APurifier: A Paradigm Shift in Hate Speech Normalization Using Unsupervised Language Models

Dong Yeon Kim, Hyeon Jae Kim, Seung Min Baek, Ho Jung Shin, Jae Wook Lee
Computer Software Department, Hanyang University

Abstract

In the online world, many individuals engage in hate speech while hiding behind the veil of anonymity. Consequently, extensive efforts are being made to eliminate such hate speech. However, most of these efforts are centered on simply cleansing the hate speech, inadvertently distorting the original intent of the statements. While hate speech is problematic, the inherent intent is significant, as freedom of speech must also be respected. Considering this importance, various studies have been conducted in the area of hate speech normalization. Particularly, recent research has leveraged the powerful capabilities of large language models (LLMs). These approaches have typically required clean-hate sentence pairs or ground truth labels to indicate the level of hate. Drawing inspiration from adversarial learning, this study introduces a new LLM-based methodology, **APurifier**, which does not require such pairs or labels.

1 Introduction

In today's digital era, people connect to the internet daily, posting their opinions on various communities, and reacting—sometimes with joy, other times with sorrow or anger—to the views of others. However, as society becomes increasingly polarized, there's a growing disconnect between interest groups, leading to misunderstandings and conflicts. Adding to this, recommendation algorithms, which are designed to show users content they like, exacerbate this divide by creating echo chambers or "filter bubbles." Consequently, these bubbles deepen the rift between groups, leading individuals to resort to hate speech. Numerous studies have been undertaken to address this issue. However, most focus solely on sanitizing hate speech, inadvertently distorting the original intent of the messages. This approach does not aid in bridging the gap between groups. To reconcile these divides, it is crucial to facilitate communication, understanding the existing problems, and how interests clash at certain points to foster productive discourse. Furthermore, freedom of expression must be respected. In any community where this freedom is guaranteed, it implies the coexistence of a variety of viewpoints and respect for diversity. If all statements deemed as hate are simply cleansed, the speaker's intent is lost, which is detrimental to communication. However, most existing studies require either pairs of hate speech and corresponding 'clean' sentences or pairs of hate speech with intensity labels. The datasets available for training Large Language Models (LLMs) are severely limited, and manually creating such datasets is not practical, especially in the case of the Korean language, where the scarcity of data is more pronounced. Therefore, this paper attempts hate speech normalization using an unsupervised approach without pair data. We faced two challenges:

1. How to supplement the insufficient dataset for training?
2. How to preserve the original intent of the speaker?

To solve these problems, we propose a new framework for hate speech normalization, named Adversarial learning-based hate speech purifier (**APurifier**). We introduce an estimator to supplement the dataset with lacking hate/clean labels. By calculating the probability of how hateful a sentence is, we use this value as the intensity label for adversarial learning. The resulting purified embeddings are used to generate sentences that reasonably preserve the speaker's intent. To our knowledge, this is the first attempt at such an unsupervised methodology. Our contributions are as follows:

- We investigate unsupervised hate speech normalization in the context of the scarcity of Korean hate speech data.

- We propose APurifier, which utilizes estimated probabilities of the hatefulness of a sentence to purify it while maintaining the original intent of the message.

2 Related Works

2.1 Fairness in Graph Neural Network

In the realm of Graph Neural Networks (GNNs), a significant challenge is the potential for these models to inherit and perpetuate societal biases present in the data. This issue is particularly pronounced in GNNs due to their unique graph structures and message-passing mechanisms. To address this, the FairGNN(Dai and Wang, 2020) model was introduced, which aims to eliminate bias in GNNs while maintaining high node classification accuracy. The FairGNN model leverages the inherent graph structures and limited sensitive attribute information available in the data. It uses this information to estimate sensitive attributes for nodes in the graph. The model then employs adversarial learning to ensure that the predictions made by the GNN are independent of these sensitive attributes. This is a crucial step in eliminating bias from the predictions. It provides a theoretical analysis demonstrating that FairGNN can achieve fairness under certain conditions. This is particularly noteworthy because these conditions hold even when only a limited number of nodes have known sensitive attributes. This makes FairGNN a practical solution for ensuring fairness in GNNs. The effectiveness of FairGNN in debiasing GNNs and maintaining prediction accuracy is demonstrated through extensive experiments on real-world datasets. These experiments show that FairGNN can successfully reduce bias without a significant trade-off in accuracy. This approach is particularly relevant for applications in sensitive domains where discrimination based on attributes like gender or race is a major concern.

2.2 Hate Speech Normalization

In the online world, there is a pervasive issue of hate speech. Simply banning or censoring hate speech may not be the most effective solution, as it does not address the root cause of the problem and may even exacerbate it by creating resentment among the users whose posts are censored. Among various research efforts, one notable approach is the method of hate speech normalization, which seeks to diminish the intensity of hatred in online posts while maintaining the essence of the underlying message. this strategy does not condone or support hate speech in any way, but rather provides a pathway towards a less hateful and more respectful discourse. It introduces a model called NACL(Masud et al., 2022), which operates in three stages:

1. Measuring hate intensity
2. Identifying hate spans
3. Paraphrasing these spans to reduce hate intensity

The first stage involves using a machine learning model to measure the intensity of hate in a given post. The second stage involves identifying the specific spans of text that contain hate speech. The final stage involves paraphrasing these hateful spans of text to reduce their hate intensity while preserving their underlying meaning. It demonstrates the effectiveness of their model through extensive experiments. They show that their model is capable of accurately predicting hate intensity, identifying hate spans, and normalizing text to reduce hate intensity. The results suggest that their model could be a valuable tool for social media platforms and other online communities seeking to mitigate hate speech and foster a more respectful discourse.

3 Methodology

3.1 Approach 1

Our methodology integrates a novel approach to hate speech normalization, focusing on maintaining the original intent of sentences and enabling purified sentence generation without paired data. We employ an Estimator to initially label a large corpus of unlabeled data, a crucial step to overcome the scarcity of labeled training samples. This estimated labeling is instrumental for training the Generator and Discriminator in subsequent phases. Our model adopts adversarial learning, inspired by the FairGNN paper, to address the challenges of unsupervised language generation. The architecture is specifically designed to facilitate a dynamic interaction between the Generator and Discriminator, using the estimated labels as a foundation for generating and refining purified content.

3.1.1 Estimator

The Estimator is trained on a smaller set of labeled data and is responsible for estimating labels for a larger, unlabeled dataset. The loss function for the

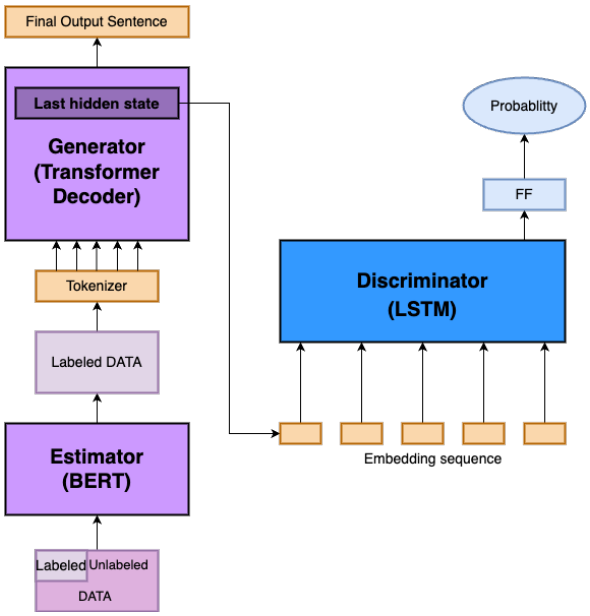

Figure 1: Overall Framework of Approach 1

Estimator, typically a form of cross-entropy loss is designed to maximize the accuracy of these predictions. These estimated labels enrich our original dataset for training the Generator and Discriminator.

3.1.2 Generator

Our Generator utilizes a pretrained transformer decoder, tasked with transforming hate and clean sentences into purified embedding sequences. It processes randomly paired sequences of estimated hate and clean sentences, producing embeddings that represent a neutralized version of the input. The Generator's loss function balances the adversarial goal of generating realistic sentences with the necessity of preserving the original sentence's meaning.

$$L_{content} = -\sum_{t=1}^{T}\sum_{k=1}^{K} y_{t,k} \cdot \log(p_{t,k}) \quad (1)$$

$$L_{adv} = -\sum_{i=1}^{N} y_i \cdot \log(p_i) + (1-y_i) \cdot \log(1-p_i) \quad (2)$$

$$L_{generator} = L_{content} - \alpha L_{adv} \quad (3)$$

3.1.3 Discriminator

The Discriminator, an integral part of our model, is constructed using an LSTM network. Its primary function is to analyze the embedding sequences generated by the Generator and classify them based on their origins, which are labeled as either hate or clean. However, to enhance the Discriminator's effectiveness, we initially pretrain it using a dataset that includes not just clean and hate speech but also offensive speech. This preliminary training phase is designed to familiarize the Discriminator with the broader spectrum of language, particularly the nuances that exist between clean and hate speech, effectively learning the characteristics of offensive content. After this pretraining phase, the Discriminator undergoes a focused adversarial learning phase using only the clean and hate data. This transition is crucial, as it shifts the Discriminator's focus from a broader understanding of language nuances to a more binary classification task. The loss function guides this training, emphasizing the Discriminator's ability to accurately distinguish between purified outputs derived from hate speech and those from clean speech. In this architecture, the Discriminator works in tandem with the Estimator and the Generator, forming a cohesive system. The Estimator provides the initial labeling, the Generator creates the neutralized content, and the Discriminator ensures the integrity and accuracy of the final purified output. This synergy enhances the model's overall ability to effectively purify hate speech while maintaining the semantic integrity of the original sentences.

$$L_{discriminator} = L_{adv} \quad (4)$$

3.2 Approach 2

Initially, our approach 1 encountered limitations inherent in unsupervised learning, failing to yield satisfactory results. This led us to contemplate a second approach. We began by exclusively training the decoder using hate-labeled data, employing teacher forcing as the training method. In this scenario, our decoder, having been trained solely on hate speech, tended to produce biased, hate-leaning outputs even when presented with clean sentences. By using the hate-biased outputs as a counterpart to the original clean sentences, we generate synthetic pair data. Subsequently, this artificially constructed paired dataset could be utilized to retrain the decoder, aiming to transform hate sentences into clean ones.

3.2.1 Decoder Training with Hate-Labeled Data

The initial training phase of the decoder focused on hate-labeled data. Teacher forcing, a technique often used in training recurrent neural networks, was applied. This method involves using the actual

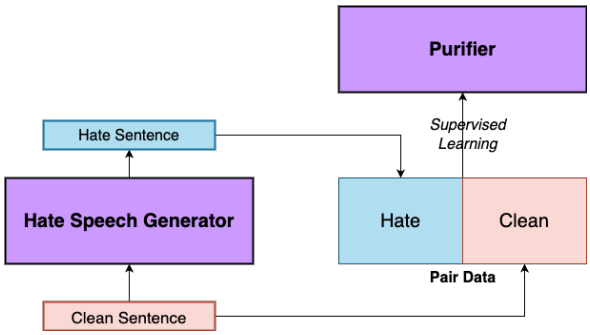

Figure 2: Overall Framework of Approach 2

output from the previous time step as an input for the current step, rather than using the predicted output. This strategy ensures that the decoder learns the structure and patterns of hate speech more effectively.

3.2.2 Synthetic Pair Data Generation

Leveraging the decoder's bias towards hate speech, we then input clean sentences, anticipating that the decoder would transform them into hate-leaning sentences. This process generates synthetic (clean, hate) pairs, which were then used as new training data. We expect this approach will cleverly circumvent the scarcity of paired samples in hate speech datasets.

3.2.3 Retraining the Decoder for Purification

With the newly created paired dataset, the next step involved retraining the new decoder to perform the opposite task: converting hate sentences into clean ones. This retraining aimed to teach the decoder to purify hate speech.

4 Experiments and Results

4.1 Dataset

We employ the Korean Hate Speech Dataset, a resource for analyzing toxic speech in Korean. The dataset, drawn from a Korean entertainment news platform, consists of three parts: labeled, unlabeled, and news title. The labeled section contains 9,381 annotated comments, categorized for social bias (gender, others, none) and hate speech (hate, offensive, none), offering insights into the interplay between bias and hate speech in online discourse. The extensive unlabeled corpus, with over 2 million comments, supports pretraining. Contextual understanding is enhanced by the news title section, providing titles of associated news articles.

4.2 Estimator Analysis

For effective adversarial learning with our LLM, a substantial amount of data was required. Our initial resources comprised a limited quantity of labeled data and a much larger pool of unlabeled data. To address this, we introduced an Estimator into our workflow. The Estimator was trained using the small set of labeled data, to label the extensive collection of unlabeled data.

Furthermore, for successful adversarial learning, it was crucial to train with distinct hate or clean data. Data that was ambiguously offensive—those near the borderline between hate and clean—posed potential complications for the learning process. To mitigate this, we trained a BERT-based Estimator using our labeled data, which included clean (label: 0), hate (label: 1), and offensive (label: 0.5) categories.

After training, the Estimator was used to label the unlabeled data. We then applied probability distribution-based filtering to remove data points that were close to the offensive category. This process ensured the exclusion of ambiguous data, leaving us with a more clearly defined dataset comprising either clean or hate speech.

This refined dataset, free from the ambiguities of offensive content, was then utilized for adversarial learning. By focusing on distinctly labeled data, we enhanced the quality and effectiveness of the training, setting a solid foundation for the adversarial learning process of our LLM.

4.3 Experiment with Llama2

Our initial experiments involved deploying Llama2, a pretrained transformer decoder language model. While Llama2 showed promise due to its advanced architecture, we encountered significant challenges due to its large model size. The training process was hindered by computational limitations until we managed to procure an NVIDIA A100 GPU, which facilitated the training. However, the training duration was excessively long, and we faced substantial delays. Further complications arose during inference, which took approximately 45 minutes per instance. This delay was a significant bottleneck in our experiment. Moreover, the quality of the generated outputs was unsatisfactory, as the model tended to repeat the same words.

4.4 Transition to Koalpaca

Given the impracticality of Llama2 for our requirements, we shifted our focus to Koalpaca, another pretrained transformer decoder language model. Koalpaca stood out due to its effective quantization and specialization in the Korean language. The model's learnable parameters were more manageable compared to Llama2. We fine-tuned Koalpaca with our dataset of hate speech comments. Although It was good at generating coherent responses, Koalpaca struggled with reproducing or closely mirroring the original sentences. The model's tendency to deviate from the source material indicated that our approaches were insufficient to achieve our objective of accurate and relevant language generation in the context of hate speech normalization.

5 Conclusion

In this paper, we embarked on a challenging journey to harness the power of the latest decoder models in the field of natural language processing, specifically for normalizing Korean hate speech. Our efforts were primarily conducted within the confines of Google Colab Pro+, a platform that we hoped would enable us to leverage these advanced models without the need for quantization.

Despite our best efforts, we encountered a series of obstacles that significantly hindered our progress. The primary challenge was the computational limitations inherent in the Google Colab Pro+ environment, which proved inadequate for handling the demands of state-of-the-art decoder models. This limitation was a significant barrier, as we aspired to utilize these sophisticated models in their full capacity to achieve the most accurate and effective results.

Our journey was marked by extensive exploration and experimentation with various methodologies and approaches. We invested considerable time and effort in seeking alternative solutions, rigorously testing different configurations and setups. However, each path we pursued led to a dead end.

The constraints of time were another critical factor that played into our challenges. The clock was not on our side, and as the time for our research drew to a close, we had to come to terms with the fact that our current environment, resources, and approaches were inadequate to fulfill our ambitious goals.

In reflection, our endeavor, though not culminating in the desired outcome, was far from fruitless. The journey was replete with learning experiences, deepening our understanding of the complexities involved in deploying advanced NLP models. It also highlighted the critical importance of computational resources in cutting-edge AI research.

Looking forward, we remain hopeful and determined. The challenges we faced have only bolstered our resolve to revisit this endeavor under more favorable circumstances. Our goal remains steadfast: to contribute meaningful advancements in the normalization of hate speech in the Korean language.

In closing, we extend our heartfelt gratitude to all those who have supported us in this journey. Our efforts, though met with setbacks, were driven by a deep commitment to our research goals and a passion for advancing the field of natural language processing. We look forward to the day when we can overcome these hurdles and realize the full potential of our research.

6 Acknowledgement

This document was prepared with the assistance of OpenAI's GPT language model, whose capabilities were instrumental in composing and refining the content of our paper.

References

Enyan Dai and Suhang Wang. 2020. Fairgnn: Eliminating the discrimination in graph neural networks with limited sensitive attribute information. *CoRR*, abs/2009.01454.

Sarah Masud, Manjot Bedi, Mohammad Aflah Khan, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Proactively reducing the hate intensity of online posts via hate speech normalization.