specific biases, and that combining training data from diverse sources significantly improves generalisation — directly supporting our argument that the lexical patterns found in the Davidson dataset reflect broader training-data limitations [19].

Waseem and Hovy (2016) annotate 16,000 tweets for racism and sexism, finding that annotator demographics significantly affect labelling, establishing the foundational basis for later disagreement-aware research and explaining annotation inconsistencies in the Davidson dataset [20].

III. METHODOLOGY

A. Dataset Description

We used the Davidson dataset, which was originally put together by Davidson et al. (2017) and has since become one of the most commonly used datasets for hate speech research. It contains 24,783 tweets that were collected from Twitter and labelled by human annotators through the CrowdFlower platform. Each tweet was given one of three labels:

- Label 0: Hate Speech: tweets that are clearly targeting someone or a group with hateful intent.
- Label 1: Offensive Language: tweets that use offensive or vulgar language but are not quite hate speech.
- Label 2: Neither: tweets that are neither hateful nor offensive.

Crucially, this dataset retains the full annotator breakdown per tweet, not just the majority-vote label, allowing disagreement to be measured directly. Class imbalance across the three labels has been identified as a persistent challenge for hate speech classifiers trained on this and similar datasets [15], and distinguishing hate speech from offensive language is widely regarded as the hardest sub-problem in three-class detection [4]. Prior reviews have also documented that poor annotation agreement compounds these challenges, particularly for multilingual and domain-specific settings [2].

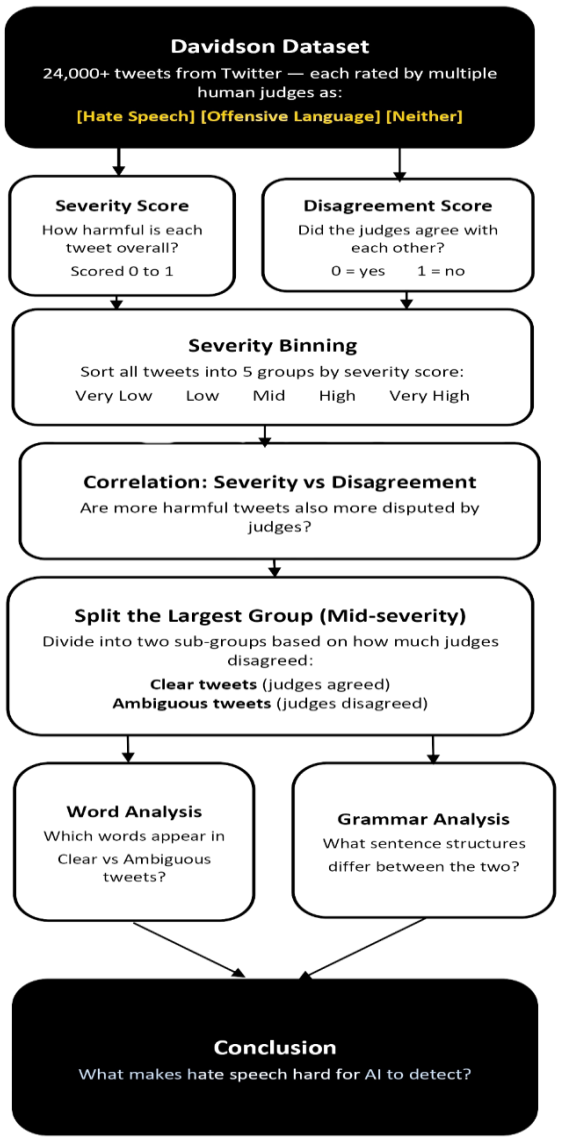

Fig. 1. Methodology pipeline overview, illustrating the full data flow from the Davidson dataset through feature engineering, statistical analysis, and ML-based classification to the final conclusions.

B. Feature Engineering: Severity Score

The original labels only give us three categories, which is quite limiting. To get a better sense of how harmful each tweet was perceived to be, we created a severity score using the following formula:

severity = (2×hs + 1×ol + 0×n) / count

Here, hate_speech, offensive_language, and neither are the number of annotators who picked each label, and count is the total number of annotators for that tweet. The weights (2, 1, 0) are just reflecting the idea that hate speech is more severe than offensive language, which is in turn more severe than neither. This gives a score between 0 and 2, which we then normalised to be between 0 and 1 by dividing by 2.

To check that the score was working as expected, we looked at the average severity per original label. The results were exactly what we would hope for, the mean severity went up from neither (0.12) to offensive language (1.03) to hate speech (1.70), confirming the score reflects the right order. This kind of continuous representation helps surface the annotation biases that categorical schemes tend to obscure, including lexical and sampling biases that can unfairly penalise marginalised communities [3].

C. Annotator Disagreement Metric

We defined a per-tweet disagreement score:

disagreement = 1 − max (hs, ol, n) / count

A score of 0 means all annotators agreed on the same label. A higher score means more of them disagreed. This lets us figure out which tweets were controversial or unclear and compare them to tweets that everyone agreed on. This is important because, as Dehghan and Yanikoglu (2025) observe, even detailed annotation guidelines cannot fully eliminate interpretive disagreement, the subjectivity of hate speech means that well-trained annotators will still diverge on borderline content [12].

D. Severity Binning

The normalised severity score was split into five equal bins:

TABLE II *Severity bins and their corresponding ranges.*

Bin Label	Severity Range
very_low	[0.0, 0.2]
low	(0.2, 0.4]
mid	(0.4, 0.6]
high	(0.6, 0.8]
very_high	(0.8, 1.0]

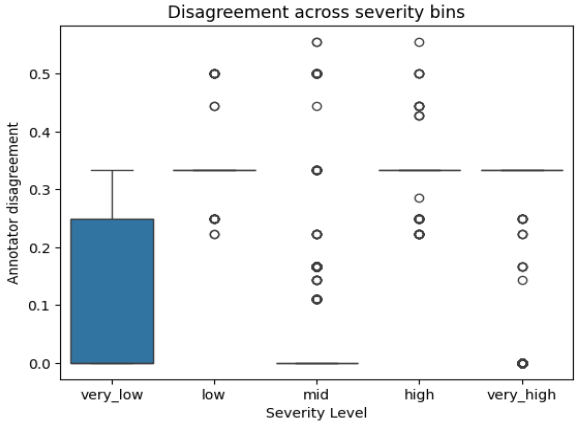

Fig. 2. Distribution of hate speech severity scores across severity bins.

E. Spearman Correlation

Spearman rank correlation (chosen over Pearson given non-normal distributions) was used to test whether severity and disagreement are related.

F. LOWESS Smoothing

A LOWESS curve (smoothing fraction 0.2) was fitted to the disagreement vs. severity scatter to reveal the overall non-parametric trend.

G. Lexical Analysis

Within the mid-severity bin, tweets were split into two groups for lexical comparison:

- Ambiguous: tweets with disagreement > 0.30 (71 tweets)
- Clear: tweets with disagreement < 0.05 (14,347 tweets)

We then trained a Logistic Regression model on TF-IDF features (using unigrams and bigrams) to try to tell the two groups apart. We looked at which words had the highest coefficients for each group and also

checked whether any words appeared in both lists. These lexical patterns are worth examining carefully, since structural and lexical biases in training data have been shown to be a primary driver of poor cross-dataset generalisation [9].

H. Syntactic Analysis

To complement the lexical findings, we analysed sentence structure. We used spaCy (en_core_web_sm) to tag each word in a tweet with its part-of-speech (POS) label, like NOUN, VERB, ADJ, etc. We then treated these POS tag sequences like text and trained another Logistic Regression model using bigrams to 4-grams of POS tags. This helped us see whether ambiguous tweets are built differently from clear ones, structurally speaking. This structural approach is motivated by the documented limitations of bag-of-words representations, which discard word order and therefore cannot encode grammatical context or syntactic framing [16].

IV. OBSERVATIONS

A. Distribution of Severity

The distribution of normalised severity scores is clearly bimodal. Most tweets had a very low severity score (close to 0), which makes sense because a lot of them were labelled as Neither by all annotators. There was also a noticeable cluster around the 0.4–0.6 range, corresponding mostly to offensive language tweets. Very few tweets had a severity of 1.0 (unanimous hate speech), but they were still present.

This imbalance means ML models have far fewer hate speech examples to learn from, making them systematically less reliable on the rarer class.

B. Relation between Severity and Disagreement

The Spearman correlation between severity and disagreement gave us this result:

$\rho = 0.291$, $p < 0.001$.

The LOWESS curve made this clearer. As severity goes up from 0, disagreement rises quickly, then flattens out, then dips around severity = 0.5, rises again, and finally drops back to near-zero at severity = 1.0. At both extremes, annotators agreed with each other: either the tweet was clearly harmless, or it was clearly hate speech. The messy middle is where things get complicated. This pattern also implies that model performance on this dataset is likely to degrade as social media language evolves, since the borderline cases are the most volatile over time [5].

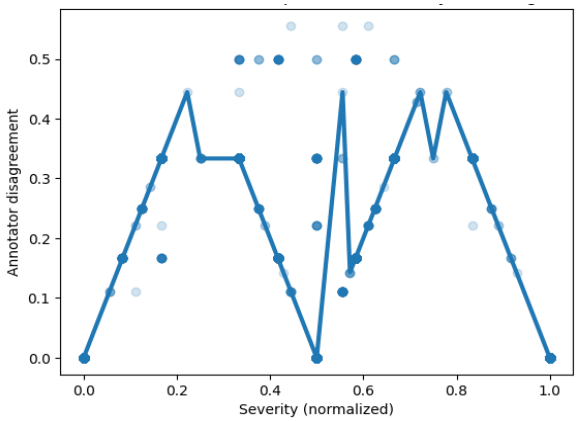

Fig. 1. LOWESS-smoothed scatter plot of annotator disagreement against normalized severity score (smoothing fraction = 0.2).

The dip around severity = 0.5 is interesting. It seems to represent tweets that everyone agreed were offensive language, not hateful enough to be hate speech, but clearly offensive enough that annotators did not argue about it. So even in the middle of the scale, there is a pocket of clarity surrounded by ambiguity.

C. Disagreement Across Severity Bins

Looking at the bin-level statistics gave us a bit of a surprise:

TABLE III *Disagreement statistics per severity bin.*

Severity Bin	Mean	Std. Dev.	Count
very_low	0.086	0.144	3,924
low	0.334	0.017	1,646
mid	0.007	0.041	14,833
high	0.333	0.014	3,078
very_high	0.265	0.134	1,302

The mid bin had the lowest mean disagreement (0.007), even though it was the largest group with nearly 15,000 tweets. This seems counterintuitive at first. But this is because most mid-severity tweets are actually very clear: annotators all agreed they were offensive language. When we flagged tweets with disagreement > 0.30 as ambiguous, only 71 out of 14,833 mid-severity tweets (about 0.48%) were

flagged. So, the mid bin is mostly made up of clear cases, with a small number of genuinely tricky ones inside it. This finding echoes the broader observation that classification accuracy alone is a poor proxy for model quality, models may score well overall while failing on the small, hard cases that matter most [10].

D. Words Associated with Ambiguity

The TF-IDF classifier achieved 99.5% training accuracy, which shows that ambiguous and clear tweets do use noticeably different words. The words most linked to ambiguous tweets included: queer, negro, dykes, retard, niggah, nicca, and trash. The words most linked to clear tweets included: bitch, bitches, pussy, fuck, fucking, hoe, hoes, and shit. And when we checked for overlap between the two lists, there was none, zero words in common.

This is a pretty meaningful result. The words associated with clear tweets are common profanity and sexual slurs that most people would agree are offensive. Annotators do not argue about them much. The words associated with ambiguous tweets are identity-based terms, some of which are slurs, some of which have been reclaimed by certain communities. This zero-overlap finding also demonstrates how structural and lexical biases embedded in training data can cause models trained on one type of content to fail when they encounter the other [19].

E. Sentence Structures Associated with Ambiguity

The POS n-gram classifier also did very well, achieving 99.79% training accuracy. This confirms that ambiguous and clear tweets are not just using different words, they are also built differently at a grammatical level.

To explain what these grammatical patterns look like, think of POS tagging as replacing every word in a sentence with its grammatical role. For example, the sentence "She quickly called John" becomes PRON ADV VERB PROPN. The model then looks for recurring sequences of these role labels to find structural patterns that separate ambiguous from clear tweets.

The POS patterns most linked to ambiguous tweets included:

- SCONJ ADP

A subordinating conjunction followed by a preposition.

Example: "That's impressive, even for someone like them.", The word "even" kicks off a dependent clause ("even for someone like them") that carries the negative judgment. The insult is not stated outright; it is tucked inside a side clause. A model that only looks at individual words sees "impressive" and might think the sentence is positive.

- PROPN PROPN VERB PROPN

A structure heavy with proper nouns, possibly referring to a group indirectly.

Example: "Black Twitter destroyed America again.", This sentence names two entities (Black Twitter, America) and uses a strong verb. Whether this reads as hateful or as commentary depends heavily on context, tone, and whether the speaker belongs to the referenced group. The sentence structure alone does not make the intent clear.

The POS patterns most linked to clear tweets included:

- NOUN PRON AUX

A direct nominal predicates.

Example: "That girl is trash.". Short, direct, no ambiguity. A noun, a pronoun, and an auxiliary verb put the target and the insult right next to each other. Annotators have very little to argue about here.

- PRON VERB NOUN

A standard subject–verb–object sentence.

Example: "I hate foreigners.". The classic direct structure: who, does what, to whom. No room for misinterpretation. Annotators essentially always agree on tweets like this.

In short, clear offensive tweets say what they mean directly. Ambiguous tweets say it indirectly, through subordinate clauses, vague references, or implied targets. This kind of structural indirection is something that a bag-of-words model simply cannot pick up on, because it throws away all word order information. Notably, even fine-tuned transformer models such as ELECTRA, which achieve state-of-the-art F1 scores on large hate speech datasets, continue to struggle with exactly these kinds of sarcastic, coded, and structurally indirect examples [17].

V. CONCLUSION

Our analysis of the Davidson dataset yields four main findings:

- Turning the original three-class labels into a continuous severity score revealed that most tweets are either clearly harmless or clearly offensive, with a smaller group in the middle that is more complex.
- Annotator disagreement is only weakly correlated with severity ($\rho = 0.291$). This means that being in the moderate range of severity does not automatically make a tweet ambiguous, most mid-severity tweets are agreed upon by annotators.
- Ambiguous tweets tend to use identity-based or reclaimed terms whose meaning is context-dependent, while clear tweets use straightforward profanity. There was no overlap at all between the top word features for each group.
- Ambiguous tweets use indirect grammatical constructions where the offensive content is buried or implied, while clear tweets are more direct. A bag-of-words ML model ignores sentence structure entirely, which means it fundamentally cannot learn this difference.

Based on all of this, we think future work should look into models that actually consider sentence structure, like dependency parsers or transformer-based models that implicitly encode word order and context, building on recent work showing that domain-specific BERT fine-tuning can achieve strong results on hate speech classification [1]. It would also be worth experimenting with training on soft labels (where the model learns from the full distribution of annotator opinions rather than just the majority vote), an approach that Uma et al. (2021) show consistently outperforms gold-label training when substantial disagreement data is available [18]. Additionally, hate speech modelling should move beyond detection alone toward broader prevention-focused and ethically grounded frameworks that address the full content moderation pipeline [8]. Using the disagreement score to flag uncertain samples for extra human review rather than forcing a hard prediction is another practical step in this direction. Hate speech detection will not improve just by making models bigger, they need to become more aware of how language actually works.

REFERENCES

[1] H. Saleh, A. Alhothali, and K. Moria, "Detection of hate speech using BERT and hate speech word embedding with deep model," *Applied Artificial Intelligence*, vol. 37, no. 1, 2023, doi: 10.1080/08839514.2023.2166719.

[2] A. S. Parihar, S. Thapa, and S. Mishra, "Hate speech detection using natural language processing: Applications and challenges," in *Proc. 5th Int. Conf. Trends in Electronics and Informatics (ICOEI)*, 2021.

[3] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, "Handling bias in toxic speech detection: A survey," *ACM Computing Surveys*, vol. 55, no. 13s, Art. no. 264, 2023, doi: 10.1145/3580494.

[4] S. Biere, "Hate speech detection using natural language processing techniques," Master's thesis, Faculty of Science, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, 2018.

[5] K. Florio, V. Basile, M. Polignano, P. Basile, and V. Patti, "Time of your hate: The challenge of time in hate speech detection on social media," *Applied Sciences*, vol. 10, no. 12, p. 4180, 2020, doi: 10.3390/app10124180.

[6] P. Fortuna, M. Domínguez, L. Wanner, and Z. Talat, "Directions for NLP practices applied to online hate speech detection," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[7] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. 5th Int. Workshop Natural Language Processing for social media*, 2017.

[8] M. Chaudhary, C. Saxena, and H. Meng, "Countering online hate speech: An NLP perspective," *arXiv preprint* arXiv:2109.02941, 2021.

[9] W. Yin and A. Zubiaga, "Towards generalisable hate speech detection: A review on obstacles and solutions," *PeerJ Computer Science*, 2021, doi: 10.7717/peerj-cs.598.

[10] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A benchmark dataset for explainable hate speech detection," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 35, no. 17, pp. 14867–14875, 2021.

[11] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proc. 57th Annu. Meeting Assoc. Computational Linguistics (ACL)*, pp. 1668–1678, 2019.

[12] A. Dehghan and B. Yanikoglu, "Annotator disagreement in hate speech detection," *arXiv preprint* arXiv:2501.05495, 2025.

[13] P. Röttger, B. Vidgen, D. Hovy, and J. Pierrehumbert, "Two contrasting data annotation paradigms for subjective NLP tasks," in *Proc. 2022 Conf. North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 175–190, 2022.

[14] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, and D. Yang, "Latent hatred: A benchmark for understanding implicit hate speech," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 345–363, 2021.

[15] S. Rawat, V. Soni, and P. Jain, "Hate speech detection: Techniques and challenges," *Artificial Intelligence Review*, vol. 57, Art. no. 97, 2024, doi: 10.1007/s10462-024-10728-9.

[16] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, p. 126232, 2023, doi: 10.1016/j.neucom.2023.126232.

[17] S. Trivedi, A. Agrawal, and S. Sharma, "Comparative exploration of deep learning architectures for hate speech detection," *arXiv preprint* arXiv:2502.09338, 2025.

[18] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio, "Learning from disagreement: A survey," *Journal of Artificial Intelligence Research*, vol. 72, pp. 1385–1470, 2021, doi: 10.1613/jair.1.12752.

[19] D. Antypas and J. Camacho-Collados, "Robust hate speech detection in social media: A cross-dataset empirical evaluation," in *Proc. 7th Workshop on Online Abuse and Harms (WOAH)*, Toronto, Canada, pp. 231–242, 2023.

[20] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. NAACL Student Research Workshop*, pp. 88–93, 2016.