Tuning Hypothesis Creation: Combining Discrete and Continuous Spaces for Zero-Shot Hate Speech Detection

Lorenzo Puppi Vecchi[1][0009−0003−9483−2026],
Sylvio Barbon Junior[2][0000−0002−4988−0702], and
Emerson Cabrera Paraiso[1][0000−0002−6740−7855]

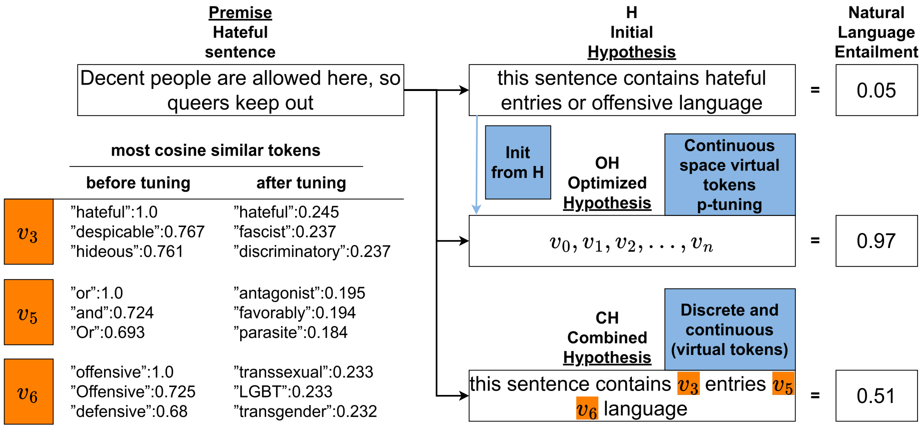
[1] Graduate Program in Informatics - Pontificia Universidade Católica do Paraná - Brazil {lorenzo.vecchi,paraiso}@ppgia.pucpr.br
[2] Department of Engineering and Architecture - University of Trieste - Italy
sylvio.barbonjunior@units.it

Abstract. Hate speech detection is a crucial endeavor in maintaining the safety of online spaces, but the effectiveness of supervised approaches hinges mainly on the availability of annotated data. Prior research has explored the utility of natural language inference (NLI) models for zero-shot hate speech detection (ZSHSD), which leverages the capacity of these models to learn semantic relationships and adapt to downstream task without relying on large annotated datasets. NLI models assess if a premise sentence logically entails a hypothesis sentence, relying on precise hypothesis design to achieve adequate downstream task performance. Existing frameworks that use NLI model for ZSHSD rely on multiple inferences with different hypothesis to extract characteristics to achieve desirable outcomes. In light of the challenges surrounding ZSHSD and the method of relying on discrete tokens to design hypotheses, we aim to optimize and identify ideal intermediate representations by applying p-tuning techniques. On HateCheck, a fully tuned hypothesis led to a 18.8 percentage point (pp) accuracy improvement, compared to a discrete designed hypothesis. Our work surpassed prior work by achieving a 5.6 pp accuracy enhancement, outperforming previous approaches that required multiple inferences. Also, the optimized tokens uncover relations to broader aspects of hate speech, offering insights for hypothesis design.

Keywords: Hate Speech · Natural Language Inference · Zero-Shot Text Classification · P-tuning

1 Introduction

Recently, there has been a rise in harmful content like hate speech and misinformation on the internet. This increase may not just be due to more people using blogs and social media but could also result from global political polarization and social media algorithms that amplify divisive content to drive engagement.


Premise
Hateful
sentence
Decent people are allowed here, so queers keep out
H
Initial
Hypothesis
this sentence contains hateful entries or offensive language
=
Natural
Language
Entailment
0.05
Init
from H
OH
Optimized
Hypothesis
Continuous
space virtual
tokens
p-tuning
$v_0, v_1, v_2, \ldots, v_n$
=
0.97
CH
Combined
Hypothesis
Discrete and
continuous
(virtual tokens)
this sentence contains v_3 entries v_5 v_6 language
=
0.51
most cosine similar tokens

	before tuning	after tuning
v_3	"hateful":1.0 "despicable":0.767 "hideous":0.761	"hateful":0.245 "fascist":0.237 "discriminatory":0.237
v_5	"or":1.0 "and":0.724 "Or":0.693	"antagonist":0.195 "favorably":0.194 "parasite":0.184
v_6	"offensive":1.0 "Offensive":0.725 "defensive":0.68	"transsexual":0.233 "LGBT":0.233 "transgender":0.232

Fig. 1. Combining standard hypothesis and p-tuned virtual tokens

This makes online safety and content moderation crucial to maintaining civility and protecting users (19; 9).

Before, detecting hate speech mainly relied on classification models that needed lots of labeled data (21; 2). Recently, Natural Language Inference (NLI) models, which are used for tasks like zero-shot classification (7), are being explored. NLI models show promising results for zero-shot hate speech detection (ZSHSD) because they capture relationships between sentences, which helps adapt content moderation to the ever-changing online world. In our study, we're using the term "hate speech" to cover a wide range of offensive content created by users. We chose this term because it reflects hurtful communication that spreads hostility through stereotypes (15).

In NLI models, generating hypotheses is crucial and affects tasks like ZSHSD (20). Even small changes in how a hypothesis is worded can greatly impact the model's performance. For example, compare "Climate change is a pressing global issue" to "Addressing climate change requires urgent global cooperation." Both statements express concern about climate change, but the second one emphasizes the need for global collaboration, which could lead to different conclusions and outcomes in ZSHSD.

To better handle the complexity of forming hypotheses, current frameworks using NLI models take a step-by-step approach. They use multiple hypotheses to understand different aspects of text, like identifying hate speech in a tweet. For example, one hypothesis might spot explicit language ("This tweet has offensive words"), while another digs deeper into implied meanings ("The tone of this tweet encourages hostility"). This method helps the model grasp various aspects of text, which is crucial for accurate ZSHSD performance (7).

Traditionally, approaches used simple hypotheses like "That's hate speech" which, while somewhat effective, often oversimplify the complexities of identifying hate speech. These basic hypotheses struggle to capture the wide range of nuances, subtleties, and context-dependent variations in hate speech (13; 7). This has led to a growing recognition of the need for more sophisticated hypothesis generation in ZSHSD.

In our work, we aim to tackle the difficulties linked with hypothesis generation in NLI models. We introduce a method that optimizes hypothesis creation through a process known as "p-tuning" (10), which starts with a standard hypothesis prompt but iteratively refines it to suit the complexities of ZSHSD. P-tuning allows us to adapt the hypothesis to the specific characteristics and nuances of the task, ensuring that it encapsulates a broader and more representative range of hate speech expressions.

General objective: This article aims to tackle the challenge in hypothesis generation for Natural language Inference when used for Zero-shot classification. It uses "p-tuning" as a method to optimize hypothesis creation for addressing the nuances involved in Hate Speech Detection.

In this process, we start with a initial Hypothesis (H), choosing the most accurate one, out of different hypothesis sentence designed to detect hate speech using NLI. Results can be seen in Table 2.

Next, the Optimized Hypothesis (OH) is generated, which starts from the best performed hypothesis of H, and is optimized using p-tuning. Different strategies of using the resulting virtual tokens can be seen in Table 4 and Table 5.

Finally, the Combined Hypothesis (CH) is formed by integrating the most relevant tokens from OH with the original H. The accuracy of different virtual tokens inserted into the original hypothesis can be seen in Table 4 and Table 5, and a general benchmark, comparing with previous works, can be seen in Table 6. A visual scheme of this idea is illustrated in Figure 1.

Our findings highlight the effectiveness of our optimized hypothesis (OH), as certain virtual tokens within the continuous embedding space closely approximate broader hate speech concepts. This suggests that including these virtual tokens (OH) in our original hypothesis (H) formulation can enhance the model's capacity to detect and address hate speech more comprehensively. This opens up three promising avenues for future research:

First: researchers can directly use our virtual token weights alongside RoBERTa Large MNLI to enhance their models' ZSHSD[3].

Second: they can employ selected model tokens from our study to adapt their models for improved classification.

Lastly: our research encourages the creation of hypotheses based on concepts' mean, expanding the method's applicability.

2 Background

The evaluation of hate speech detection models often suffers from overestimation issues, impacting both state-of-the-art (SOTA) and baseline models. When tested on different datasets, these models typically show reduced performance, highlighting a gap between test sets and real-world data (18; 1; 24; 5; 23).

[3] https://anonymous.4open.science/r/Combining-Discrete-and-Continuous-Spaces-for-Zero-Shot-Hate-Speech-Detection-FB34/README.md

Various approaches to hate speech detection have been explored, including traditional machine learning models such as character n-gram Logistic Regression (8), Support Vector Machines (17), and shallow networks with pre-trained embeddings like MLP with Byte-Pair Encoding (BPE) (22). Generally, these simpler models underperform compared to deep neural networks.

Prior to 2019, recurrent neural networks were commonly used as SOTA models (8). The introduction of BERT (Bidirectional Encoder Representations from Transformers) by (3) marked a significant advancement, establishing BERT and its variants as the new SOTA in hate speech detection (6; 4).

Although BERT and its derivatives have improved in capturing contextual information and linguistic nuances, our study focuses on NLI (Natural Language Inference) models. This choice is based on the rationale that explicit logic learning through textual entailment can reduce bias and enhance the recognition of social communities (12). NLI models excel in hate speech detection by leveraging logical training, which improves context-aware zero-shot hate speech detection (ZSHSD) systems. Unlike pre-trained sentence encoders that may perpetuate stereotypes, NLI models help mitigate bias present in training data (14).

Using NLI models for predictive tasks involves transforming the target task into an NLI format, essentially a fine-tuning task. The model assesses a premise and a hypothesis to determine their logical relationship—whether the premise implies, contradicts, or is neutral to the hypothesis. A recent method applies NLI models for zero-shot topic classification, where the input text serves as the premise and each topic has a corresponding hypothesis like "This text is about <topic>." Here, "neutral" and "contradiction" labels merge into "not-entailment," meaning an "entailment" prediction indicates relevance to the topic, while "not-entailment" suggests irrelevance. It's important to note that an entailment prediction indicates alignment with certain model artifacts, not necessarily correctness (25).

3 Proposed Approach

Our method focused on identifying the most effective hypothesis for assessing the existence of hate speech in a given premise. Rather than initiating the optimization process from entirely ran'dom weights, we leveraged the best-performing results obtained from prior experiments, as seen in Table 2. This strategic starting point aimed to expedite the optimization process and build upon existing knowledge.

3.1 Hypothesis Optimization

In the realm of NLI models, a hypothesis is a sentence, that is encoded in a sequence of tokens. Each token is represented as a integer and, after passing thought the models embedding layer, every token becomes a vector. When working with p-tuning, instead of changing the initial hypothesis sentence, the continuous vectors as optimized directly. This optimization creates whats is called as "virtual tokens", since then dont perfectly resemble a proper token anymore.

This means that the algorithm learns to adjust the continuous vector representation of initially passed hypothesis to improve its performance in ZSHSD. A general scheme of the process used to optimize the initial hypothesis H and test the different combinations of hypothesis tokens, can be seen in Table 2.

Let's denote H as initial hypothesis vectors (selected from the best results of Table 2) and OH as the Optimized hypothesis vectors. The learning process involves adjusting the weights θ of the model to find an optimal hypothesis vector OH, that minimizes a specific loss function, typically associated with ZSHSD. Mathematically, this can be represented as:

$$OH = argmin_{\theta} L(P, H, \theta) \tag{1}$$

Here, L represents the loss function, P denotes the task-specific premise, H signifies the initial hypothesis vectors, and θ symbolizes the model's weights. The objective is to find the optimal set of weights that minimizes the loss function, thereby producing the optimized hypothesis vectors OH.

NLI models typically feature a softmax layer at the end, which serves to produce three primary outputs denoting the potential relationships between a given premise P and hypothesis H: contradiction, neutrality, and entailment. These outputs, akin to probabilities, are normalized to ensure that their collective sum equals 1, effectively indicating the likelihood of each possible relationship.

The goal of training an NLI model is to minimize the error in predicting these probabilities, ensuring that the model accurately identifies the logical relationship between the premise and hypothesis. This is typically achieved through the use of a categorical cross-entropy loss L, which is defined as:

$$L = -\sum_{c=1}^{3} y_i \log(p_i) \tag{2}$$

In the context of binary hate speech detection, the loss function in this case is optimized to increase entailment output when hate is present, i.e. $c = 3$, and maximize contradiction output when hate is not present, i.e. $c = 1$. Given that $c = [1, 2, 3]$ represents, contradiction, neutrality and entailment, respectively.

After optimizing our hypothesis, we proceeded with a structured sequence of experiments to assess both the individual and combined performance of optimized tokens. These experiments were designed to gauge the efficacy of our approach in diverse contexts, thereby illuminating its capacity to discern hate speech.

In our study, we used the RoBERTa Large model[4] (11), which had undergone fine-tuning on a diverse range of text sources, including Wikipedia, the Book Corpus, and the MultiNLI dataset. The training of this model was conducted on an NVIDIA 4090 RTX GPU, spanning a duration of approximately 12 hours, with a training regimen extending over 20 epochs. For optimization, we employed the AdamW optimizer and implemented a linear decay learning rate schedule, setting the initial learning rate at 1e-3. It is noteworthy that our training process

[4] https://huggingface.co/roberta-large-mnli

required significantly less time and computational resources in comparison to previous work (19). Notably, we optimized our approach to achieve efficient resource utilization by employing p-tuning solely on the hypothesis, in contrast to traditional whole-model fine-tuning methodologies. This strategic adaptation likely contributed to the reduced time and computational resources required for our training process.

3.2 Hypothesis Similarity Analysis

Our initial phase involved scrutinizing the optimized hypothesis and assessing its similarity to the original tokens before optimization. This allowed us to gauge the extent to which our optimizations retained the core meaning of the original text. We measured this distance using cosine similarity with the base model's token embeddings.

3.3 Token Importance Assessment

In the second phase, we delved into the importance of each individual optimized token. We examined the outcomes of using these tokens in three distinct manners: isolated (one at a time), inserted individually, and cumulatively. Following the idea of **CH** presented in Figure 1.

Isolated (Token-Level): In the "Isolate" experiment, we isolated one token at a time from the optimized hypothesis. The primary objective was to determine whether the isolated tokens exhibited improved performance, particularly concerning terms associated with hate speech, such as "hateful" and "offensive."

Individual (Contextual Insertion): In the "Individual" experiment, once the tokens have been optimized in this new continuous space, they are reinserted back into the standard hypothesis structure. This means replacing the original tokens with their tuned representations in the hypothesis. Let's say the token "hateful" is initially represented as $[0.2, 0.5, -0.3]$, and after tuning, it becomes $[0.8, -0.1, 0.7]$. After reinserting the tuned tokens, the hypothesis may look like this:

"this sentence contains $[hateful]$ **entries or offensive language"**

"this sentence contains $[hateful]$ **entries or** $[offensive]$ $[language]$**"**

Here, $[hateful]$, $[offensive]$ and $[language]$ represent the transformed versions of the original tokens "hateful", "offensive" and "language".

The motivation behind this approach is to leverage a more nuanced and contextually aware representation of the key terms in the hypothesis. By doing so, this study aims to measure whether this enhanced understanding of specific language cues can lead to improved performance in hate speech detection as a NLI tasks. The idea is that the tuned representations may capture subtleties and context better than the original, static tokens.

Cumulative (Sequential Insertion): The "Cumulative" experiment involved the incremental addition of the optimized tokens in the natural order of the sentence. While we anticipated an overall improvement, the objective here was to analyze performance peaks corresponding to the introduction of each new optimized token. This step provided insights into the cumulative impact of our method on ZSHSD.

3.4 Benchmark

In the final phase of our experimentation, our focus shifted towards comparative analysis. Specifically, we sought to assess the efficacy of our methodology, which incorporated the optimized hypothesis and tokens, in contrast to established benchmarks as presented in prior studies (19; 16; 7).

Dataset The dataset utilized in this study comprises two distinct sources, each contributing valuable insights into ZSHSD. In aggregate, the total dataset employed in our research comprises 40,772 instances. It consists of 22,571 instances labeled as "hateful" and 18,201 instances categorized as "non-hateful." This diverse and multifaceted dataset serves as a foundation for our explorations into hate speech detection, allowing us to analyze and evaluate the performance of the method across a wide range of real-world hate speech scenarios.

Table 1. Dataset - Data sources and number of examples

	hateful	non-hateful	Total
Learning from the Worst: Train	17740	15184	32924
Learning from the Worst: Test	2268	1852	4120
HateCheck: Test	2563	1165	3728
Total	22571	18201	40772

Learning from the Worst (19) is a process to create datasets for training better hate detection models. It uses both humans and models. The dataset has 37,044 instances: 20,008 are labeled "hateful", and 17,036 are “non-hateful." We followed the authors' setup, training the model on the four rounds made available by the dataset (R1, R2, R3, and R4) and testing on each rounds splits separately.

HateCheck (16) is a suite of tests for evaluating hate speech detection models. It has 29 test functionalities based on previous research and civil society input. The dataset contains 3,728 instances: 2,563 are "hateful," and 1,165 are "non-hateful." We used the best model accuracy scores for individual tests based on the original work (16). For the average best performance we used the accuracy of "All" the different hypotheses combined as proposed by (7).

4 Results

Through an evaluation of the optimized hypothesis and token representations, as well as their impact on ZSHSD performance, we aim to provide a detailed insight into the capabilities and limitations of our approach.

4.1 Zero-Shoot Hate speech hypothesis

In this hypothesis test, we employed an NLI model to assess the effectiveness of various hypotheses in measuring hate speech and offensive language. Table 2 presents the results, including the average performance.

The best hypothesis with an average accuracy result of 0.669 was *"this sentence contains hateful entries or offensive language"*. This result suggests that explicitly addressing hateful entries and offensive language in the hypothesis effectively captures a broader range of harmful content. The model excels here because it doesn't need to make a strict distinction between the two types of content, which can be challenging given their nuanced overlap. We conducted this brief experiment using the testing split of "Learning from the Worst" and "HateCheck" datasets.

Table 2. Hypothesis Test (Best performing results are highlighted)

	Learning	Hatecheck	Average
this sentence contains hateful entries or offensive language	0.597	0.741	0.669
that sentence contains offensive language	0.601	0.728	0.665
that sentence contains hateful entries or offensive language	0.603	0.727	0.665
this sentence contains offensive language	0.600	0.728	0.664
that sentence contains hateful entries	0.580	0.745	0.663
that contains hate speech	0.590	0.728	0.659
this contains hate speech	0.590	0.727	0.658
this sentence contains hateful entries	0.568	0.737	0.653
hateful	0.566	0.728	0.647
offensive language	0.569	0.714	0.641
hate speech	0.564	0.707	0.635
hate	0.524	0.678	0.601
Average	0.579	0.724	0.652

The worst hypothesis with an average accuracy result of 0.601 was *"Hate"*. This result could be attributed to the extreme brevity and lack of context in the hypothesis. It does not provide sufficient information for the model to accurately classify hate speech or offensive language. Furthermore, the term "hate" on its own is highly ambiguous, making it challenging for the model to make accurate predictions.

These results show that the performance of the NLI model in ZSHSD is closely tied to the clarity and specificity of the hypothesis. Hypotheses that encompass a broader spectrum of harmful content and provide context tend to yield better results. Conversely, overly brief or ambiguous hypotheses may hinder the model's ability to make accurate classifications, as evidenced by the lower scores observed in this test. Therefore, it is imperative to carefully craft hypotheses that encompass the diversity and context of hate speech to enhance the effectiveness of NLI-based ZSHSD models.

4.2 Hypothesis Similarity Analysis

In Table 3 below, we present the original token representations alongside their optimized counterparts. We will examine the changes in each token and discuss why the optimized version might be more or less effective for predicting hate speech. Furthermore, we will highlight the tokens that are likely to be more useful for this task. It's important to note that these transformations occur within a continuous embedding space, so the similarity of the optimized tokens to specific words might not fully capture the subtleties of their representation. Additionally, even if tokens seem less interpretable from a human perspective, they can still impact the model's internal weights in ways that may positively or negatively affect its performance on the task.

this - The optimized version adds unique characters and symbols, like "[...]" and "🙂", which might enhance the model's ability to detect nuanced hate speech or sarcasm. However, it makes direct interpretation harder.

sentence - The optimized version includes unrelated words like "edits", "inscription" and "subtitles" which don't seem to help with ZSHSD but rather shift towards unrelated terms.

contains - The optimized version contains random characters and sequences ("pha", "nton", etc.) that don't align with the original token's meaning. This transformation doesn't relate to ZSHSD.

hateful - The optimized version keeps "hateful" and "despicable" while adding terms like "fascist" and "discriminatory", broadening its capacity to detect hate speech with more discriminatory and extremist terms.

entries - The optimized version introduces unrelated terms ("ansk", "owitz", etc.) that don't positively contribute to ZSHSD or relate to the original token's meaning.

or - The optimized version adds terms like "antagonist", "favorably" and "extremists" which aren't directly related to the original token but may help identify hate speech involving comparisons, contrasts, or extremist viewpoints.

offensive - The optimized version includes terms like "transsexual", "LGBT", and "ethnic" which are relevant for detecting hate speech involving sensitive topics and discriminatory language.

language - The optimized version introduces unrelated terms ("andra", "eto", etc.) that don't contribute positively to ZSHSD.

The evaluation of the optimized tokens and their contribution to ZSHSD can vary significantly depending on the specific NLI model and its objectives. The analysis provided here represents a human-driven evaluation of the tokens' relevance to ZSHSD. While some tokens like "hateful" and "offensive" appear to strengthen the model's ability to identify hate speech by introducing relevant terms, others seem unrelated to the task and may introduce noise.

Table 3. Original Token and Optimized/p-tuned Token Cosine Similarity

	Original		Optimized/p-tuned
		...	
contains	"contains":1.0 "contain":0.786 "containing":0.752 "contained":0.727		"pha":0.221 "nton":0.217 "pton":0.204 "beit":0.199
hateful	"hateful":1.0 "despicable":0.767 "hideous":0.761 "vile":0.755		"hateful":0.245 "fascist":0.237 "discriminatory":0.237 "despicable":0.233
		...	
or	"or":1.0 "and":0.724 "Or":0.693 "OR":0.64		"antagonist":0.195 "favorably":0.194 "parasite":0.184 "extremists":0.183
offensive	"offensive":1.0 "Offensive":0.725 "defensive":0.68 "offensive":0.654		"transsexual":0.233 "LGBT":0.233 "transgender":0.232 "ethnic":0.232
		...	

Table 4. Original and Virtual Tokens Combinations Performance

	Cumulative	Individual	Isolated	Average
v_{tokens}				
CH this	0.38	0.61	0.38	0.45
CH sentence	0.39	0.39	0.37	0.38
CH contains	0.63	0.61	0.51	0.59
CH hateful	0.70	0.45	0.46	0.54
CH entries	0.71	0.50	0.43	0.55
CH or	0.68	0.63	0.38	0.57
CH offensive	0.85	0.68	0.49	0.67
CH language	0.87	0.61	0.43	0.63
Average	0.65	0.56	0.43	

It's essential to note that the effectiveness of token optimization is context-dependent and should be evaluated within the broader framework of the NLI model's goals. Further experimentation and fine-tuning may be necessary to determine the most suitable token representations. Ultimately, the choice of optimized tokens should align with the specific nuances and objectives of the problem at hand, and the assessment should be guided by both human judgment and quantitative performance metrics.

4.3 Token Importance Assessment

The performance of three different methods for incorporating optimized tokens into the hypothesis for ZSHSD—"Cumulative," "Individual," and "Isolate"—is shown in Table 4. The accuracy scores over all test data reveal the contributions of specific tokens:

"this" and "sentence": Consistently contribute across all methods, with "Individual" showing a slightly higher impact.

"contains": Significantly boosts performance in the "Cumulative" approach, indicating its value in enhancing ZSHSD.

"hateful": Has a strong impact in the "Cumulative" method but decreases in "Individual" and "Isolate," suggesting its lesser standalone effect.

"entries" and "or": Vary in contribution, with "Cumulative" showing the highest values. "Or" contributes the least in the "Isolate" approach.

"offensive" and "language": Like "hateful," these tokens are crucial in the "Cumulative" approach but less impactful individually.

The "cumulative" approach shows the most significant performance improvement as tokens are sequentially added, underscoring the value of considering tokens cumulatively for enhanced ZSHSD. However, the impact of individual tokens varies, and not all tokens contribute equally.

Table 5. Optimized Virtual Tokens Inserted Into Standard Hypothesis Performance

	Accuracy
CH - contains hateful	0.578
H - pure hypothesis	0.665
CH - contains offensive	0.698
CH - hateful offensive	0.759
CH - hateful or offensive	0.828
CH - contains hateful or offensive	0.834
CH - contains hateful offensive	0.837
OH - full optimized	0.870

The "Cumulative" approach shows the most significant performance improvement, highlighting the importance of considering tokens collectively for better ZSHSD.

For further analysis, the tokens "contains," "hateful," "or," and "offensive" were selected due to their pivotal roles. The "Individual" method was used for additional tests:

"contains": Directly relates to hate speech identification.

"hateful" and "offensive": Explicitly linked to hate speech, enhancing the model's discriminatory power.

"or": Despite being a common conjunction, it introduces nuanced interpretations with terms like "antagonist", "favorably" and "extremists" demonstrating its adaptability in capturing varied hate speech nuances.

Table 5 shows the "full_optimized" hypothesis achieving the highest accuracy (0.870), indicating its effectiveness. Combinations including "hateful" and "offensive" also exhibit high accuracy, close to the "full_optimized" hypothesis, suggesting that a selected combination of tokens can achieve similar results.

The statistical analysis reveals a significant improvement of the "full optimized" sample over the "pure hypothesis". Considering the first value and second value as the significance to "full optimized" and "pure hypothesis" respectively, among the combinations analyzed, "contains hateful offensive" (0.321 vs. 0.001), "hateful or offensive" (0.216 vs. 0.003), and "contains hateful or offensive" (0.262 vs. 0.003) show significant improvements when compared to "H - Pure Hypothesis." The analyses were conducted with a 95% confidence interval, considering p-values below 0.05 as significant, thereby confirming the alternative hypothesis.

4.4 Previous work comparation

Table 6 displays accuracy results from various hate speech tests, highlighting the performance of different hypothesis configurations, from pure hypotheses to in-

Table 6. Datasets Detailed Results

Significance to full optimization (OH)	✗	✗	✗		✓	✓		
Virtual tokens inserted on standard hypothesis	CH contains hateful	**H**	CH contains offensive	...	CH contains hateful or offensive	CH contains hateful offensive	**OH**	Previous work (19) (16)
LearningFrom R2	0.519	0.578	0.669		0.805	0.796	0.834	0.779
			...					
LearningFrom R3	0.635	0.574	0.706		0.75	0.756	0.793	0.768
HateCheck derog dehum h	0.493	1.0	0.843		1.0	0.993	1.0	0.986
HateCheck derog impl h	0.229	0.914	0.486		0.964	0.936	0.921	0.85
			...					
HateCheck spell space add h	0.769	0.78	0.983		0.919	0.971	0.74	
HateCheck spell space del h	0.738	0.851	0.702		0.901	0.865	0.943	0.801
HateCheck TOTAL	0.613	0.741	0.725		0.884	0.895	0.929	0.873
Average	0.596	0.672	0.737		0.869	0.88	0.908	

termediate insertions to fully optimized hypotheses. Results exceeding previous work are marked in green. The individual insert method was used for intermediate steps, as depicted in Figure 1.

Key observations include an increased accuracy with the addition of optimized tokens to the pure hypothesis and a general trend of increasing accuracy from initial to fully optimized hypotheses, underscoring the effectiveness of the optimization procedure in enhancing the model's ZSHSD capabilities.

Some tests show that specific intermediate steps outperform others, indicating the need to tailor the optimization process to different hate speech categories. Intermediate representations such as "Hateful or Offensive", "Contains Hateful or Offensive", and "Contains Hateful Offensive" consistently yield high accuracy across various tests, reinforcing their importance in improving.

Performance enhancements were notable when tested on the Learning from the Worst test dataset (19), with accuracy gains of 5.5 percentage points (pp) and 1.2 pp on test splits R2 and R4, respectively. Unlike previous efforts that relied on fully fine-tuned classification models, our approach optimizes the input hypothesis of a NLI model. On the Hatecheck dataset (16), optimized tokens led to an average accuracy improvement of 7.74 pp, with a best-case scenario improvement of 18.8 pp using a fully tuned hypothesis (OH) compared to (H). Our fully tuned single hypothesis also outperformed previous work by 5.6 pp.

These findings highlight the effectiveness of our optimization procedure in enhancing the model's ZSHSD capabilities. Both fully tuned hypotheses and significant token insertions on standard hypotheses showed improved accuracy, suggesting future directions for broader hypothesis design within NLI models.

5 Conclusion

Our study underscores the significance of thoughtful token selection and the synergy between different aspects of hypothesis optimization. By improving how hypotheses are formulated, we can enhance NLI model performance in ZSHSD, contributing to the vital work of online content moderation and digital safety.

Our results highlighted the importance of specific tokens like "hateful" and "offensive" in enhancing model accuracy. By optimizing these tokens, we expanded the model's capability to detect a wider range of hateful language and ideologies. For instance, the optimized version of "hateful" introduced terms like "fascist" and "discriminatory" enriching the model's understanding of hatred and intolerance. Similarly, optimizing "offensive" incorporated terms like "transsexual", "LGBT" and "ethnic" which are relevant to hate speech detection. These additions improved the model's ability to recognize subtle nuances in discriminatory language.

Future work should focus on expanding and diversifying token pools, adapting models for multilingual contexts, and developing real-time adaptation mechanisms to keep up with evolving hate speech. Enhancing contextual analysis, incorporating user feedback, addressing ethical and bias issues, are crucial for improving accuracy and practical utility in online content moderation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Arango, A., Pérez, J., Poblete, B.: Hate speech detection is not as easy as you may think: A closer look at model validation. In: Proceedings of the 42nd international acm sigir conference on research and development in information retrieval. pp. 45–54 (2019)

[2] Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the international AAAI conference on web and social media. vol. 11, pp. 512–515 (2017). https://doi.org/10.1609/icwsm.v11i1.14955

[3] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT. vol. 1, p. 2 (2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423

[4] Fortuna, P., Dominguez, M., Wanner, L., Talat, Z.: Directions for NLP practices applied to online hate speech detection. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 11794–11805. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.809, https://aclanthology.org/2022.emnlp-main.809

[5] Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. **51**(4) (jul 2018). https://doi.org/10.1145/3232676, https://doi.org/10.1145/3232676

[6] Fortuna, P., Soler-Company, J., Wanner, L.: How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? Information Processing Management **58**(3), 102524 (2021). https://doi.org/https://doi.org/10.1016/j.ipm.2021.102524, https://www.sciencedirect.com/science/article/pii/S0306457321000339

[7] Goldzycher, J., Schneider, G.: Hypothesis engineering for zero-shot hate speech detection. In: Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022). pp. 75–90. Association for Computational Linguistics, Gyeongju, Republic of Korea (Oct 2022), https://aclanthology.org/2022.trac-1.10

[8] Gröndahl, T., Pajola, L., Juuti, M., Conti, M., Asokan, N.: All you need is" love" evading hate speech detection. In: Proceedings of the 11th ACM workshop on artificial intelligence and security. pp. 2–12 (2018)

[9] Gruzd, A., Soares, F.B., Mai, P.: Trust and safety on social media: Understanding the impact of anti-social behavior and misinformation on content moderation and platform governance. Social Media+ Society **9**(3), 20563051231196878 (2023)

[10] Liu, X., Ji, K., Fu, Y., Tam, W.L., Du, Z., Yang, Z., Tang, J.: P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks (2022)

[11] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)

[12] Luo, H., Glass, J.: Logic against bias: Textual entailment mitigates stereotypical sentence reasoning. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 1235–1246 (2023)

[13] MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: Challenges and solutions. PloS one **14**(8), e0221152 (2019)

[14] Nadeem, M., Bethke, A., Reddy, S.: StereoSet: Measuring stereotypical bias in pretrained language models. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 5356–5371. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.416, https://aclanthology.org/2021.acl-long.416

[15] Paz, M.A., Montero-Díaz, J., Moreno-Delgado, A.: Hate speech: A systematized review. SAGE Open **10**(4), 2158244020973022 (2020). https://doi.org/10.1177/2158244020973022, https://doi.org/10.1177/2158244020973022

[16] Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., Pierrehumbert, J.: Hatecheck: Functional tests for hate speech detection models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 41–58 (2021)

[17] Sevani, N., Soenandi, I.A., Wijaya, J., et al.: Detection of hate speech by employing support vector machine with word2vec model. In: 2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE). pp. 1–5. IEEE (2021)

[18] Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., Margetts, H.: Challenges and frontiers in abusive content detection. In: Proceedings of the third workshop on abusive language online. Association for Computational Linguistics (2019)

[19] Vidgen, B., Thrush, T., Waseem, Z., Kiela, D.: Learning from the worst: Dynamically generated datasets to improve online hate detection. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.132, https://aclanthology.org/2021.acl-long.132

[20] Wang, S., Fang, H., Khabsa, M., Mao, H., Ma, H.: Entailment as few-shot learner (2021)

[21] Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In: Proceedings of the NAACL Student Research Workshop. pp. 88–93. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/N16-2013, https://aclanthology.org/N16-2013

[22] Waseem, Z., Thorne, J., Bingel, J.: Bridging the gaps: Multi task learning for domain transfer of hate speech detection. Online harassment pp. 29–55 (2018)

[23] Wiegand, M., Ruppenhofer, J., Kleinbauer, T.: Detection of Abusive Language: the Problem of Biased Datasets. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 602–608. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1060, https://aclanthology.org/N19-1060

[24] Yin, W., Zubiaga, A.: Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Computer Science **7**, e598 (2021)

[25] Yin, W., Hay, J., Roth, D.: Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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). Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1404, https://aclanthology.org/D19-1404