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 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.

Introduction

In the digital age, growing polarization and recommendation algorithms create online echo chambers and lead to hate speech. Existing solutions mainly focus on removing hate speech but overlook the original intent behind messages, hindering understanding between groups. To bridge divides and respect freedom of expression, a more balanced approach is needed. However, current studies rely on limited datasets with paired hate speech and clean sentences or intensity labels for training Large Language Models (LLMs). This paper introduces an unsupervised approach to hate speech normalization, overcoming data scarcity challenges.

Related Works

Fairness in Graph Neural Network is a model designed to mitigate bias in Graph Neural Networks (GNNs) while preserving accurate node classification. It utilizes graph structures and limited sensitive attribute data to estimate attributes, employing adversarial learning to ensure GNN predictions are independent of these attributes, promoting fairness. Empirical results demonstrate that FairGNN effectively reduces bias in GNNs without compromising prediction accuracy, making it valuable in sensitive domains where attribute-based discrimination is a concern.
Hate speech normalization addresses online hate speech by reducing its intensity while retaining the core message, offering an alternative to outright censorship. NACL, a model introduced in this context, consists of three stages: measuring hate intensity, identifying hate spans, and paraphrasing them to diminish hate while preserving meaning. Extensive experiments confirm NACL's effectiveness in predicting hate intensity, identifying hate speech, and normalizing text, making it a valuable tool for promoting respectful discourse on social media and online platforms.

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.

Methodology

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 over- come 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.

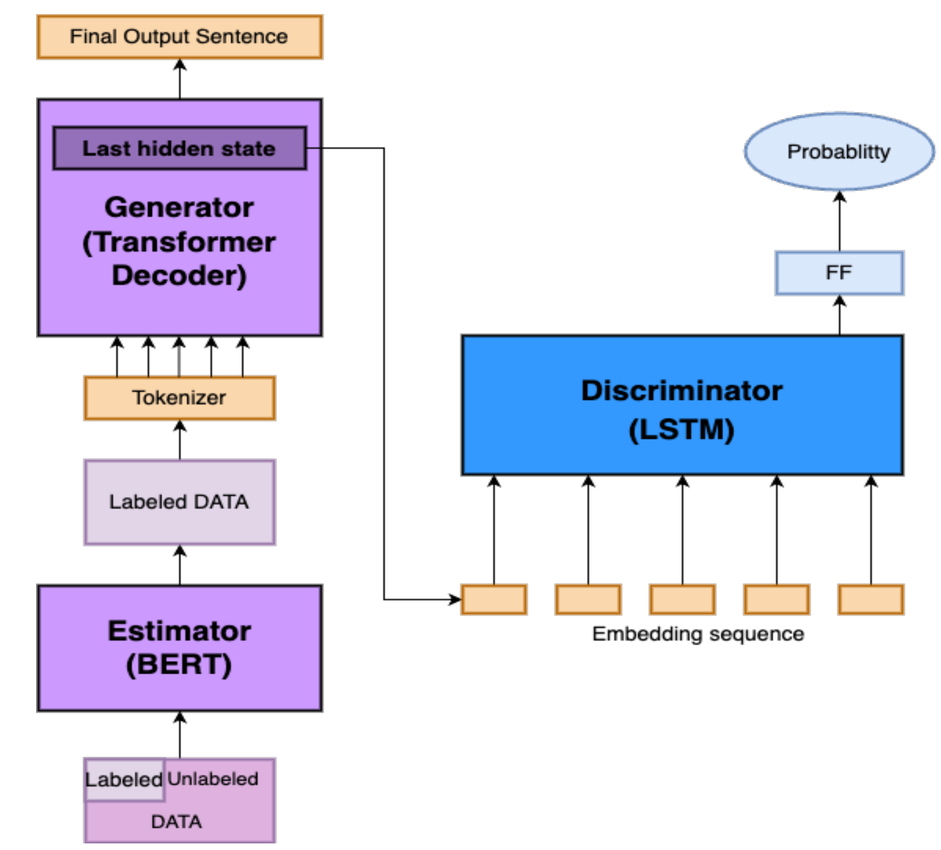
Approach 2:
Initially, our approach 1 encountered limitations inherent in unsupervised learning, failing to yield satisfactory results. This led us to contemplate a second approach. We began by exclusively training the decoder using hate-labeled data, employing teacher forcing as the training method. 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.

Conclusion

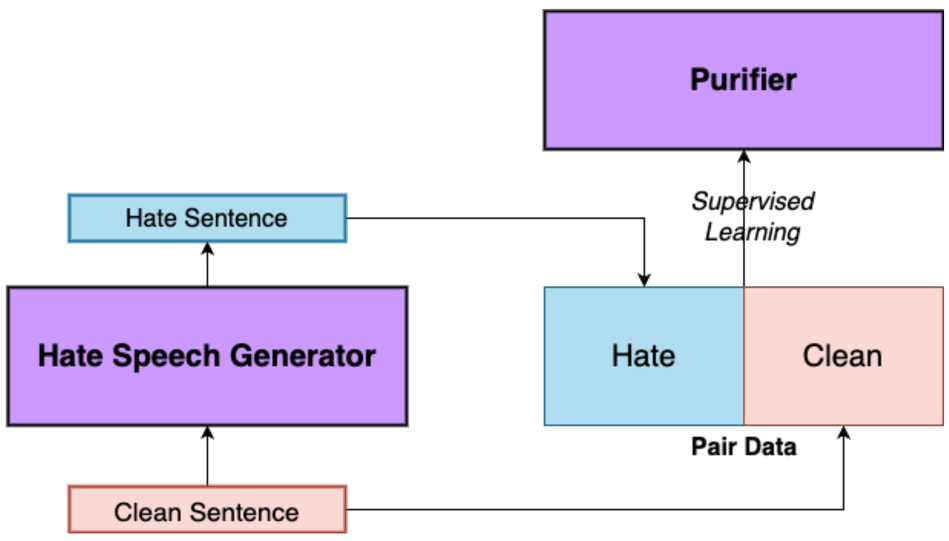
The paper chronicles our ambitious attempt to employ cutting-edge decoder models for the purpose of normalizing hate speech in the Korean language, conducted within the confines of Google Colab Pro+. The journey was marked by significant challenges, notably the computational limitations of this environment, which were insufficient to meet the resource demands of state-of-the-art decoder models. Despite extensive exploration and experimentation with various methodologies, configurations, and setups, the team faced numerous dead ends, exacerbated by the constraints of time. Consequently, we had to acknowledge that our current resources and approaches were not adequate to achieve their ambitious goals.

However, the research experience was far from unproductive. It deepened our understanding of the complexities involved in deploying advanced natural language processing models and underscored the critical importance of computational resources in the realm of cutting-edge AI research. With unwavering determination, we expressed our commitment to revisiting this endeavor under more favorable circumstances in the future, with the ultimate aim of contributing significantly to the normalization of hate speech in the Korean language. We extend heartfelt gratitude to our supporters and look forward to the day when we can overcome these obstacles and fully realize the potential of this research.

Model Diagram



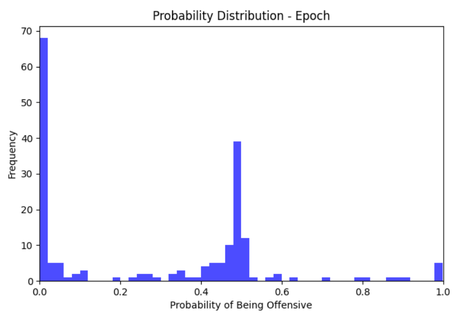
<Approach 1>



<Approach 2>

Experiments and Results

Training Estimator with Korean Hate Speech Dataset



To improve adversarial learning for our Large Language Model (LLM), the team used an Estimator trained on limited labeled data to label a larger pool of unlabeled data, ensuring a clear distinction between clean and hate speech categories by filtering out ambiguous content. This refined dataset was then utilized to enhance the quality and effectiveness of the adversarial learning process, establishing a strong foundation for our LLM.

Llama2 Generation
Initial experiments with Llama2, a large pretrained transformer language model, faced challenges due to its size, requiring an NVIDIA A100 GPU for training, but still resulting in long training times, slow inference (45 minutes per instance), and poor output quality with word repetition issues.

Koalpaca Generation
Due to impracticality, the focus shifted to Koalpaca, a specialized Korean language pretrained transformer model with manageable parameters; however, it struggled to reproduce original sentences accurately, hindering the objective of hate speech normalization.