An Evaluation of the Tree of Thoughts Prompting Strategy for Implicit Hate Speech Detection

Giovanna Tonazzo
giovanna.tonazzo01@icatt.it – 5216819

Abstract

Implicit hate speech detection is a significant challenge in natural language processing. While large language models have demonstrated success in explicit hate speech detection, their effectiveness in identifying implicit hate speech is still an open question. In this study, we explore the application of the Tree of Thoughts (ToT) prompting strategy to improve LLM performance in detecting implicit hate speech. We conducted three experiments using Mistral LLM: (1) integrating ToT into prompting, (2) implementing a graph-based solution aligned with ToT methodology, and (3) optimizing prompts with the DSPY library. Our results indicate that ToT strategies improve binary classification performance compared to baseline models. However, for fine-grained classification and target identification, ToT introduces complexity that sometimes reduces accuracy. Optimized prompting, combining ToT with fine-tuned examples, gives the best performance in all tasks. A qualitative analysis highlights that the model frequently relies on linguistic cues, struggling to go beyond the literal meaning, and shows bias in the classification of target groups.

1 Introduction

Implicit hate speech is one of the most difficult challenges for the detection of abusive content on social media. While explicit hate speech is readily identifiable through overt language and specific keywords, implicit hate speech employs coded language, subtle inferences, and contextual cues to convey discriminatory messages, making it difficult for automatic detection methods to identify. Several theoretical frameworks have been proposed to address the complexities of implicit hate speech. Waseem et al. (2017) emphasize the interplay of denotation and connotation in understanding such language. Wiegand et al. (2021) propose a detailed analysis of the subtypes of implicit abuse, including stereotypes, dehumanization, and euphemistic constructions, each requiring context-aware analysis. Sap et al. (2020) highlight the importance of identifying implied meanings, as it is often the unspoken message that frames judgments about others. These studies underscore the need for systems capable of interpreting subtle and contextually dependent language.

Research in NLP is paying more and more attention to implicit hate speech and new datasets have been created to address the problem, for instance Caselli et al. (2020), ElSherief et al. (2021), and Ocampo et al. (2023). Despite advances in NLP, implicit hate speech detection remains a challenge. Ocampo et al. (2023) show that state-of-the-art neural network architectures, while effective for explicit content, struggle with the nuanced nature of implicit and subtle hate speech. These models have also a limited capacity to adapt to the dynamic nature of online hate. As shown by Vishwamitra et al. (2024), new waves of online hate, driven by rapidly changing sociopolitical events such as the COVID-19 pandemic and geopolitical conflicts, introduce new targets and coded language. Traditional machine learning models falter in these scenarios due to their reliance on static datasets, leading to issues of concept drift and outdated classifications. Recently, the rise of large language models (LLMs) has been a significant development in the field of NLP and prompt-based strategies have been found to effectively guide

Stage	Task	Task Description	Classification Type	Classes
1	High-Level Categorization	Classify tweets as implicit hate or not hate.	Binary	implicit_hate, not_hate
2	Fine-Grained Implicit Hate Categorization	Categorize implicit hate tweets using a six-class taxonomy.	Multi-Class	white_grievance, incitement, inferiority, irony, stereotypes_misinformation, threats_intimidation
3	Target Group Detection & Implied Statement Explanation	Identify the targeted demographic group and generate a natural language explanation of the implied meaning.	Free-Text	Target group (e.g., immigrants, women, racial groups, etc.), implied statement (paraphrased in natural language)

Table 1 – Tasks from ElSherief et al. (2021)

LLMs in hate speech detection (Gao et al., 2023). Yet, the performance of LLMs in implicit hate speech detection is still an open question.

To address these challenges, we propose an approach to implicit hate speech detection that leverages the Tree of Thoughts (ToT) prompting strategy introduced by Yao et al. (2023). This method adopts a deliberate problem-solving framework inspired by the dual process theory of human cognition and revisits foundational concepts of artificial intelligence, conceptualizing problem-solving as a tree-based search through combinatorial possibilities. Unlike chain-of-thought prompting, ToT allows for the exploration of diverse reasoning paths, evaluation of their promise, and backtracking when necessary. Our research hypothesis is that by adopting the ToT prompting strategy, it is possible to address the limitations of current NLP models and improve the performance of LLMs in implicit hate speech detection. The paper is organized as follows: after presenting the data and the tasks performed, we introduce the experiments devised to test the validity of our hypothesis, we then present a quantitative and qualitative analysis of our findings, and we conclude with final remarks and considerations for further work.[1]

2 Data

We utilize the dataset introduced by ElSherief et al. (2021)[2], a large-scale benchmark corpus designed to analyze and detect implicit hate speech. The dataset was collected from Twitter between 2015 and 2017, focusing on tweets from accounts associated with major U.S. hate groups, as identified by the Southern Poverty Law Center. The groups include black separatist, white nationalist, neo-Nazi, anti-Muslim, racist skinhead, Ku Klux Klan, anti-LGBTQ, anti-immigrant. Tweets were categorized by crowdsourced annotators into implicit hate speech and non-hate speech, while expert annotators provided fine-grained labels for implicit hate speech, using a six-class taxonomy (see Table 6 for the details). The dataset includes also natural language explanations for implied meanings and identified target groups. The final dataset contains 22,584 tweets, with 6,346 labeled as implicit hate speech.

3 Workflow

We generate a baseline by prompting an LLM to perform each task proposed in ElSherif et al. (2021). The tasks are summarized in Table 1. We conduct our experiments with Mistral Large Language Model (as of December 31, 2024) and we set the temperature parameter to 0 to enhance reproducibility. We include in

[1] The data and code used in this research are available at github.com/tonazzog/implicit-hate

[2] The dataset is available at github.com/SALT-NLP/implicit-hate

the prompts the same definition of hate speech provided to annotators in the original paper[3], as well as the guidelines and examples used in the instructions for Amazon Mechanical Turk workers. We also require the model to provide an explanation and confidence level for the classifications. We design three experiments:

Experiment 1: we add the Tree of Thought instructions to the prompts used to generate the baseline. The ToT prompt is reported in Figure 1.

Experiment 2: we design a graph-based solution that aligns with the ToT methodology, with each node representing a specific stage of thought decomposition, generation, evaluation, and refinement[4]. Following the analysis by Wiegand et al. (2021), we require the model to decompose reasoning into two branches: on one branch we ask the model to identify the subtypes of implicit abuse (e.g., stereotyping, perpetrators, euphemistic constructions), and on the other to reason about world knowledge and inferences. Based on the thoughts produced in these steps, we require the model to generate alternative explanations for the meaning of the text before proceeding to the classification. In this step, we follow the analysis by Sap et al. (2020), who highlight the importance of identifying the implied meaning of a statement. The following node corresponds to the evaluation stage, where the model ranks the generated branches based on their likelihood of representing implicit abuse. If the model is uncertain or detects ambiguity, it can deepen and refine the analysis. This involves expanding less explored branches or considering additional context or external factors to improve interpretation. We finally require the model to produce the classification as implicit hate speech or not hate speech and to provide an explanation and a confidence level for the conclusion. If the post is classified as implicit hate speech, the model proceeds to stages 2 and 3 (evaluation of implicit hate classification and identification of hate target and implied meaning) and we provide the explanation generated in the previous step as a further source of knowledge. The graph is presented in Figure 2. Our strategy is similar to the one proposed by Vishwamitra et al. (2024), but their solution follows a decision-tree like approach, and we strongly rely on passing information to downstream nodes.

Experiment 3: we perform prompt optimization with DSPY[5] library described in Khattab et al. (2023) on the prompts used in the first experiment. We use the MIPROv2 optimizer for both prompt and samples optimization (Opsahl-Ong et al., 2024). The settings for each stage are summarized in Table 7.

4 Results and Discussion

For Stage 1 and 2 results are evaluated in terms of F1 macro score (F1), precision (P), recall (R), and accuracy (A). For Stage 3, we evaluate target and meaning similarity in terms text similarity metrics, including BLEU unigram (B1), BLEU bigram (B2), BLEU (B), Meteor (M), NIST (NI), f1-score of Rouge-1 (R1), Rouge-2 (R2), and Rouge-L (R-L), f1-score of BERTScore (BS) and BLEURT (BL). The size of our test sets is around 15% of the original datasets (3000, 1000, and 1000 posts respectively for stage 1, 2 and 3).

4.1 Results for Stage 1

Strategy	F1	P	R	A
Baseline	0.600	0.701	0.683	0.602
Prompt ToT	0.641	0.710	0.709	0.641
Graph ToT	0.676	0.719	0.730	0.676
Optimized	0.729	0.735	0.759	0.736

Table 2 – Results for stage 1

The results for stage 1 (binary classification) show that introducing the ToT methodology

[3] X/Twitter's policy on hateful conduct

[4] We implement the solution with LangGraph library

[5] DSPY official documentation site

improves classification performance over the baseline, with the Optimized strategy yielding the best results across all metrics, surpassing the F1 score obtained by ElSherief et al. (2021) fine-tuning BERT. The Graph ToT approach also shows improvements, particularly in recall, suggesting that structured reasoning paths enhance the model's ability to capture implicit hate speech (see also Figure 3 for a comparison of the metrics). Our results differ from Zhang et al. (2024), who concluded that prompting strategies did not lead to a noticeable difference in performance. However, we reach the same conclusions regarding the model's high confidence in the generated answers (overall confidence is 0.84, 0.79 for misclassified posts) and for the high rate of implicit hate false positives (see the confusion matrix for the best performing strategy in Figure 4). A closer examination of these posts highlights common patterns. The model tends to over-rely on specific keywords or phrases that are frequently associated with hate speech, even when the broader context does not support such a classification. For instance, the post *"since your great white hope is already backpedaling will you go back to being silent ? #trump #workingclass #14words"* was flagged due to the presence of the phrase "great white hope" and the hashtag "#14words," which are associated white supremacy. However, the intent of the post could have been more about political critique than explicit hate. Another pattern we observed is the misinterpretation of historical discussions as hate speech. The post *"white people have been through slavery though"* could be interpreted as only making a historical comparison. The model also struggles to differentiate between criticism and hate speech. For example, the post *"shame on hindus who saw / supported pk"* could be interpreted as criticism of specific actions, rather than an attack on the group's identity. The model has difficulty recognizing satire or irony. The post *"a make america white evangelical male event ."* could be a sarcastic critique of conservative positions rather than an endorsement of exclusion. Finally, the model often assumes negative intent in posts that might be only neutral or informative. The post *"i have .. and have also read the crimes that zionist bankers inflicted on germans in weimar republic .."* was flagged for its reference to "Zionist bankers," but annotators did not see it to endorse harmful stereotypes.

4.2 Results for Stage 2

Strategy	F1	P	R	A
Baseline	0.470	0.493	0.468	0.552
Prompt ToT	0.408	0.418	0.419	0.492
Graph ToT	0.367	0.424	0.382	0.448
Optimized	0.516	0.558	0.510	0.581

Table 3 – Results for stage 2

Unlike Stage 1, where ToT strategies consistently improved results, fine-grained classification presented a more complex scenario. The Optimized approach achieved the highest performance, but other ToT strategies exhibited a decrease in classification accuracy compared to the baseline (see Table 3 and Figure 5). The results are lower compared to ElSherief et al. (2021). This suggests that while multi-branch reasoning benefits high-level classification, it may introduce noise when dealing with fine-grained categories that require precise examples rather than complex reasoning paths. This conclusion is supported by the fact that the optimizer did not modify the original prompt, thus showing that only the examples influenced the result. The confusion matrix for the Optimized strategy highlights that the model particularly struggles to distinguish between "white grievance" and "incitement" (Figure 6). The misclassification of these posts can be attributed to several recurring patterns in the model's interpretation. The model tends to focus heavily on identifying expressions of grievance, frustration, or victimhood related to white identity, often overlooking the presence of language that annotators interpreted as incitement to violence or promotion of hate groups. For instance, in the post *"are you talking about the same white supremacists*

Strategy	B1	B2	B	R1	R2	R-L	M	NI	BS	BL
Baseline	0.558	0.264	0.126	0.648	0.169	0.648	0.365	0.207	0.947	0.600
Prompt ToT	0.514	0.254	0.120	0.598	0.162	0.597	0.354	0.202	0.938	0.553
Graph ToT	0.493	0.234	0.114	0.593	0.143	0.592	0.342	0.180	0.935	0.553
Optimized	0.600	0.328	0.142	0.682	0.253	0.682	0.416	0.264	0.953	0.593

Table 4 – Results for stage 3 – Targeted Group similarity

Strategy	B1	B2	B	R1	R2	R-L	M	NI	BS	BL
Baseline	0.312	0.201	0.111	0.425	0.190	0.413	0.308	0.651	0.914	0.486
Prompt ToT	0.276	0.170	0.084	0.376	0.158	0.364	0.304	0.609	0.905	0.468
Graph ToT	0.271	0.167	0.083	0.376	0.155	0.363	0.297	0.592	0.906	0.473
Optimized	0.410	0.286	0.164	0.502	0.303	0.489	0.420	0.895	0.925	0.521

Table 5 – Results for stage 3 – Implied Statement Similarity

who saved your people from genocide?", the model interprets the defense of white supremacists as a grievance rather than recognizing it as a defense of extremist behavior, which is a form of incitement. Another pattern is the model's struggle with forms of incitement where the call to violence or support for extremist groups is implied through context or historical references. In the post *"that's why white consciousness has exploded. whites are realizing this is an existential moment,"* the model sees the "explosion" of white consciousness as a grievance, missing the urgency and potential call to action implied by the term "existential moment" which could incite group unity and action. The model's reliance on certain keywords or phrases that are strongly associated with "white grievance" can bias its classification. In the post *"who is promoting 'white privilege'? #withjewswelose"*, the model focuses on the grievance against "white privilege" and the association with Jewish people, missing the potential incitement in the hashtag "#withjewswelose," which could promote anti-Semitic sentiments.

4.3 Results for Stage 3

As in Stage 2, in Stage 3 the Optimized approach achieved the highest performance, but other ToT strategies exhibited a decrease in text similarity scores compared to the baseline (see Table 4 and Table 5). Comparing our results for implied statement similarity with Huang et al. (2023), we see that while we don't improve on n-gram based metrics, we obtain better results in embeddings-based metrics, showing that the model can capture adequately the overall meaning of a statement. For target group identification, the best performing method identifies very well some subgroups (i.e. Jews and Muslims), while other subgroups are specified in even more detail (i.e. immigrants are qualified as illegal immigrants, Mexicans). For other groups, the model gives completely different interpretations (see Figures 7 and 8), as in cases where it identifies "non-white people" or "black people" instead of "white people" as target. Posts that mention both "white" and "black" groups often result in the model assigning negative implications to black people, instead of recognizing criticism directed at white people. For example, in the post: *"maybe you should build your thriving black communities far away where white people can't get to them"*, the model focused on the explicit mention of "black communities" rather than interpreting it as a critique of white interference. Another common issue is that instead of capturing statements that imply

white fear or responsibility, the model shifts the blame onto black people. For instance, *"new video: 'madness at #mizzou: black hysteria and white cowardice' #tcot #altright"*, the phrase "black hysteria" likely triggered the model to classify black people as the target rather than recognizing that the post implies white fear. Some posts involve white people being criticized or portrayed as the dominant group, but the model reverses the target. In this post: *"No white women are on average closer to white men in intelligence than black men or women,"* the statement asserts white superiority, but the model reframes it as an attack on black inferiority.

4.4 Conclusions

Our results demonstrate that the performance of ToT strategies varies depending on the complexity of the task and the level of reasoning required. Our findings suggest that ToT prompting strategies improve performance in tasks that require multi-step reasoning but may introduce unnecessary complexity in more fine-grained classification problems, where example-based learning proves more effective. We also showed that prompt optimization with DSPY yielded the best performance on all tasks. A qualitative analysis of misclassified posts shows systematic errors, including the model's over-reliance on the presence of specific linguistic cues without considering broader context, and its struggle to differentiate between criticism, discussion of historical or political issues, irony and satire and actual hate speech. The model also demonstrated biases in target group identification, sometimes attributing negative implications to marginalized groups instead of recognizing critiques of dominant groups. Future work should explore generalizability across different LLMs and assess the impact of temperature variations on reasoning diversity, for example setting a higher temperature for the generative steps to enhance creativity, combined with a lower temperature in the decision tasks. Another promising direction is experimenting with alternative DSPY configurations, as our study relied exclusively on the "heavy" auto-setting without attempting parameter optimization techniques.

References

T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, and M. Granitzer. 2020. I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

M. ElSherief, C. Ziems, D. Muchlinski, V.i Anupindi, J. Seybolt, M.De Choudhury, and D. Yang. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

K. Guo, A. Hu, J. Mu, Z. Shi, Z. Zhao, N. Vishwamitra and H. Hu. 2023. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pp. 1568-1573. IEEE.

F. Huang, H. Kwak, and J. An. 2023. Chain of Explanation: New Prompting Method to Generate Quality Natural Language Explanation for Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 90–93.

O. Khattab, A. Singhvi, P.i Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan A, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, H. Miller, M. Zaharia, C. Potts. DSPy: Compiling Declarative Language Model Calls into State-of-the-Art Pipelines. In *The Twelfth International Conference on Learning Representations*, 2024.

N. B. Ocampo, E. Sviridova, E. Cabrio, and S. Villata. 2023. An In-depth Analysis of Implicit and Subtle Hate Speech Messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

K. Opsahl-Ong, M. J Ryan, J. Purtell, D. Broman, C. Potts, M.Zaharia, and Omar K.. 2024. Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366, Miami, Florida, USA. Association for Computational Linguistics.

M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

N. Vishwamitra, K. Guo, F. T. Romit, I. Ondracek, L. Cheng, Z. Zhao. 2024. Moderating New Waves of Online Hate with Chain-of-Thought Reasoning in Large Language Models. In *2024 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, 2024, pp. 788-806.

Z. Waseem, T. Davidson, D. Warmsley, and I. Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

M. Wiegand, J. Ruppenhofer, and E. Eder. 2021. Implicitly Abusive Language – What does it actually look like and why are we not getting there?. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. 2023. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 517, 11809–11822.

M. Zhang, J. He, T. Ji, and C. Tien Lu. 2024. Don't Go To Extremes: Revealing the Excessive Sensitivity and Calibration Limitations of LLMs in Implicit Hate Speech Detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12073–12086, Bangkok, Thailand. Association for Computational Linguistics.

Appendix

Hate Category	Definition	Example
White Grievance	Includes frustration over a minority group's perceived privilege and casting majority groups as the real victims of racism. This language is linked to extremist behavior and support for violence.	*"Black lives matter and white lives don't? Sounds racist."*
Incitement to Violence	Includes flaunting ingroup unity and power or elevating known hate groups and ideologies.	*"White brotherhood"* (ingroup unity) or *"Hitler was Germany – Germans shall rise again!"* (elevating nationalism and Nazism).
Inferiority Language	Implies one group or individual is inferior to another, including dehumanization and toxification. Related to assaults on human dignity, dominance, and declarations of superiority of the in-group.	*"It's not a coincidence the best places to live are majority white."*
Irony	Use of sarcasm, humor, and satire to attack or demean a protected class or individual. Commonly used by modern online hate groups to mask hatred and extremism.	*"Horrors... Disney will be forced into hiring Americans."*
Stereotypes and Misinformation	Associates a protected class with negative attributes like crime or terrorism. Also includes misinformation that feeds stereotypes, such as Holocaust denial.	*"Can someone tell the black people in Chicago to stop killing one another before it becomes Detroit?"*
Threatening and Intimidation	Conveys a speaker's commitment to a target's pain, injury, damage, loss, or violation of rights. Includes implicit threats to rights, freedoms, and opportunities.	*"All immigration of non-whites should be ended."*
Other	None of the above was identified.	*N/A*

Table 6 – Fine Grained Taxonomy of Implicit Hate Speech

```
Follow this procedure:

Imagine three different experts are answering this question.
They will brainstorm the answer step by step reasoning carefully and taking all facts into consideration.
All experts will write down 1 step of their thinking, then share it with the group.
They will each critique their response, and the all the responses of others.
They will check their answer on based on the nature of the language and intent.
Then all experts will go on to the next step and write down this step of their thinking.
They will keep going through steps until they reach their conclusion taking into account the thoughts of the other experts.
If at any time they realise that there uncertainty in their logic they will backtrack to where that uncertainty occurred.
If any expert realises they're wrong at any point then they acknowledges this and start another train of thought.
Each expert will assign a likelihood of their current assertion being correct.
Continue until the experts agree on the single most likely classification.
```

Figure 1 – Tree of Thoughts prompt instructions for experiment 1

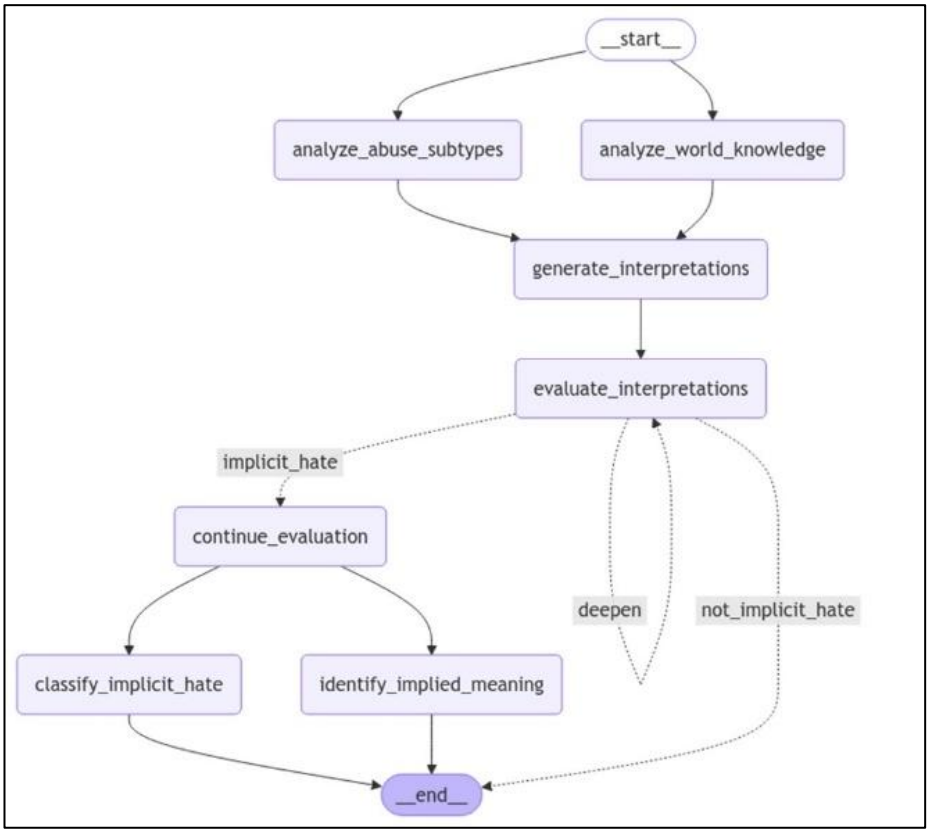


Figure 2 – Graph of Thoughts implementation for experiment 2

	Tempera-ture	**Opti-mizer**	**Sam-ples**	**Auto Setting**	**Labeled Demos**	**Boot-strapped Demos**	**Metric**
Stage 1	0.0	MiproV2	100	Heavy	8	8	Accuracy
Stage 2	0.0	MiproV2	300	Heavy	20	20	Accuracy
Stage 3	0.0	MiproV2	100	Heavy	20	20	BERT F1 Score > 0.9

Table 7 – Settings of DSPY MIPRO optimizer in experiment 3

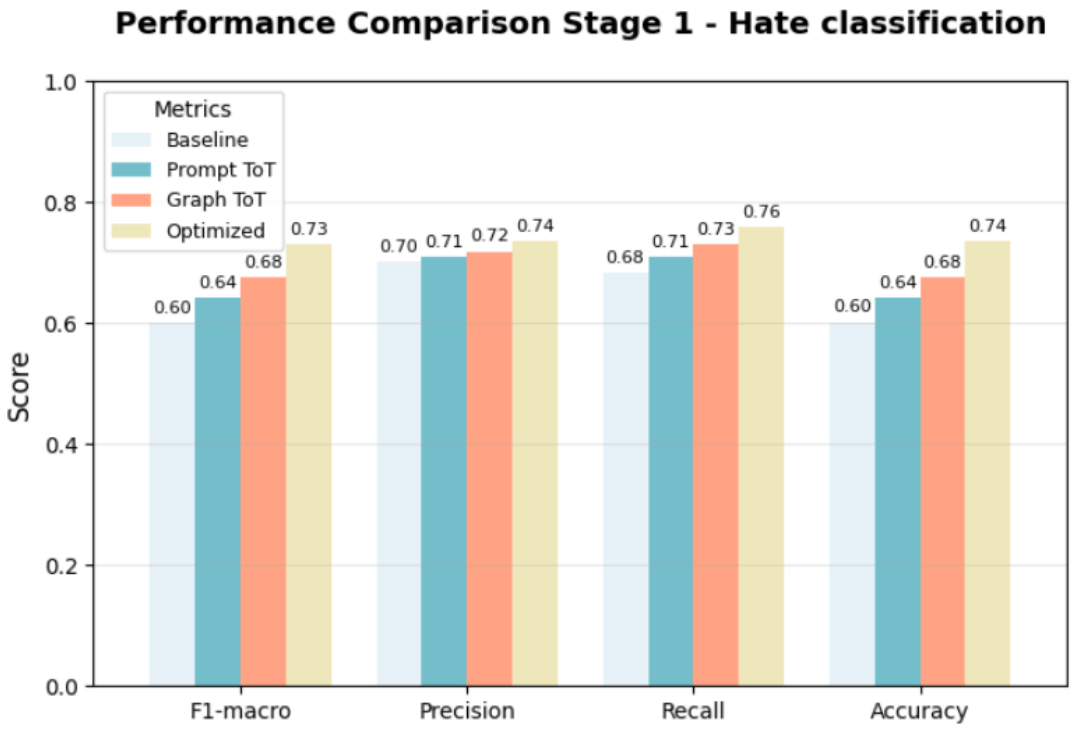


Figure 3 – Performance comparison stage 1

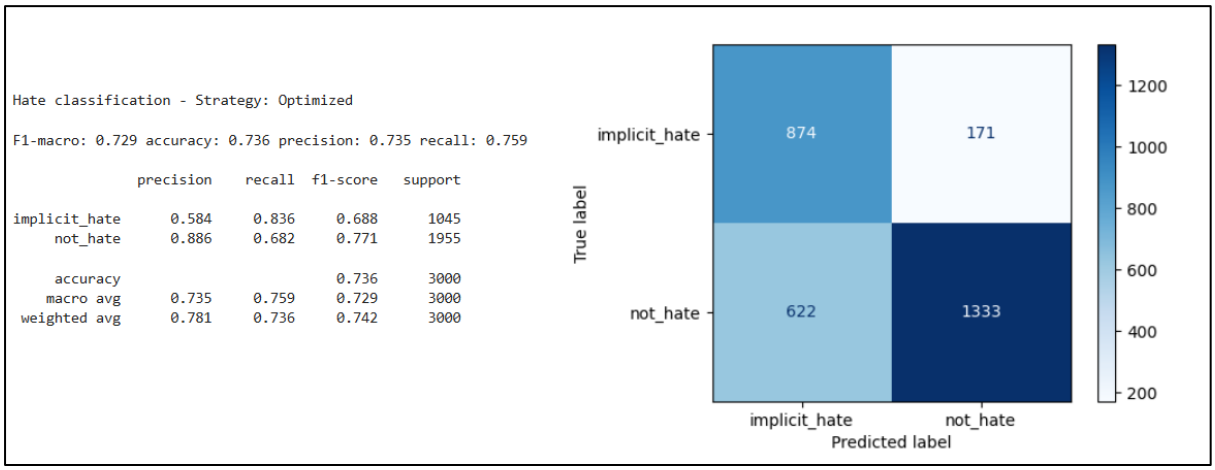


Figure 4 – Confusion matrix stage 1 (Optimized strategy)

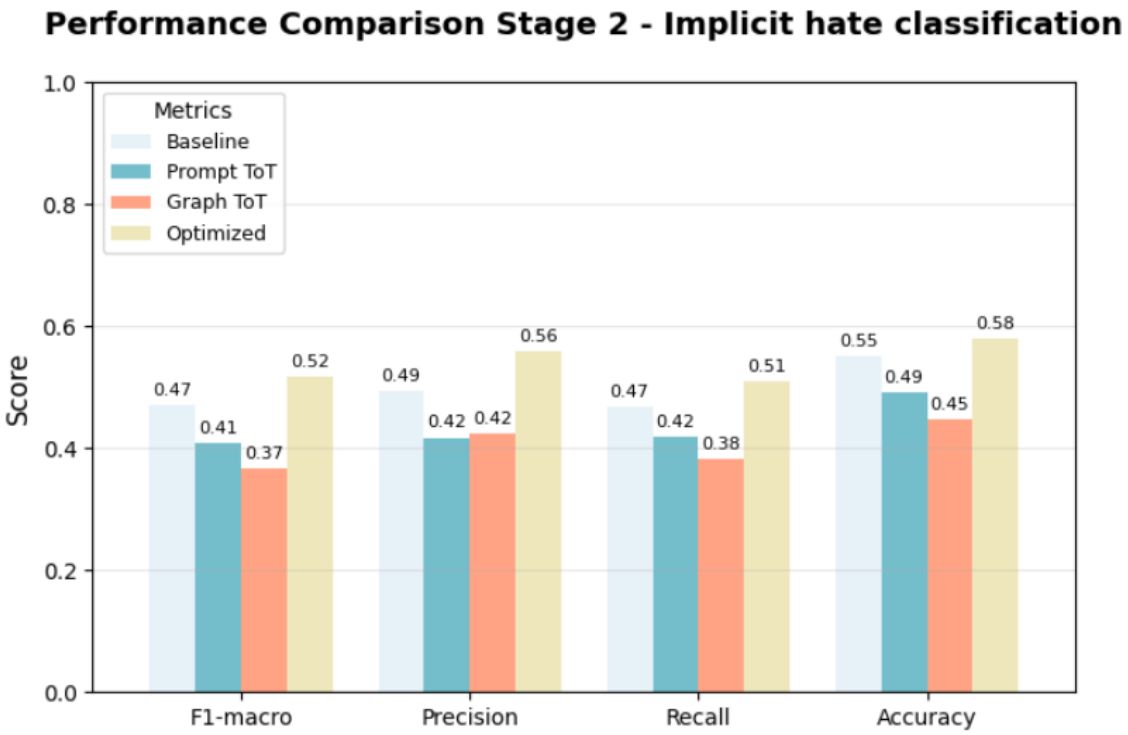
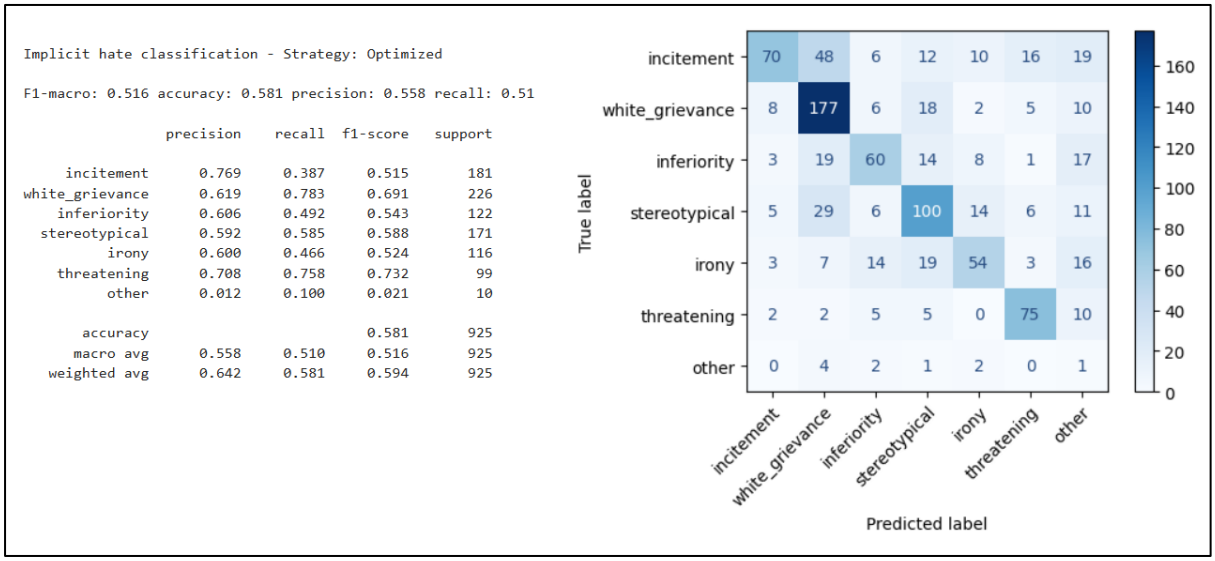


Figure 5 – Performance comparison stage 2



Implicit hate classification - Strategy: Optimized

F1-macro: 0.516 accuracy: 0.581 precision: 0.558 recall: 0.51

	precision	recall	f1-score	support
incitement	0.769	0.387	0.515	181
white_grievance	0.619	0.783	0.691	226
inferiority	0.606	0.492	0.543	122
stereotypical	0.592	0.585	0.588	171
irony	0.600	0.466	0.524	116
threatening	0.708	0.758	0.732	99
other	0.012	0.100	0.021	10
accuracy			0.581	925
macro avg	0.558	0.510	0.516	925
weighted avg	0.642	0.581	0.594	925

Figure 6 – Confusion matrix for stage 2 (Optimized strategy)

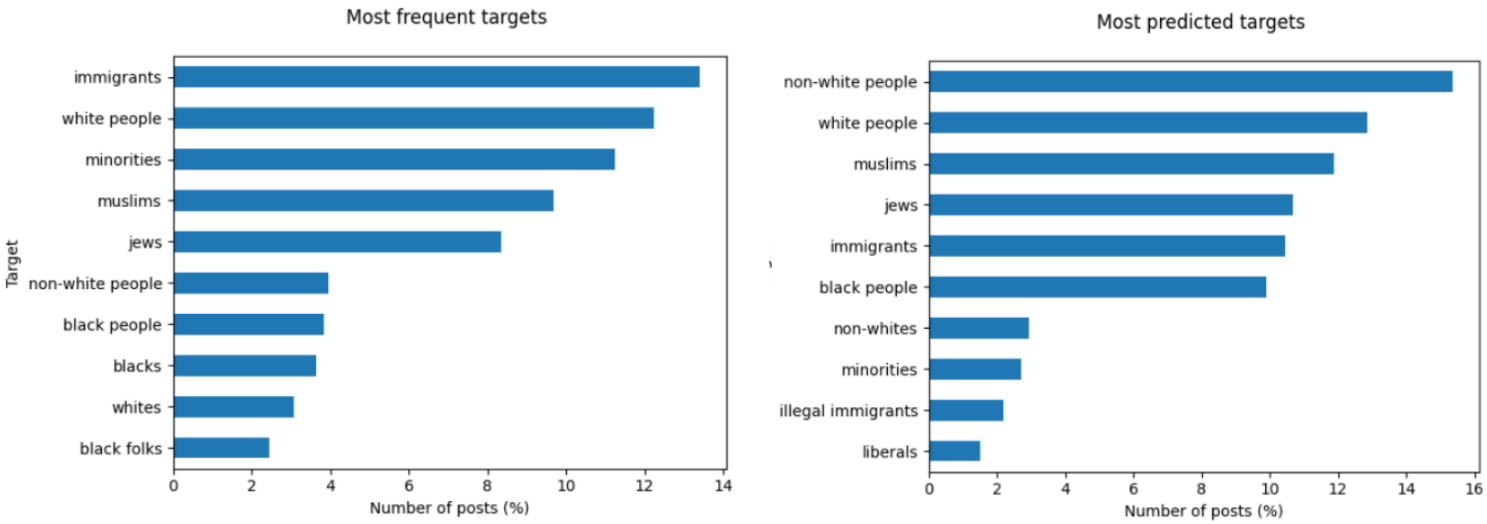


Figure 7 – Most frequent target groups in gold standard vs predicted (Optimized strategy)

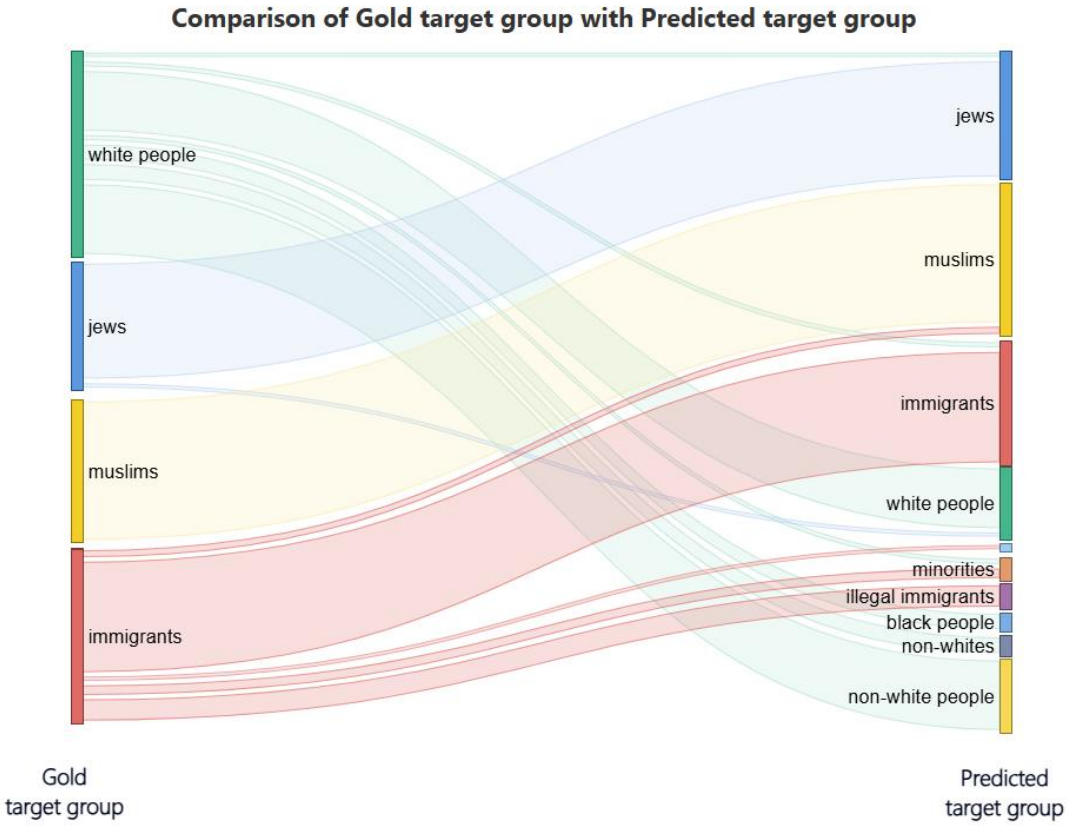


Figure 8 – Differences in classification of targeted group (Optimized strategy)