When Words Wear Masks: Detecting Malicious Intents and Hostile Impacts of Online Hate Speech

Priyansh Singhal and Piyush Joshi
Department of Computer Science & Engineering
Indian Institute of Information Technology Sri City, India
Correspondence: Priyansh Singhal

Abstract

Hate speech on social media poses significant challenges for content moderation and user safety. While various datasets exist for hate speech detection, existing approaches treat hate speech as a monolithic phenomenon, detecting hateful content by using simple categorical labels such as hate, offensive, or toxic. This approach fails to distinguish between the speaker's underlying motivations and the content's potential societal consequences. This paper introduces I2-HATE, a novel dataset with a dual taxonomy that separately captures *Intent* (why the speaker produced hate speech) and *Impact* (what harm it may cause to individuals and communities) of online hateful posts. This dual-taxonomy approach enables moderation systems to differentiate hateful content based on underlying motivation and potential harm, supporting more nuanced intervention strategies. We release the I2-HATE dataset[1] and code[2] publicly.

1 Introduction

Hate speech on social media platforms inflicts psychological harm on individuals and contributes to offline violence and social disruption (Naslund et al., 2020; Okpala and Cheng, 2025; Saha et al., 2019; Dreißigacker et al., 2024; Deep Singh et al., 2025; Akomeah, 2023). The volume of user-generated content makes manual moderation unsustainable, necessitating automated detection systems (Kapil and Ekbal, 2024; Schmidt and Wiegand, 2017). However, current moderation systems predominantly operate on binary removal decisions, which Goldman (2021) argues "*has hindered the consideration of other remedial options.*" Effective moderation requires graduated responses where remedies are proportionate to the violation (Goldman, 2021; Scheuerman et al., 2021).

Early research framed hate speech detection as binary or ternary classification. Foundational datasets like Waseem and Hovy (2016) and Davidson et al. (2017) provided crucial benchmarks, though systematic reviews revealed pervasive issues including low inter-annotator agreement, dataset degradation, and definitional inconsistencies (Fortuna et al., 2020; Madukwe et al., 2020; Vidgen and Derczynski, 2020). The field subsequently pursued greater nuance through multi-label taxonomies capturing intersectional hate (Lee et al., 2022; Mollas et al., 2022; Mathew et al., 2021), multi-aspect frameworks deconstructing messages along axes like hostility type and target group (Ousidhoum et al., 2019), and implicit hate speech detection (ElSherief et al., 2021; Ocampo et al., 2023; Gao et al., 2017).

Despite this progress, existing approaches share a fundamental limitation: they do not separately model speaker **Intent** (why hate is produced) and societal **Impact** (what harm it causes). This conflation is consequential because, as Wilson and Land (2020) argue, "*the full meaning and potential effect of any speech act lies with the intent of the speaker, the content of the expression, and the context in which it is uttered.*" Platform policies already consider author intent as a criterion for moderation decisions, yet Wang et al. (2025b) identify a critical disconnect: current detection models "*typically lack efforts to capture intent.*" Meanwhile, frameworks like MLMA (Ousidhoum et al., 2019) include annotator sentiment (a proxy for impact) but not intent, while "Measuring Hate Speech" (Sachdeva et al., 2022) captures perceived harm without separate intent judgments. These dimensions remain conflated into a single spectrum of "hatefulness," preventing models from distinguishing malicious threats from ignorant microaggressions, a distinction essential for proportionate enforcement aligned with platform severity frameworks (O'Kane, 2021; Scheuerman et al., 2021).

To address this gap, we introduce **I2-Hate**, a dataset built on a dual-taxonomy that explicitly models Intent and Impact as independent, multi-label classification tasks. Drawing from Wang et al. (2025b)'s application of Weberian concepts to online hate, our Intent taxonomy captures motivations such as ideological expression and strategic incitement (Section 2.3), while our Impact taxonomy captures harms ranging from psychological damage to incitement to violence (Section 2.4).

Our contributions are as follows:

1. **Novel Dual-Taxonomy Framework:** We introduce the first theoretically grounded framework separating speaker Intent from societal Impact as independent, multi-label classification tasks for hate speech analysis.

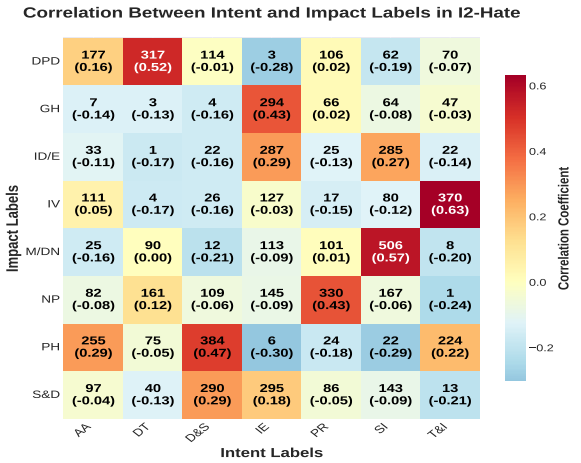
[1] https://huggingface.co/datasets/PS4Research/I2-Hate

[2] https://github.com/ps-research/I2-Hate

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 136–153

Label	Train	Val	Test	Total
Intent Label Distribution				
Affective Aggression [**AA**]	251	90	100	441
Derisive Trolling [**DT**]	229	75	79	383
Dominance & Subjugation [**D&S**]	316	107	97	520
Ideological Expression [**IE**]	422	145	150	717
Performative Reinforcement [**PR**]	258	80	86	424
Strategic Incitement [**SI**]	455	154	144	753
Threat & Intimidation [**T&I**]	278	84	83	445
Total Intent	**2209**	**735**	**739**	**3683**
Impact Label Distribution				
Disruption of Public Discourse [**DPD**]	445	149	152	746
Glorification of Hate [**GH**]	268	98	78	444
Incitement to Discrimination/Exclusion [**ID/E**]	374	117	123	614
Incitement to Violence [**IV**]	394	134	122	650
Misinformation/Disinformation Nexus [**M/DN**]	449	161	150	760
Normalization of Prejudice [**NP**]	543	173	190	906
Psychological Harm [**PH**]	524	167	162	853
Stigmatization & Dehumanization [**S&D**]	496	184	183	863
Total Impact	**3493**	**1183**	**1160**	**5836**

(a) Distribution of labels across splits.



(b) Correlation matrix between Intent and Impact labels. Numbers show co-occurrence counts (top) and Pearson correlation coefficients (bottom).

Figure 1: Overview of I2-Hate label distributions and cross-taxonomy correlations

2. **I2-Hate Dataset:** We present 3,296 Twitter (presently known as 𝕏) posts annotated with 7 Intent labels and 8 Impact labels, enabling systems to match moderation actions to both speaker motivation and harm severity.
3. **Comprehensive Model Benchmark:** We evaluate 19 transformer models across 6 architectural families (encoders, domain-specific encoders, advanced encoders, encoder-decoders, and lightweight variants), establishing performance baselines.
4. **State-of-the-Art LLM Evaluation:** We benchmark 4 frontier LLMs (Llama 4 (Scout) (Meta AI, 2025), ChatGPT-5 (OpenAI, 2025), Gemini 2.5 Pro (Google DeepMind, 2025), Claude Sonnet 4.5 (Anthropic, 2025)) in zero-shot and 3-shot settings.

We release the **I2-Hate** dataset publicly on Hugging Face to support content moderation teams, platform safety researchers, and policymakers in developing context-aware hate speech detection systems.

2 I2-Hate Dataset

2.1 Data Collection and Annotation

We collected 3,296 posts from Twitter (now 𝕏). **Inclusion criteria:** Posts must (1) target contentious social, political, cultural, or racial issues; (2) focus on identity groups or social categories; and (3) contain hate directed at collective identities rather than individuals. **Exclusion criteria:** Personal attacks on specific public figures, brand or company criticism, and interpersonal grievances unrelated to protected characteristics. These criteria ensured the dataset captures hate speech targeting marginalized identities within broader societal debates, where Intent and Impact distinctions matter most for moderation.

Three annotators underwent rigorous two-week training on the taxonomy framework and completed practice annotation rounds before independently labeling all posts (detailed methodology in Appendix A). Inter-annotator agreement measured by Fleiss' kappa (Fleiss, 1971) yielded κ=**0.74** for Intent and κ=**0.79** for Impact, indicating substantial agreement. Figure 1a presents the distribution of labels across splits.

2.2 Taxonomy Development Process

Our dual taxonomy was developed through **literature-informed iterative refinement rather than strict theoretical derivation**.

Intent Taxonomy Development

We began with Wang et al. (2025a)'s application of Weberian concepts to hate speech, which frames hateful expression as purposive action with distinct motivations.

Categories from Wang et al.'s Weberian framework:

- *Affective Aggression* ← Affectual action: "*an emotional response, such as anger or frustration*"
- *Ideological Expression* ← Value-rational action: "*motivated by values and beliefs*"
- *Strategic Incitement* ← Goal-rational action: "*used strategically to achieve political or ideological goals*"

Categories from domain-specific literature:

- *Threat & Intimidation*: Marsters (2019)'s distinction between "Howlers" (intimidation) and "Hunters" (actionable threats)
- *Dominance & Subjugation*: Carlson (2020)'s framing as "*hate speech... a mechanism for colonization... the practice of domination involving the subjugation of one group*"

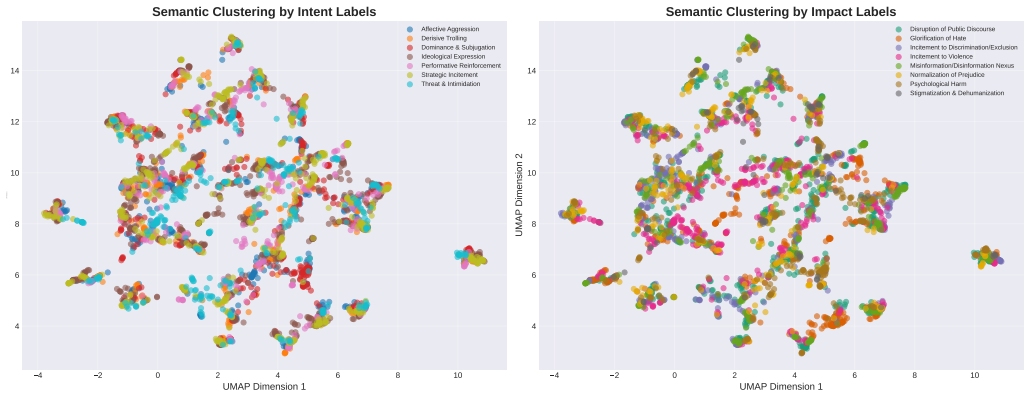


Figure 2: UMAP visualization using all-MiniLM-L6-v2 embeddings shows semantic overlap across Intent and Impact categories. Posts with different dominant labels occupy shared embedding space, indicating absence of discrete classification boundaries.

- *Derisive Trolling*: Coles and West (2016)'s definition as "*manipulating another individual into losing their temper... causing deliberate offence*"

- *Performative Reinforcement*: Kunst et al. (2021)'s work on in-group signaling where "*individuals may be encouraged to engage if they expect social recognition*"

We iteratively refined to 7 categories within our target range of 5-8 for annotator reliability.

Impact Taxonomy Development

While Intent captures speaker motivation, Impact captures societal harm regardless of intent.

Categories with guidance from Wang et al.:

- *Incitement to Violence*: "*online hate speech has been recognised... for its potential to incite and propagate offline violence*"

- *Misinformation/Disinformation Nexus*: hate that "*rel[ies] on coded language and misinformation*"

Categories from domain-specific research:

- *Normalization of Prejudice*: Soral et al. (2018) explain how exposure leads content to be "*interpreted by an individual as less negative and harmful... increasing prejudice*"

- *Psychological Harm*: Judge and Nel (2018) document "*the psychological hurt and the psychological harm of hate speech is undeniable.*"

- *Stigmatization & Dehumanization*: Saffari et al. (2024) define this as "*content which perceives or treats people as less than human*"

- *Disruption of Public Discourse*: Cover (2023) describe how hate effects "*the withdrawal of relationality... a right to speak or participate socially*"

2.3 Intent Taxonomy

The Intent dimension captures seven distinct speaker motivations behind hate speech production:

(1) Affective Aggression: Reactive emotional expression driven by anger, frustration, or outrage, characterized by impulsive hostile language without strategic planning (Ray and George, 2021; Mane et al., 2025). As Mane et al. (2025) note, "*the Frustration-Aggression Theory suggests that frustration leads to aggression, which can be exacerbated by social media platforms.*"

(2) Derisive Trolling: Deliberate provocation for amusement or disruption, employing mockery, sarcasm, or feigned ignorance to elicit reactions (Coles and West, 2016).

(3) Dominance & Subjugation: Assertion of power and social hierarchy through degradation and belittling language that positions target groups as inferior (Carlson, 2020).

(4) Ideological Expression: Articulation of hateful worldviews or political ideologies that position certain groups as threats to valued institutions or social order (Lee, 2010).

(5) Performative Reinforcement: In-group signaling and solidarity building through shared hateful rhetoric, reinforcing group identity and boundaries (Kunst et al., 2021).

(6) Strategic Incitement: Calculated language crafted to achieve specific political, ideological, or social objectives, including mobilizing followers or coordinating hostile actions (Fyfe, 2017).

(7) Threat & Intimidation: Direct or implied threats designed to instill fear, silence targets, or warn of impending harm (Marsters, 2019).

2.4 Impact Taxonomy

The Impact dimension captures eight domains of potential societal harm inflicted by hate speech, regardless of speaker intent:

(1) Psychological Harm: Emotional distress, anxiety, fear, or trauma experienced by target individuals or

Model Name	Subset Accuracy		Macro F1		Weighted F1		Micro F1		Hamming Loss	
	Intent	Impact	Intent	Impact	Intent	Impact	Intent	Impact	Intent	Impact
BERT-base	0.8012	0.4643	0.8919	0.7859	0.8826	0.7818	0.8826	0.7838	0.0369	0.0924
BERT-large	0.8118	0.5266	0.8987	0.8118	0.8895	0.8076	0.8897	0.8085	0.0347	0.0823
RoBERTa-base	**0.8149**	0.5357	0.9007	0.8194	**0.8899**	0.8142	0.8903	0.8144	0.0345	0.0823
RoBERTa-large	0.7436	0.4598	0.8634	0.7809	0.8536	0.7773	0.8548	0.78	0.0464	0.0943
HateBERT	0.7891	0.4219	0.8810	0.7616	0.8706	0.7569	0.8702	0.7575	0.0405	0.1064
Twitter-RoBERTa-Hate	0.7982	0.5038	0.8801	0.7998	0.8704	0.7959	0.8714	0.7952	0.0403	0.0903
ToxicBERT	0.7951	0.4583	0.8807	0.7819	0.8724	0.7779	0.8721	0.7790	0.0403	0.0941
DistilBERT	0.7936	0.4461	0.8875	0.7721	0.8779	0.7667	0.8776	0.7679	0.0386	0.1000
DistilRoBERTa	0.7906	0.4461	0.8850	0.7677	0.8744	0.7640	0.8747	0.7645	0.0397	0.1007
DistilBART-cnn	0.8027	0.4917	0.8843	0.7932	0.8764	0.7882	0.8764	0.7904	0.0390	0.0909
DeBERTa-v3-small	0.7860	0.4568	0.8799	0.7734	0.8715	0.7691	0.8716	0.7701	0.0408	0.0998
DeBERTa-v3-base	0.8134	0.5144	**0.903**	0.8047	0.8942	0.7984	**0.8939**	0.7991	**0.0338**	0.0878
XLM-RoBERTa-base	0.7785	0.4461	0.8718	0.7742	0.8597	0.7683	0.8585	0.7711	0.0447	0.0986
ELECTRA-small	0.7602	0.3733	0.8401	0.7209	0.8319	0.7138	0.8324	0.7210	0.0512	0.1144
ELECTRA-base	0.7982	0.5053	0.8932	0.7985	0.8855	0.7937	0.8849	0.7954	0.0364	0.0882
ConvBERT-base	0.7891	0.5296	0.8802	0.8092	0.8733	0.8053	0.8735	0.8056	0.0401	0.0848
ALBERT-base-v2	0.7739	0.4810	0.8711	0.7880	0.8629	0.7836	0.8628	0.7836	0.0434	0.0933
BART-base	0.8103	0.5083	0.8962	0.8015	0.8871	0.7958	0.8872	0.7989	0.0356	0.0869
BART-large	0.7921	**0.5539**	0.8773	**0.8253**	0.8695	**0.8201**	0.8703	**0.8212**	0.0403	**0.078**

Figure 3: Performance of 19 transformer models on Intent and Impact classification tasks. Best results per metric are shown in bold.

groups (Judge and Nel, 2018).

(2) Stigmatization & Dehumanization: Systematic stripping of human dignity through language that portrays groups as less than human (Saffari et al., 2024).

(3) Normalization of Prejudice: Contribution to mainstreaming discriminatory attitudes by making hateful views appear acceptable or widespread (Soral et al., 2018).

(4) Glorification of Hate: Celebration extol the crimes of genocide, historical atrocities, hate crimes, or extremist figures (Cortés, 2021).

(5) Incitement to Violence: Language that encourages, justifies, or provides rationale for physical harm against individuals or groups (Kopytowska and Baider, 2017).

(6) Incitement to Discrimination/Exclusion: Promotion of discriminatory practices or social exclusion targeting specific groups (Leader et al., 2009).

(7) Disruption of Public Discourse: Derailment of constructive dialogue through hate speech that silences marginalized voices (Cover, 2023).

(8) Misinformation/Disinformation Nexus: Intersection of hate speech with false or misleading claims about target groups (Cinelli et al., 2021; Kim and Kesari, 2021).

Both taxonomies allow multi-label annotation, as posts frequently exhibit multiple intents and impacts simultaneously. Complete label definitions with examples are provided in Appendix B. Figure 1b illustrates the correlation structure between Intent and Impact dimensions: while largely orthogonal, certain combinations occur more frequently, such as Derisive Trolling with Disruption of Public Discourse ($r = 0.52$), Threat & Intimidation with Incitement to Violence ($r = 0.63$), and Strategic Incitement with Misinformation/Disinformation Nexus ($r = 0.57$). Figure 2 visualizes semantic embedding space using sentence transformers (Transformers). The extensive overlap between label categories reveals that hate speech does not fall into discrete semantic clusters; posts with different dominant labels occupy shared embedding regions. The absence of separable clusters validates our multi-label framework over binary classification. Additional exploratory analysis is provided in Appendix C.

3 Experiments and Results

3.1 Models and Training

We evaluate 19 transformer models across multiple scales and architectural families. **Fine-tuned models** include: (1) General encoders: BERT-base (Devlin et al., 2018), BERT-large, RoBERTa-base (Liu et al., 2019), RoBERTa-large, DistilBERT, DistilRoBERTa (Sanh et al., 2019); (2) Domain-specific hate speech models: HateBERT (Mathew et al., 2020), ToxicBERT (Hanu and Unitary team, 2020), Twitter-RoBERTa-Hate (Barbieri et al., 2020); (3) Advanced encoders: DeBERTa-v3-small (He et al., 2021), DeBERTa-v3-base, ELECTRA-small and ELECTRA-base (Clark et al., 2020), ALBERT-base-v2 (Lan et al., 2019), XLM-RoBERTa-base (Conneau et al., 2019), ConvBERT-base (YituTech, 2021); (4) Encoder-decoders: BART-base (Lewis et al., 2019), BART-large, DistilBART-cnn (Shleifer, 2020). Additionally, we evaluate **state-of-the-art LLMs** in zero-shot and 3-shot settings: Llama 4 (Scout variant), ChatGPT-5, Gemini 2.5 Pro, and Claude Sonnet 4.5.

The dataset was split 60/20/20 (1,978/659/659 posts) using stratified sampling. All transformer models were fine-tuned for multi-label classification using binary cross-entropy loss (Zhang and Sabuncu, 2018) and AdamW optimizer (Loshchilov and Hutter, 2017).We report micro-F1, macro F1 (Takahashi et al., 2022) (primary metric, handles class imbalance), weighted F1 (Hi-

Model Name	Subset Accuracy		Macro F1		Weighted F1		Micro F1		Hamming Loss	
	Intent	Impact	Intent	Impact	Intent	Impact	Intent	Impact	Intent	Impact
Llama4_ZeroShot	0.05	0.01	0.5933	0.5484	0.5906	0.5489	0.5424	0.5216	0.3543	0.415
Llama4_3Shot	0.01	0	0.5895	0.5195	0.5866	0.5198	0.5348	0.4932	0.3729	0.465
ChatGPT5_ZeroShot	0.08	0.01	0.6386	0.5249	0.6361	0.5255	0.5772	0.5042	0.3014	0.4375
ChatGPT5_3Shot	0.06	0.01	0.6307	0.4892	0.6277	0.49	0.5642	0.4706	0.32	0.5062
gemini-2.5-pro_ZeroShot	0.05	0.01	0.6075	0.5201	0.6045	0.5207	0.5462	0.5007	0.3229	0.4338
gemini-2.5-pro_3Shot	0.04	0.01	0.6176	0.5013	0.6148	0.5019	0.5534	0.4816	0.3229	0.4925
ClaudeSonnet4.5_ZeroShot	0.08	0.05	0.6173	0.5697	0.6147	0.5697	0.5588	0.5525	0.2843	0.3038
ClaudeSonnet4.5_3Shot	0.08	0	0.6243	0.5819	0.6216	0.5823	0.5641	0.5528	0.2914	0.3337
BART-large (Reference)	**0.7921**	**0.5539**	**0.8773**	**0.8253**	**0.8695**	**0.8201**	**0.8703**	**0.8212**	**0.0403**	**0.078**

Figure 4: Performance of SOTA LLMs on Intent and Impact classification tasks. For Llama 4 we used the Scout version.

nojosa Lee et al., 2024), subset accuracy (Nam et al., 2017) (exact label match), and Hamming Loss (Wu and Zhu, 2020) (fraction of incorrectly predicted labels in multi-label classification). LLMs received complete taxonomy definitions from Appendix B (all 15 labels with definitions and key indicators). Details about experimental setup, hyperparameters, and reproducibility are in Appendix D.

3.2 Overall Performance

Figure 3 presents test set performance for fine-tuned transformer models. DeBERTa-v3-base achieves the strongest Intent classification performance with macro F1 of 0.903 and subset accuracy of 0.8134, closely followed by RoBERTa-base (0.9007 macro F1, 0.8149 subset accuracy). BERT-large ranks third with 0.8987 macro F1, demonstrating that scaling model size provides marginal gains. For Impact classification, BART-large leads with macro F1 of 0.8253 and subset accuracy of 0.5539, followed by RoBERTa-base (0.8194 macro F1, 0.5357 subset accuracy). Notably, larger model variants (BERT-large, RoBERTa-large) do not consistently outperform their base counterparts, RoBERTa-large achieves only 0.8634 Intent macro F1 compared to RoBERTa-base's 0.9007, suggesting task-specific optimization matters more than parameter count.

Intent classification proves substantially more tractable than Impact detection across all models, with average macro F1 scores of 0.8824 and 0.7879 respectively, a 12% relative difference. Detailed per-label precision, recall, and F1 scores for Intent and Impact classification are provided in Appendix E.

3.3 Domain-Specific Model Performance

Domain-specific hate speech models show mixed results. Twitter-RoBERTa-Hate marginally outperforms general encoders on Impact classification (0.7998 vs DistilBERT 0.7721, +3.6%) but underperforms on Intent (0.8801 vs RoBERTa-base 0.9007, -2.3%). However, more specialized variants struggle significantly: HateBERT suffers -1.2% degradation on Intent and 3.1% on Impact compared to BERT-base, while ToxicBERT shows similar under-performance (-1.2% Intent, -0.6% Impact).

3.4 Large Language Model Evaluation

Figure 4 presents SOTA LLM performance on I2-HATE. Despite receiving complete taxonomy definitions with 15 labels, operational definitions, and key indicators, all LLMs dramatically underperformed fine-tuned transformer models. The best-performing configuration Claude Sonnet 4.5 in 3-shot setting achieved only 0.6243 Intent macro F1 and 0.5819 Impact macro F1, representing 30.7% and 29% degradation compared to RoBERTa-base (0.9007 and 0.8194 respectively). Subset accuracy collapsed catastrophically: LLMs achieved 0.00-0.08 on Intent (vs. 0.8149 for RoBERTa-base) and 0.00-0.05 on Impact (vs. 0.5539 for BART-large).

Error analysis revealed systematic over-labeling: LLMs assigned substantially more labels per post than ground truth annotations. When we presented LLM predictions to our original annotators, they frequently disagreed with the label assignments. Few-shot learning provided minimal benefit and sometimes degraded performance (e.g., ChatGPT-5 Impact dropped from 0.5249 to 0.4892 macro F1). This suggests the task requires learning nuanced label boundaries from substantial training data rather than in-context pattern matching.

These results validate I2-HATE's contribution: **the dataset enables capabilities that neither scaling nor prompting can achieve**. The LLMs performance gap demonstrates that our dual-taxonomy framework captures genuine linguistic subtlety beyond what current prompting paradigms can address.

4 Conclusion and Future Work

Hate speech detection requires moving beyond binary toxicity classification to understand both communicative intent and societal consequences. I2-Hate provides the first dataset annotating Intent (7 labels) and Impact (8 labels) in 3,296 social media posts, enabling content moderation systems to distinguish speaker motivations from downstream harms. Future work should extend this dual-taxonomy framework multilingually, validate real-world moderation interventions that leverage Intent-Impact distinctions for proportional responses, and apply it to other harmful content domains, establishing generalizable methods for context-aware content policy.

Limitations

Our dataset is drawn exclusively from Twitter, limiting generalizability to platforms with different communication norms (Reddit's pseudonymity, TikTok's video format). The English-only content prevents multilingual analysis and cross-cultural hate speech pattern detection. The dual taxonomy introduces subjectivity in distinguishing Intent from Impact: while annotators achieved strong agreement (Fleiss' $\kappa = 0.74$ Intent, 0.79 Impact), borderline cases remained challenging. The dataset captures 2024-2025 hate speech; evolving slang and coded language require periodic retraining. Computational constraints limited evaluation to dual T4 GPU infrastructure, precluding larger architectures and advanced techniques like multi-task learning that could improve performance.

Ethical Considerations

All posts in I2-Hate are sourced from publicly accessible Twitter in compliance with platform Terms of Service. While this constitutes public data, we recognize the sensitive nature of hate speech content. We implement privacy protections including removal of usernames, profile images, timestamps, and potentially identifying information. Posts containing threats of imminent violence were reported to platform moderators prior to dataset inclusion. The annotation process involved three mental health professionals with expertise in hate speech impacts, who received comprehensive training and access to mental health support resources throughout the annotation period. I2-Hate is designed exclusively for research into hate speech detection, content moderation improvement, and understanding online toxicity patterns—**and is not intended for surveillance, profiling, suppression of marginalized voices, or any punitive applications that could harm vulnerable communities.** We acknowledge dual-use risks: while our work aims to improve detection systems, the dataset could theoretically be misused. To mitigate this, we will implement a gated public access when me make the dataset public.

References

Kwabena Odame Akomeah. 2023. Hate Speech Detection beyond plain Natural Language Processing. In *Proceedings of the First Workshop on Foundational Models and Disinformation (FDIA@ESSIR 2023)*.

Anthropic. 2025. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Caitlin Ring Carlson. 2020. Hate speech as a structural phenomenon. *First Amendment Studies*, 54(2):217–224.

Matteo Cinelli, Andraž Pelicon, Igor Mozetič, Walter Quattrociocchi, Petra Kralj Novak, and Fabiana Zollo. 2021. Dynamics of online hate and misinformation. *Scientific reports*, 11(1):22083.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Bryn Alexander Coles and Melanie West. 2016. Trolling the trolls: Online forum users constructions of the nature and properties of trolling. *Computers in Human Behavior*, 60:233–244.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Ismael Cortés. 2021. Hate speech, symbolic violence, and racial discrimination. antigypsyism: what responses for the next decade? *Social Sciences*, 10(10):360.

Rob Cover. 2023. Digital hostility, subjectivity and ethics: Theorising the disruption of identity in instances of mass online abuse and hate speech. *Convergence*, 29(2):308–321.

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.

Daman Deep Singh, Ramanuj Bhattacharjee, and Abhijnan Chakraborty. 2025. Rethinking hate speech detection on social media: Can llms replace traditional models? *arXiv e-prints*, pages arXiv–2506.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Arne Dreißigacker, Philipp Müller, Anna Isenhardt, and Jonas Schemmel. 2024. Online hate speech victimization: consequences for victims' feelings of insecurity. *Crime Science*, 13(1):4.

Mai ElSherief, Caleb Ziems, and et al. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794.

Shannon Fyfe. 2017. Tracking hate speech acts as incitement to genocide in international criminal law. *Leiden Journal of International Law*, 30(2):523–548.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. *arXiv preprint arXiv:1710.07394*.

Eric Goldman. 2021. Content moderation remedies. *Mich. Tech. L. Rev.*, 28:1.

Google DeepMind. 2025. Gemini 2.5 pro model card. Technical report.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Maria Cristina Hinojosa Lee, Johan Braet, and Johan Springael. 2024. Performance metrics for multilabel emotion classification: comparing micro, macro, and weighted f1-scores. *Applied Sciences*, 14(21):9863.

Melanie Judge and Juan A Nel. 2018. Psychology and hate speech: A critical and restorative encounter.

Prashant Kapil and Asif Ekbal. 2024. A survey on combating hate speech through detection and prevention in english. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 485–501.

Jae Yeon Kim and Aniket Kesari. 2021. Misinformation and hate speech: The case of anti-asian hate speech during the covid-19 pandemic. *Journal of Online Trust and Safety*, 1(1).

Monika Kopytowska and Fabienne Baider. 2017. From stereotypes and prejudice to verbal and physical violence: Hate speech in context. *Lodz Papers in Pragmatics*, 13(2):133–152.

Marlene Kunst, Pablo Porten-Cheé, Martin Emmer, and Christiane Eilders. 2021. Do "good citizens" fight hate speech online? effects of solidarity citizenship norms on user responses to hate comments. *Journal of Information Technology & Politics*, 18(3):258–273.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Tirza Leader, Brian Mullen, and Diana Rice. 2009. Complexity and valence in ethnophaulisms and exclusion of ethnic out-groups: What puts the" hate" into hate speech? *Journal of Personality and Social Psychology*, 96(1):170.

Jean Lee, Taejun Lim, and et al. 2022. K-MHaS: A multi-label hate speech detection dataset in Korean online news comment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3530–3538. International Committee on Computational Linguistics.

Steven P Lee. 2010. Hate speech in the marketplace of ideas. In *Freedom of Expression in a Diverse World*, pages 13–25. Springer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the fourth workshop on online abuse and harms*, pages 150–161.

Swapnil Sanjaykumar Mane, Suman Kundu, and Rajesh Sharma. 2025. A survey on online aggression: Content detection and behavioral analysis on social media. *ACM Computing Surveys*, 57(7):1–36.

Alexandria Marsters. 2019. *When hate speech leads to hateful actions: A corpus and discourse analytic approach to linguistic threat assessment of hate speech*. Georgetown University.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*.

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.

Meta AI. 2025. Llama 4 scout: Natively multimodal intelligence. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. 2017. Maximizing subset accuracy with recurrent neural networks in multi-label classification. *Advances in neural information processing systems*, 30.

John A Naslund, Ameya Bondre, John Torous, and Kelly A Aschbrenner. 2020. Social media and mental health: benefits, risks, and opportunities for research and practice. *Journal of technology in behavioral science*, 5(3):245–257.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *EACL 2023-17th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2023, pages 1997–2013. Association for Computational Linguistics.

Ruby O'Kane. 2021. Meta's private speech governance and the role of the oversight board: Lessons from the board's first decisions. *Stan. Tech. L. Rev.*, 25:167.

Ebuka Okpala and Long Cheng. 2025. Large language model annotation bias in hate speech detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1389–1418.

OpenAI. 2025. Gpt-5 system card. Technical report.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. 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 4675–4684. Association for Computational Linguistics.

Argha Ray and Joey George. 2021. Online hate and its routes to aggression: A research agenda. *Social Impact and Information Systems*.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94.

Hamidreza Saffari, Mohammadamin Shafiei, Hezhao Zhang, Lasana Harris, and Nafise Sadat Moosavi. 2024. Beyond hate speech: Nlp's challenges and opportunities in uncovering dehumanizing language. *arXiv preprint arXiv:2402.13818*.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–33.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Sam Shleifer. 2020. DistilBART-CNN-6-6: Distilled BART Model for Summarization. https://huggingface.co/sshleifer/distilbart-cnn-6-6. HuggingFace Model Hub.

Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. 2018. Exposure to hate speech increases prejudice through desensitization. *Aggressive behavior*, 44(2):136–146.

Kanae Takahashi, Kouji Yamamoto, Aya Kuchiba, and Tatsuki Koyama. 2022. Confidence interval for micro-averaged f 1 and macro-averaged f 1 scores. *Applied Intelligence*, 52(5):4961–4972.

Sentence Transformers. all-minilm-l6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Luna Wang, Andrew Caines, and Alice Hutchings. 2025a. Web (er) of hate: A survey on how hate speech is typed. *arXiv preprint arXiv:2506.16190*.

Xinyu Wang, Sai Koneru, Pranav Narayanan Venkit, Brett Frischmann, and Sarah Rajtmajer. 2025b. The unappreciated role of intent in algorithmic moderation of abusive content on social media. *Harvard Kennedy School Misinformation Review*.

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, San Diego, California. Association for Computational Linguistics.

Richard Ashby Wilson and Molly K Land. 2020. Hate speech on social media: Content moderation in context. *Conn. L. Rev.*, 52:1029.

Guoqiang Wu and Jun Zhu. 2020. Multi-label classification: do hamming loss and subset accuracy really conflict with each other? *Advances in Neural Information Processing Systems*, 33:3130–3140.

YituTech. 2021. Convbert-base. Github. https://huggingface.co/YituTech/conv-bert-base.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Appendix A : Annotation Process and Guidelines

The project proposal received Institutional Review Board (IRB) approval prior to data collection and annotation. All annotation procedures adhered to ethical guidelines for research involving potentially traumatic content.

Annotation Guidelines

Annotators were trained to identify both explicit and implicit manifestations of Intent and Impact labels. The annotation process emphasized multi-label assignment, as hate speech posts frequently exhibit multiple intents and impacts simultaneously. Annotators were instructed to assign all applicable labels rather than forcing a single category, reflecting the multifaceted nature of online hate.

Intent Disambiguation

Distinguishing between similar intents required careful attention to pragmatic cues, discourse markers, and functional characteristics:

- **Strategic Incitement vs. Ideological Expression:** Strategic Incitement employs instrumental language aimed at mobilizing action, recruiting, or polarizing (e.g., "Share this before they censor it," coded language, calls to action); Ideological Expression articulates hateful beliefs as statements of conviction without explicit mobilization goals (e.g., "I believe [group] are inferior," declarative statements of worldview).

- **Affective Aggression vs. Dominance & Subjugation:** Affective Aggression features spontaneous emotional outbursts with high valence and reactive positioning (e.g., profanity-laden insults in heated exchanges); Dominance & Subjugation employs calculated language aimed at systematic degradation and power assertion (e.g., persistent dehumanization, organized harassment campaigns).

- **Performative Reinforcement vs. Derisive Trolling:** Performative Reinforcement uses ingroup signaling and shared references to gain social approval within hateful communities (e.g., memes, jargon); Derisive Trolling aims to provoke reactions for entertainment, often crossing community boundaries (e.g., deliberately inflammatory statements designed to "trigger" targets).

- **Threat & Intimidation identification:** Both explicit threats ("You should be [violent act]") and implicit threats ("People like you won't be safe here much longer") qualify. Veiled threats using conditional language ("If I see you...") or implied consequences ("Better watch your back") warrant this label.

Impact Assessment Criteria

Annotators were trained to infer impacts from linguistic patterns and contextual evidence, recognizing that harm manifests through both explicit and coded language:

- **Coded language and dog whistles:** Implicit impacts could be labeled when hate employs plausible deniability. For example, "[Group] have lower IQs—it's just biology" constitutes *Stigmatization & Dehumanization* and *Normalization of Prejudice* despite superficially neutral framing.
- **Severity threshold:** An impact must be substantively evidenced, not merely implied by single words. Mild criticism ("I disagree with [group]'s politics") is insufficient for *Psychological Harm* unless it includes targeted degradation, severe insults, or language likely to cause significant distress to members of the targeted group.
- **Multi-impact posts:** A single passage often warrants multiple impact labels. For instance, "[Group] are parasites spreading disease. They should be eliminated from society" exhibits *Stigmatization & Dehumanization* (animal/disease metaphors), *Incitement to Violence* ("eliminated"), and *Incitement to Discrimination/Exclusion* (removing from society).
- **Historical context:** Annotators considered whether language echoes historical genocidal rhetoric, white supremacist tropes, or extremist propaganda, which elevates impact severity even when not explicitly violent (e.g., replacement theory narratives, blood libel references).

Annotator Recruitment and Training

Annotator Background

Three undergraduate computer science students from our institution served as annotators. All were fluent English speakers with familiarity with harmful online content. Annotators were recruited through departmental announcements and selected based on: (1) availability for the 3-month project duration, (2) demonstrated attention to detail in preliminary screening tasks, (3) familiarity with online discourse norms and social media platforms, and (4) capacity to maintain emotional wellbeing while engaging with hateful content.

Ethical Protections for Annotators

Given the traumatic nature of hate speech content, comprehensive wellbeing supports were implemented:

- **Exposure limits:** Annotation sessions were limited to 2-hour blocks with mandatory breaks to prevent vicarious trauma accumulation.
- **Mental health resources:** Annotators had access to confidential counseling services throughout the project and for 3 months post-completion.
- **Right to refuse:** Annotators could skip posts they found excessively distressing without penalty, with such posts reviewed by the research team.
- **Peer support:** Weekly group debriefing sessions allowed annotators to process emotional impacts collectively.

Training Protocol

Annotators underwent a structured 2-week training program:

1. **Week 1 – Theoretical Foundations:** In-depth study of the dual taxonomy, including Weberian social action theory (Intent) and harm-based frameworks (Impact). Training materials covered historical hate speech patterns, coded language evolution, and contemporary extremist rhetoric to contextualize annotation decisions.
2. **Week 2 – Calibration:** Three practice annotation rounds, each consisting of 15 example posts (45 total) representing diverse hate types, targets, and severity levels. After each round, annotators participated in group discussions to resolve ambiguities, clarify edge cases (particularly Strategic Incitement vs. Ideological Expression), and establish consistent interpretation guidelines. Discrepancies were adjudicated through consensus-building rather than imposed by researchers. Special attention was given to implicit hate, dog whistles, and multi-label decision criteria.

Data Collection and Annotation

Following training, annotators spent 2 weeks collecting posts meeting inclusion criteria: English-language, hate speech directed at protected characteristics (race, religion, gender, sexual orientation, disability, nationality), minimum 20 characters, from public Twitter accounts. This process yielded 3,296 posts. Posts containing imminent threats of violence were flagged and reported to platform moderators before inclusion in the dataset. Each post was independently annotated by all three annotators. The final dataset reflects consensus labels, with disagreements resolved through structured discussion protocols prioritizing lived experience perspectives when annotators' backgrounds differed.

Informed Consent

All annotators provided informed consent and were explicitly informed that:

- Their work would contribute to a research publication
- The resulting dataset would be publicly released under an open license
- The content involved financially distressing and potentially emotionally difficult material

- They could withdraw from the project at any time without penalty

No personally identifying information about annotators is included in any released materials.

Compensation

Annotators were compensated at 10 Indian Rupees (INR) per post labeled. Each annotator independently labeled all 2,408 posts, receiving a total of 24,080 INR (approximately $270 USD at time of compensation) for approximately 70 hours of work (collection and annotation combined). This corresponds to an hourly rate of 344 INR/hour ($3.88 USD/hour).

We acknowledge that this compensation rate is below minimum wage standards in high-income countries such as the United States ($7.25 USD/hour federal minimum). However, this rate aligns with standard compensation for undergraduate research assistants in our geographic context (India), where the cost of living is substantially lower and typical undergraduate research assistant rates range from 200–400 INR/hour. We recognize ongoing debates in the NLP community regarding equitable compensation for annotation labor and commit to adhering to local fair-wage standards while acknowledging the global disparities in compensation norms.

Annotator Well-Being

Given the potentially distressing nature of financial hardship narratives, we implemented the following safeguards:

- Annotators were encouraged to take breaks as needed.
- Weekly check-ins were conducted to assess emotional and mental impact.
- No annotator reported significant distress requiring intervention.

Appendix B: Taxonomy Details and Examples

This appendix provides detailed operational definitions, key indicators, and illustrative examples for each category within the dual-level taxonomy of Speaker Intent and Potential Impact. It is intended as a reference for annotators and researchers applying this framework. All examples are mentioned here are *sanitized* for academic publication.

Part A: Speaker Intent Taxonomy

This section details the seven categories used to classify the observable functional intent of a speaker's message.

1. **Strategic Incitement**

   Definition: Language used as a calculated means to achieve a political or ideological goal, such as radicalizing others, polarizing opinion, or mobilizing a group. Often uses coded language, misinformation, or strategic narratives.

   Key Indicators:

   - Use of coded language or dog whistles to maintain plausible deniability.
   - Strategic deployment of misinformation or conspiracy theories targeting a group.
   - Language designed to radicalize, recruit, or mobilize followers.
   - An instrumental tone focused on achieving specific political outcomes.
   - Appeals to in-group solidarity against a perceived external threat posed by another group.

2. **Ideological Expression**

   Definition: Language that articulates a hateful worldview or belief system as a matter of conviction. The primary goal is the expression of the value itself, not a further strategic objective.

   Key Indicators:

   - Declarative statements presenting hateful beliefs as facts or truths.
   - Language with a principled or moralistic tone, justifying hate based on an ideology.
   - The act of stating the belief is the primary communicative goal.
   - References to a broader hateful worldview (e.g., white supremacy, religious extremism).

3. **Performative Reinforcement**

   Definition: Language intended to signal in-group belonging and gain social approval within a community that shares hateful norms. Reinforces social bonds through shared hateful expression.

   Key Indicators:

   - Use of in-group jargon, memes, or shared hateful references.
   - Language that seeks validation or agreement from like-minded peers.
   - The speech act serves to reinforce the speaker's identity as a member of the group.
   - Often relies on inside jokes or tropes that are hateful but may be opaque to outsiders.

4. **Affective Aggression**

   Definition: Language driven by a spontaneous and intense emotional response, such as anger or frustration. Typically reactive and occurs in the context of an interpersonal conflict.

   Key Indicators:

   - High emotional valence; language is laden with anger, rage, or frustration.
   - Reactive nature, often appearing as a direct reply in a heated exchange.

- Use of profanity, slurs, and direct insults.
- Lacks the calculated or strategic tone of other categories.

5. **Dominance & Subjugation**

   Definition: Language aimed at asserting power over a target by humiliating, insulting, or degrading them. Reinforces a social hierarchy and the target's perceived inferiority.

   Key Indicators:

   - Language that explicitly frames the target as inferior or subordinate.
   - Use of derogatory and humiliating insults targeting a person's identity.
   - Aims to diminish the target's social standing and assert the speaker's superiority.
   - Objectification or mockery designed to degrade the target.

6. **Threat & Intimidation**

   Definition: Language that explicitly or implicitly threatens a target with harm to instill fear and coerce them into silence or inaction.

   Key Indicators:

   - Explicit calls for violence or physical harm.
   - Veiled or coded threats suggesting future harm.
   - Language intended to create a sense of fear or danger.
   - Coercive statements aimed at silencing the target.

7. **Derisive Trolling**

   Definition: Language intended to provoke a strong emotional reaction for the speaker's amusement. May also serve to disrupt conversations.

   Key Indicators:

   - Inflammatory, off-topic, or insincere comments.
   - Use of bad-faith questions or arguments ("just asking questions").
   - Clear intent to elicit an angry or emotional response ("get a rise").
   - May use hateful language instrumentally to maximize the provocative effect.

Part B: Potential Impact Taxonomy

This section details the eight categories used to classify the potential harms or consequences of a hateful message.

1. **Psychological Harm**

   Definition: Language likely to cause significant emotional distress, fear, or trauma in a targeted individual. Includes severe insults, humiliation, and targeted harassment.

   Key Indicators:

   - Intense, personal, and degrading insults targeting an individual's identity.
   - Content that creates a reasonable fear for one's safety.
   - Repeated and unwelcome harassment.
   - Language intended to cause maximum humiliation or emotional pain.

2. **Incitement to Violence**

   Definition: Language that directly or indirectly calls for, encourages, or glorifies physical violence against a person or group based on their protected characteristics.

   Key Indicators:

   - Direct calls to "attack," "kill," or "harm" members of a group.
   - Glorification of past acts of violence against a group.
   - Coded language that advocates for violence (e.g., "day of the rope").
   - Dehumanizing language used to justify physical harm.

3. **Incitement to Discrimination/Exclusion**

   Definition: Language that advocates for denying a group their rights or excluding them from social, economic, or political life (e.g., jobs, housing, public services).

   Key Indicators:

   - Calls to deny a group employment, housing, or access to services.
   - Advocating for segregation or the creation of "[group]-free zones."
   - Demands for the removal of a group's civil rights or political participation.
   - Promoting policies that would systematically disadvantage a protected group.

4. **Stigmatization & Dehumanization**

   Definition: Language that portrays a group as subhuman (e.g., animals, vermin), inherently flawed, or a societal blight. Strips the target of their humanity.

   Key Indicators:

   - Metaphors comparing a group to animals, insects, filth, or disease.
   - Language that describes a group as a cancer or plague on society.
   - Treating people as objects or inherently inferior beings.

- Rhetoric that denies the target group's capacity for human emotion or reason.

5. **Normalization of Prejudice**

   Definition: Language that propagates stereotypes and biased beliefs, making them seem socially acceptable and reinforcing systemic inequalities.

   Key Indicators:

   - Presenting harmful stereotypes as common sense or widely-held beliefs.
   - Using generalizations to attribute negative traits to all members of a group.
   - Hateful or biased jokes that trivialize prejudice.
   - Casual use of derogatory language in a way that implies it is normal.

6. **Disruption of Public Discourse**

   Definition: Language that makes online spaces hostile and unusable for targeted groups, chilling speech and silencing marginalized voices through intimidation or harassment.

   Key Indicators:

   - Coordinated harassment campaigns (dogpiling) to drive a user offline.
   - Creating a toxic environment where members of a group feel unsafe to participate.
   - Using threats or abuse to shut down conversations on specific topics.
   - Overwhelming a target with hateful messages to silence them.

7. **Misinformation/Disinformation Nexus**

   Definition: Hate speech that relies on or spreads harmful falsehoods about a protected group to justify hatred, discrimination, or violence.

   Key Indicators:

   - Citing false statistics or fabricated stories to portray a group as dangerous.
   - Spreading known conspiracy theories that target a specific group (e.g., QAnon, antisemitic tropes).
   - Using doctored images or videos to support a hateful narrative.
   - Falsely accusing a group of conspiring to cause societal harm.

8. **Glorification of Hate**

   Definition: Language that praises or supports hateful ideologies, violent extremist groups, historical atrocities, or perpetrators of hate crimes.

   Key Indicators:

   - Expressing admiration for historical hate figures (e.g., Hitler) or extremist groups (e.g., KKK).
   - Celebrating or minimizing historical atrocities (e.g., Holocaust denial).
   - Using slogans, symbols, or imagery associated with violent hateful movements.
   - Praising individuals who have committed hate crimes.

WARNING: Hateful Content, Reader discretion is advised.

Text: If you make a song criticizing Jews or praising Hitler, you get banned by every streaming platform.
If you make a song blaspheming God or worshipping Satan, you get to play the Grammy's, the Super Bowl, the VMA's.
To find out who rules over you. . .
Intent Labels: ["Strategic Incitement", "Ideological Expression"]
Impact Labels: ["Normalization of Prejudice", "Disruption of Public Discourse", "Misinformation/Disinformation Nexus"]

Text: Why are niggers still living in the US if it's so evil
racist? Why didn't you return to your ancestral homes? Do you have a Stockholm syndrome?
I'll tell you why: Even niggers don't want to live next to other niggers, let alone in countries run by them.
Intent Labels: ["Ideological Expression", "Affective Aggression"]
Impact Labels: ["Psychological Harm", "Incitement to Discrimination/Exclusion", "Stigmatization & Dehumanization", "Normalization of Prejudice"]

Text: Oh wow, another fascinating thread on the 'socioeconomic factors' behind the vibrant cultural expression of looting a Foot Locker. Please, enlighten me more. I'm on the edge of my seat. Is it because of redlining that someone is curb-stomping a tourist for his iPhone? Is it the legacy of slavery that causes a pack of them to beat a store clerk unconscious? My tiny brain just can't comprehend the complex historical forces at play. Or, and hear me out, maybe... just maybe... they're violent criminals who belong in cages? Nah, couldn't be. That's too simple. Let's get another 50-tweet thread from a gender studies major to explain it to me.
Intent Labels: ["Derisive Trolling"]
Impact Labels: ["Disruption of Public Discourse","Normalization of Prejudice"]

Text: I'm not a racist, I'm a realist. I'm just looking at the patterns. Why is it that when a neighborhood's demographic shifts past a certain percentage, the crime rates, school quality, and property values invariably plummet? This isn't an isolated incident; it happens with predictable regularity across the country. Are we not allowed to ask why? Are we not allowed to discuss the observable reality that forced integration consistently leads to the degradation of stable communities? The media will call you a bigot for even pointing this out. But if one variable consistently produces the same negative outcome, don't you have a responsibility to question the premise? I'm just asking questions that everyone is thinking but is too afraid to say out loud. What is the actual end game here?
Intent Labels: ["Strategic Incitement"]
Impact Labels: ["Misinformation/Disinformation Nexus","Normalization of Prejudice"]

Text: Check out this new Wojak meme I just saw. It's the 'I love living in a diverse city' soyjak. He's got his little NPR tote bag and everything. In the first panel he's smiling. In the second panel, he's looking over his shoulder nervously at a group of 'teens.' In the third, he's getting his car jacked. And in the last panel, he's on the ground saying 'At least I'm not a racist!' LMAO. It's funny because it's true. Every single one of my liberal friends who lives in the city has a story like this. They'll never admit it, but they're terrified. This is peak comedy right here.
Intent Labels: ["Derisive Trolling","Performative Reinforcement"]
Impact Labels: ["Normalization of Prejudice","Stigmatization & Dehumanization"]

Table 1: This table contains explicit examples of hate speech, presented to illustrate the application of our taxonomy.

Appendix C : Exploratory Dataset Analysis

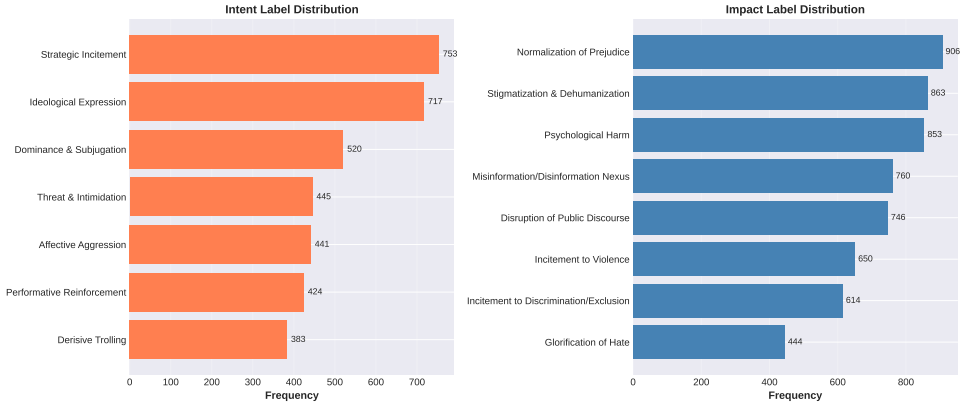


Figure 5: **Intent and Impact Label Distribution in I2-Hate Dataset.** Left: Intent label frequencies show Strategic Incitement (753 posts) and Ideological Expression (717 posts) as dominant categories, while Derisive Trolling (383 posts) is least common. Right: Impact label frequencies reveal Normalization of Prejudice (896 posts) and Stigmatization & Dehumanization (863 posts) as most prevalent consequences, with Glorification of Hate (444 posts) least frequent. The distribution reflects the multi-dimensional nature of hate speech, where single posts often exhibit multiple labels simultaneously.

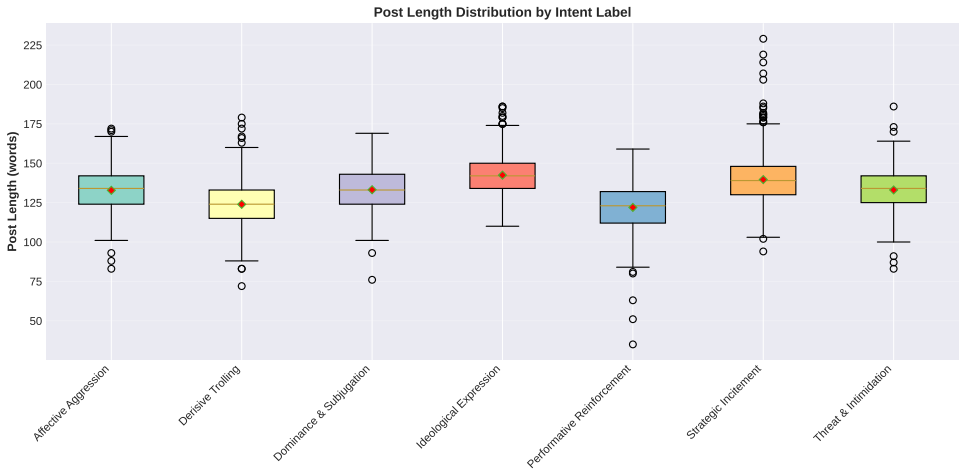


Figure 6: **Post Length Distribution by Intent Label.** Boxplots showing word count distributions for each Intent category. Strategic Incitement exhibits highest median length (142 words), reflecting verbose, manifesto-style rhetoric aimed at mobilization. Ideological Expression shows similar verbosity (median 138 words), consistent with detailed articulation of hateful worldviews. Performative Reinforcement demonstrates shortest median (121 words), aligning with social media engagement optimization strategies favoring brevity.

Figure 7: **Top 10 Most Frequent 2-grams by Intent Category.** Each subplot reveals distinct lexical patterns characterizing different Intent types. **Affective Aggression**: Reactive phrases (“just saw,” “people like,” “idiot post”) indicating spontaneous emotional responses. **Derisive Trolling**: Provocative language (“trying understand,” “just trying,” “feels like”) designed to elicit reactions. **Dominance & Subjugation**: Power-asserting phrases (“people like,” “george soros,” “deep same”) aimed at degradation. **Ideological Expression**: Declarative statements (“george soros,” “men like,” “lived nuance,” “black lives”) articulating worldviews. **Performative Reinforcement**: In-group signaling (“due lol,” “doing wake,” “stay strong”) with community jargon. **Strategic Incitement**: Mobilizing rhetoric (“george soros,” “federal reserve,” “american people,” “society foundations”) using conspiracy references. **Threat & Intimidation**: Menacing language (“day reckoning,” “american people,” “reckoning coming”). These lexical fingerprints provide interpretable features distinguishing Intent categories.

Appendix D : Experimental Setup and Reproducibility

Hyperparameter Configuration

All models were fine-tuned using identical hyperparameter configurations to ensure fair comparison across architectures. Table 2 presents the complete hyperparameter settings used for all experiments.

Parameter	Value
Optimizer	AdamW
Learning Rate	5e-5
Learning Rate Schedule	Linear with Warmup
Warmup Steps	500
Batch Size (Training)	16
Batch Size (Validation)	16
Batch Size (Test)	16
Number of Epochs	10
Early Stopping Patience	3 epochs
Weight Decay	0.01
Gradient Clipping (max norm)	1.0
Dropout Rate	0.2
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	1e-8
Max Sequence Length	512 tokens
Loss Function	Binary Cross-Entropy
Random Seed	42

Table 2: Complete hyperparameter configuration used for all models.

Training dynamics: All models were trained with a linear learning rate schedule including warmup over the first 500 steps. Early stopping was configured to halt training if validation loss did not improve for 3 consecutive epochs; however, no models triggered early stopping within the 10-epoch limit, indicating that all models converged or plateaued within this training duration.

Computational Resources

All experiments were conducted on Kaggle's cloud computing platform using the following hardware configuration:

- **GPUs:** Dual NVIDIA Tesla T4 (16GB memory each)
- **Training Strategy:** PyTorch DataParallel for models fitting memory constraints
- **CPU:** 4-core Intel Xeon (exact model varies by Kaggle allocation)
- **RAM:** 30GB system memory
- **Storage:** SSD-backed persistent storage

Training time: Total wall-clock time for all experiments (15 models $\times$ 2 tasks = 30 training runs) was approximately 10 hours.

Computational budget: Using Kaggle's free tier GPU quota (30 hours per week), the entire experimental pipeline incurred **zero direct monetary cost**. This demonstrates the feasibility of reproducing our results using freely available cloud computing resources, enhancing accessibility for researchers without institutional GPU access.

Software Environment

Experiments were conducted using Kaggle's standard Python environment (as of October 2025). All dependencies were installed via pip and conda. Key package versions are listed below:

Package	Version
Python	3.11.13
PyTorch	2.6.0+cu124
CUDA Toolkit	12.4
Transformers (Hugging Face)	4.52.4
scikit-learn	1.2.2
NumPy	1.26.4
Pandas	2.2.3
SentenceTransformers	4.1.0
tqdm	4.67.1
Matplotlib	3.7.1
Seaborn	0.12.2

Table 3: Software environment and package versions used in all experiments.

Code and Data Release

Reproducibility statement: To facilitate reproducibility and enable future research, we release the complete training and evaluation scripts, including model implementations, data preprocessing pipelines, and visualization code. Available at: https://github.com/ps-research/I2-Hate. The I2-Hate dataset will be released under the **Creative Commons Attribution-ShareAlike 4.0 International (CC-BY-SA 4.0)** license, upon acceptance.[3]. The Dataset will be released by the name of **I2-Hate on Hugging Face**.

Attribution requirement: All uses of I2-Hate must provide appropriate credit by citing this paper and acknowledging the dataset.

AI Use Statement

For grammar correction, we utilized AI based writing assistants, and Overleaf AI assistant, and for coding tasks, we employed Claude Sonnet 4. It is important to emphasize that the development of our ideas and the execution of the research were entirely independent of AI assistance.

[3] https://creativecommons.org/licenses/by-sa/4.0/

Appendix E : Detailed Per-Label Performance

Model Name	Affective Aggression			Derisive Trolling			Dominance & Subjugation			Ideological Expression			Performative Reinforcement			Strategic Incitement			Threat & Intimidation		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
BERT-base	0.9257	0.9529	0.9000	0.9150	0.8974	0.9333	0.8750	**0.9010**	0.8505	0.8369	0.8613	0.8138	0.9067	0.9714	0.8500	0.8489	0.8408	0.8571	0.9349	0.9294	0.9405
RoBERTa-base	**0.9674**	0.9468	**0.9889**	0.9333	0.9333	0.9333	0.8696	0.9000	0.8411	**0.8611**	0.8671	0.8552	0.9139	0.9718	0.8625	0.8317	0.8456	0.8182	0.9277	0.9390	0.9167
HateBERT	0.9371	0.9647	0.9111	0.9139	0.9079	0.9200	0.8529	0.8969	0.8131	0.8311	0.8146	0.8483	0.8980	0.9851	0.8250	0.8251	0.8389	0.8117	0.9091	0.9259	0.8929
Twitter-RoBERTa-Hate	0.9497	0.9551	0.9444	0.8820	0.8256	**0.9467**	0.8476	0.8641	0.8318	0.8108	**0.9211**	0.7241	0.8919	0.9706	0.8250	0.8438	0.8133	**0.8766**	0.9349	0.9294	0.9405
ToxicBERT	0.9302	0.9756	0.8889	0.8903	0.8625	0.9200	0.8585	0.8667	0.8505	0.8223	0.8310	0.8138	0.9079	0.9583	0.8625	0.8506	0.8506	0.8506	0.9048	0.9048	0.9048
DistilBERT	0.9364	0.9759	0.9000	0.9007	0.8947	0.9067	**0.8774**	0.8857	0.8692	0.8443	0.8472	0.8414	0.8980	0.9851	0.8250	0.8258	0.8205	0.8312	0.9302	0.9091	**0.9524**
DistilRoBERTa	0.9545	**0.9767**	0.9333	0.8931	0.8452	**0.9467**	0.8638	0.8679	0.8598	0.8058	0.8421	0.7724	**0.9290**	0.9600	**0.9000**	0.8469	0.8497	0.8442	0.9017	0.8764	0.9286
DistilBART-cnn	0.9297	0.9053	0.9556	0.9143	**0.9846**	0.8533	0.8597	0.8333	**0.8879**	0.8403	0.8462	0.8345	0.8904	0.9848	0.8125	0.8477	0.8649	0.8312	0.9080	0.8778	0.9405
DeBERTa-v3-small	0.9341	0.9239	0.9444	0.9200	0.9200	0.9200	0.8558	0.8519	0.8598	0.8414	0.8414	0.8414	0.8816	0.9306	0.8375	0.8339	0.8366	0.8312	0.8929	0.8929	0.8929
XLM-RoBERTa-base	0.9438	0.9545	0.9333	0.9167	0.9565	0.8800	0.8357	0.8396	0.8318	0.8153	0.7574	**0.8828**	0.8844	0.9701	0.8125	0.8067	0.8288	0.7857	0.9000	0.9474	0.8571
ELECTRA-small	0.8736	0.9048	0.8444	0.9103	0.9429	0.8800	0.8019	0.8300	0.7757	0.8112	0.8227	0.8000	0.8082	0.8939	0.7375	0.7958	**0.8692**	0.7338	0.8795	0.8902	0.8690
ELECTRA-base	0.9425	0.9762	0.9111	0.8917	0.8537	0.9333	0.8465	0.8426	0.8505	0.8472	0.8531	0.8414	0.9200	**0.9857**	0.8625	**0.8645**	0.8590	0.8701	**0.9398**	**0.9512**	0.9286
ConvBERT-base	0.9153	0.9310	0.9000	0.8974	0.8642	0.9333	0.8708	0.8922	0.8505	0.8403	0.8462	0.8345	0.8889	0.9315	0.8500	0.8449	0.8591	0.8312	0.9040	0.8602	**0.9524**
ALBERT-base-v2	0.9143	0.9412	0.8889	0.8974	0.8642	0.9333	0.8455	0.8230	0.8692	0.8281	0.8429	0.8138	0.8767	0.9697	0.8000	0.8312	0.8312	0.8312	0.9048	0.9048	0.9048
BART-base	0.9556	0.9556	0.9556	**0.9467**	0.9467	**0.9467**	0.8491	0.8571	0.8411	0.8571	0.8662	0.8483	0.8904	0.9848	0.8125	0.8516	0.8462	0.8571	0.9231	0.9176	0.9286

Figure 8: Per-label performance of 15 transformer models on Intent classification. Each row represents a model; columns show F1-score (F1), Precision (P) and Recall (R) for each of the 7 Intent labels. Best results per metric are highlighted in bold. Models are grouped by architecture family for easier comparison.

Model Name	Disruption of Public Discourse			Glorification of Hate			Incitement to Discrimination/Exclusion			Incitement to Violence			Misinformation/Disinformation Nexus			Normalization of Prejudice			Psychological Harm			Stigmatization & Dehumanization		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
BERT-base	0.7153	0.7840	0.6577	0.7735	0.8434	0.7143	0.8267	0.8611	0.7949	**0.9023**	0.9091	0.8955	0.8170	0.8621	0.7764	0.6795	0.7626	0.6127	0.8058	0.7809	0.8323	0.7674	0.8250	0.7174
RoBERTa-base	0.7668	0.7317	**0.8054**	**0.8557**	0.8646	0.8469	0.8398	0.8509	0.8291	0.8989	0.9023	0.8955	0.8411	0.9007	0.7888	0.7317	0.7742	0.6936	0.8177	0.7590	0.8862	**0.8035**	0.8726	0.7446
HateBERT	0.6866	0.7731	0.6174	0.7671	0.6942	**0.8571**	0.8070	0.8288	0.7863	0.8583	0.9083	0.8134	0.7919	0.8613	0.7329	0.6228	0.6460	0.6012	0.8024	0.8024	0.8024	0.7566	0.7371	0.7772
Twitter-RoBERTa-Hate	0.7552	0.7883	0.7248	0.8085	0.8444	0.7755	0.8270	0.8167	0.8376	0.8794	0.9187	0.8433	0.8198	**0.9508**	0.7205	0.7091	0.6809	0.7399	0.8150	0.7379	**0.9102**	0.7847	0.8581	0.7228
ToxicBERT	0.7526	0.7826	0.7248	0.8065	0.8523	0.7653	0.7773	0.7946	0.7607	0.8708	0.8613	0.8806	0.7766	0.8692	0.7019	0.6890	0.7290	0.6532	0.8232	0.8385	0.8084	0.7593	**0.8786**	0.6685
DistilBERT	0.7298	0.7647	0.6980	0.7795	0.7835	0.7755	0.7982	0.8296	0.7607	0.8971	0.8841	0.9104	0.7703	0.8444	0.7081	0.6786	0.6994	0.6590	0.7940	0.7917	0.7964	0.7295	0.8276	0.6522
DistilRoBERTa	0.7073	0.6480	0.7785	0.7485	0.8767	0.6531	0.7946	0.8318	0.7607	0.8812	0.9055	0.8582	0.8179	0.9154	0.7391	0.6581	0.7445	0.5896	0.8071	0.8000	0.8144	0.7267	0.8121	0.6576
DistilBART-cnn	0.7357	0.7863	0.6913	0.8283	0.8200	0.8367	0.8169	**0.9062**	0.7436	0.8708	0.8613	0.8806	0.7932	0.8731	0.7267	0.6928	0.7970	0.6127	0.8252	0.7912	0.8623	0.7828	0.7725	0.7935
DeBERTa-v3-small	0.7544	**0.8030**	0.7114	0.7831	0.8132	0.7551	0.7678	0.8617	0.6923	0.9000	**0.9286**	0.8731	0.7670	0.7303	**0.8075**	0.6645	0.7518	0.5954	0.7883	**0.8643**	0.7246	0.7621	0.7198	**0.8098**
XLM-RoBERTa-base	0.7431	0.7698	0.7181	0.7865	0.8750	0.7143	0.8070	0.8288	0.7863	0.8759	0.8571	0.8955	0.8038	0.8194	0.7888	0.6369	0.7092	0.5780	0.8080	0.7747	0.8443	0.7323	0.8440	0.6467
ELECTRA-small	0.6692	0.7607	0.5973	0.7186	0.8696	0.6122	0.7857	0.8224	0.7521	0.8741	0.8676	0.8806	0.7041	0.8868	0.5839	0.5106	0.6606	0.4162	0.7841	0.7459	0.8263	0.7207	0.8054	0.6522
ELECTRA-base	0.7015	0.7899	0.6309	0.8087	0.8706	0.7551	**0.8482**	0.8879	0.8120	0.8930	0.8832	0.9030	**0.8469**	0.8904	**0.8075**	0.6790	0.7285	0.6358	0.8207	0.8333	0.8084	0.7902	0.7923	0.7880
ConvBERT-base	**0.7770**	0.7823	0.7718	0.8136	**0.9114**	0.7347	0.8198	0.8667	0.7778	0.8993	0.8681	**0.9328**	0.8389	0.9124	0.7764	**0.7345**	0.7182	**0.7514**	0.8249	0.8176	0.8323	0.7656	0.8431	0.7011
ALBERT-base-v2	0.7548	0.7267	0.7852	0.7869	0.8471	0.7347	0.8182	0.8738	0.7692	0.8889	0.9134	0.8657	0.7973	0.8571	0.7453	0.6444	0.6795	0.6127	**0.8369**	0.8608	0.8144	0.7768	0.8323	0.7283
BART-base	0.7708	0.7632	0.7785	0.8367	0.8367	0.8367	0.8250	0.8049	**0.8462**	0.8769	0.9048	0.8507	0.8191	0.9091	0.7453	0.6644	**0.8276**	0.5549	0.8329	0.7903	0.8802	0.7861	0.8395	0.7391

Figure 9: Per-label performance of 15 transformer models on Impact classification. Each row represents a model; columns show Precision (P), Recall (R), F1-score (F1), for each of the 8 Impact labels. Best results per metric are highlighted in bold.