Hate Speech Classification Using ALBERT

Cinthya Rosales, Neal Vazquez
DATASCI 266

Abstract

The objective of our project is to develop a machine learning model capable of accurately identifying hate speech in online communications. Utilizing the 'Social Media Hate Comments' dataset from Kaggle, this study employs the ALBERT (A Lite BERT) algorithm for its efficiency in processing natural language.

Inspired by Zeng et. al 2021, our approach compares ALBERT's performance with other widely accepted text classification models. In our case, our ALBERT model was compared to the performance of linear regression, random forest, and CNN-LSTM binary text classification models. Through this, we aim to understand ALBERT's effectiveness and present a comprehensive view of hate speech classification techniques. Our implementation includes preprocessing steps and leveraging the 'ALBERT-base-v2' pre-trained model, chosen for its advanced attention mechanism and parameter efficiency.

Key findings indicate that ALBERT outperforms traditional models, achieving an accuracy of 80.01% in hate speech detection, compared to lower accuracies of 53.06% with random forest, 54.05% with linear regression, and 67.66% with CNN-LSTM. This underscores ALBERT's capability in handling complex language patterns.

The study contributes significantly to natural language processing and the management of online hate speech, suggesting ALBERT's potential for large-scale, ethical content moderation and informing future research and digital policy-making strategies.

Introduction

The rapid proliferation of online platforms has fundamentally transformed the way individuals communicate and express opinions. While these platforms offer unparalleled opportunities for free expression and information exchange, they also present significant challenges, notably the rise of harassment and hate speech. A recent ADL survey found that 55% of adults have experienced some form of harassment online at some point in their lives, with 33% experiencing some form within the past 12 months. The implications of such content are far-reaching, impacting not just individuals but the very fabric of digital discourse. Our project, "Hate Speech

Classification using ALBERT", addresses this pressing issue by developing a sophisticated model capable of discerning hate speech within the diverse landscape of online communication.

The ubiquity of online hate speech and its potential to incite real-world harm underscore the necessity for effective detection and moderation tools. Traditional approaches to content moderation, primarily manual, are increasingly inadequate due to the sheer volume and complexity of online content. This challenge is exacerbated by the nuanced nature of language and the contextual subtleties that differentiate offensive content from benign expressions. Consequently, there is a growing reliance on automated systems, particularly those leveraging advancements in natural language processing (NLP) and machine learning, to address this issue efficiently and at scale.

Our research contributes to this evolving domain by employing the ALBERT (A Lite BERT) model, a state-of-the-art machine learning framework renowned for its efficiency in processing natural language. ALBERT's architecture, an optimized variant of the widely-used BERT model, offers a compelling balance of computational efficiency and sophisticated language understanding capabilities. This study focuses on harnessing these attributes to accurately classify text as hate speech or non-hate speech, navigating the intricacies of linguistic patterns and contextual relevance.

Furthermore, our approach aligns with and extends beyond existing research in this field. We build upon studies such as Zeng et. al 2021 and Wullach et. al 2021, which have utilized ALBERT in similar contexts, to not only validate the model's efficacy but also to explore its adaptability and performance in comparison to other models. By conducting a comprehensive analysis, including comparisons with traditional machine learning approaches like linear regression, random forest, and CNN/LSTM architectures, we aim to present a holistic view of the capabilities and limitations of various models in hate speech detection.

In summary, the "Hate Speech Classification using ALBERT" project endeavors to make a meaningful contribution to the field of NLP and the broader societal challenge of managing online hate speech. Through this work, we seek to offer insights into the practical application of advanced machine learning techniques in creating safer and more respectful online communities.

Background and Literature Review

The escalation of hate speech on social media platforms necessitates advanced automated detection methods. Traditional manual identification is impractical due to the sheer volume and subtlety of online content. The complexity of hate speech, often veiled in irony or non-direct language, poses additional challenges for detection algorithms.

Advancements in deep learning, particularly in Natural Language Processing (NLP), have significantly improved hate speech detection. Convolutional Neural Networks (CNN), Recurrent

Neural Networks (RNN), and Attention Mechanisms have been pivotal in these advancements. The introduction of transformer models like BERT and its variants, including ALBERT, marked a turning point in text classification tasks due to their deep contextualized representations and efficiency in parameter usage.

In "ALBERT for Hate Speech and Offensive Content Identification," Zeng et. al (2021) utilized ALBERT-large for the HASOC 2021 task, focusing on hate speech and offensive language recognition in English and Hindi. Their study highlighted ALBERT's reduced model parameters and efficient training capabilities, achieving a Macro F1 score of 83.75%. The dataset, derived from Twitter and provided by HASOC 2021, included 3790 English training instances, enhancing the model's ability to classify complex and nuanced content.

Another significant contribution is the "Fight Fire with Fire" study (Wullach et. al 2021), which explored augmenting datasets with synthetic hate speech generated via GPT models. This approach aimed to address the scarcity of labeled datasets and improve generalization in hate speech detection. By fine-tuning BERT, RoBERTa, and ALBERT models with these augmented datasets, the study demonstrated notable improvements in model performance and generalization, underscoring the potential of synthetic data enrichment in overcoming dataset sparsity and biases.

These studies form the foundation of our research, guiding our methodological approach. We build upon these findings by employing ALBERT-base-v2, chosen for its balance in efficiency and performance. Our study contributes to the evolving discourse in NLP, offering insights into the application of transformer-based models and data augmentation techniques in the nuanced task of hate speech detection.

Methodology

Our methodology encompasses several key components: dataset selection and preprocessing, model selection and fine-tuning, and comparative analysis with other models.

Dataset Selection and Preprocessing

We selected the 'Social Media Hate Comments' dataset from Kaggle for its diversity and volume, containing over 41,000 comments from various social media platforms, each categorized as either hate speech or non-hate speech. This dataset provides a broad representation of online discourse, crucial for training a robust model.

Preprocessing steps included:

- **Text Normalization**: Lowercasing all text and converting numbers to strings.

- **Cleaning**: Removing non-alphanumeric characters, including emojis and special characters like '@'.
- **Stemming:** Reducing word to their base or root often by trimming off inflections and derivational affixes.
- **Lemmatization:** Reducing words to their dictionary form, considering their part of speech to ensure accurate and meaningful reduction.
- **Tokenization**: Employing the ALBERT Fast tokenizer to convert text into tokens suitable for model input.
- **Model Selection and Fine-Tuning**: The core of our methodology is the utilization of the 'ALBERT-base-v2' model. This model was chosen for its balance of efficiency and depth of learning, critical for large-scale text classification tasks like ours.

Fine-tuning involved:

- **Parameter Adjustment**: Tailoring the model parameters to suit our specific task of hate speech detection.
- **Training and Validation Split**: Dividing the dataset into training (80%) and validation sets, ensuring a broad and representative learning process.
- **Loss Function and Optimizer**: Using Pytorch's Crossentropy as the loss function and AdamW optimizer for efficient training.

Models

In our research, we employed a diverse set of machine learning models to address the complex task of hate speech detection. Each model was chosen based on its unique strengths and suitability for specific aspects of the problem.

- **Linear Regression:** Linear regression was chosen for its simplicity and interpretability. It serves as a baseline model to establish a clear understanding of how input features influence predictions. By providing a straightforward linear relationship between features and the target variable, it helps in identifying the relative importance of different features in the context of hate speech classification. We built and trained our own linear regression model. We used the 80/20 split in data for training and testing this model.
- **Random Forest:** Random forest was selected for its capability to capture complex interactions between words and phrases in textual data. By aggregating the predictions of multiple decision trees, random forest can uncover intricate patterns and dependencies within the input data. We built and trained our own random forest model. We used the 80/20 split in data for training and testing this mode
- **CNN-LSTM:** Convolutional neural networks (CNN) excel at feature extraction from sequential data, such as text. They can identify important local patterns and representations within text. Long Short-Term Memory(LSTM) networks are proficient at understanding and modeling long-term dependencies in sequential data. Hate speech often involves identifying subtle contextual nuances and linguistic constructs that span

over multiple words or phrases. LSTM helps capture such dependencies effectively. We built and trained our own random CNN-LSTMl. We used the 80/20 split in data for training and testing this mode

- **ALBERT-base-v2:** ALBERT-base-v2 was chosen as a representative of transformer-based models due to its improved efficiency compared to the original BERT while maintaining competitive performance. This model offers a favorable trade-off between computational resources and accuracy. Being pretrained on diverse and extensive data, it possesses a broad understanding of language, making it well suited for hate speech classification. Its pretraining on a wide range of topics and context allows it to capture nuances and subtleties in language that may not be explicitly present in the training data for this task.

Comparative Analysis

To benchmark the performance of ALBERT-base-v2, we compared it against traditional models of Linear Regression, Random Forest, and CNN-LSTM architectures. The comparison was based on accuracy, precision, recall, and F1 scores. This analysis helps in understanding the relative strengths and weaknesses of each model in the context of hate speech detection.

Experimental Setup:
The experiments were conducted in a controlled environment, leveraging GPU acceleration for efficient model training and evaluation. The setup ensured that each model was given equal resources and conditions for a fair comparison.

Our methodology integrates a rigorous preprocessing regime, a strategic choice of a transformer-based model, and a comprehensive comparative analysis. This approach not only allows us to evaluate ALBERT-base-v2's performance in hate speech detection but also to contribute to the broader understanding of automated content moderation using advanced NLP techniques.

Results

Linear Regression: This model had a fairly low training time of 4 minutes and 49 seconds. Its performance in hate speech detection was modest. It achieved an accuracy of 54.05%, with precision of 0.4876 and recall of 0.1008. The F1 score was 0.167. The linear regression ,odel showed a relatively high number of false negatives (3320) compared to true positives (372) indicating a challenge in identifying actual instances of hate speech.

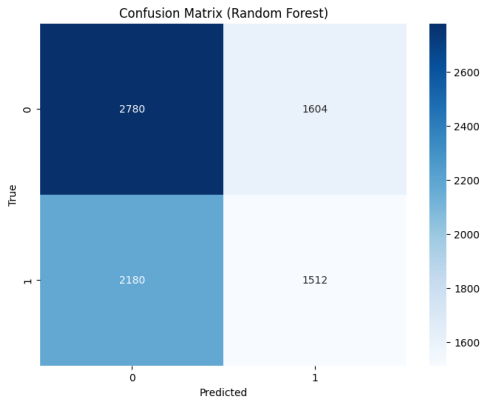
Random Forest: This model demonstrated a notably shorter training time of just 11 seconds, making it computationally efficient. In terms of performance, it achieved an accuracy of 53.22%, with a precision of 0.4861 and a recall of 0.4085. The F1 score was 0.4439. This model showed

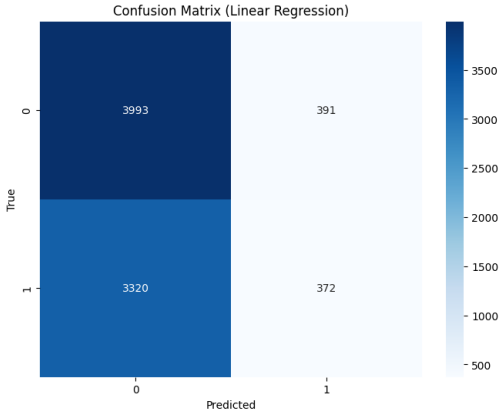
an improved ability to identify true positives (1512) compared to false negatives (1604). However, it also had a significant number of false positives (2180), which is a point of concern. The results of this model are still too poor to be considered as an appropriate tool for identifying hate speech.

CNN-LSTM: This model had a training time of 27 seconds. It demonstrated the highest accuracy out of the simpler models, achieving 68.16% accuracy. The model displayed balance precision (0.6543) and recall (0.6438) resulting in an F! Score of 0.6490. This model showed good performance in identifying both true positives (2423) and true negatives (3066) while maintaining a relatively low number of false negatives (1271) and false positives (1318). With more fine tuning, we believe that this model could achieve scores similar to our current ALBERT model. This is significant as its computational cost is very low compared to ALBERT.

ALBERT: This model had by far the largest computational cost and had a training time of 2 hours. It delivered the most impressive results in this instance, achieving an accuracy of 80.01%. It achieved a precision of 0.8032 and a recall of 0.7580. The F1 score for ALBERT was 0.7799. This model exhibited a robust ability to correctly classify both true positives (3340) and true negatives (3137) while keeping the false positives (1256) and false negatives (487).

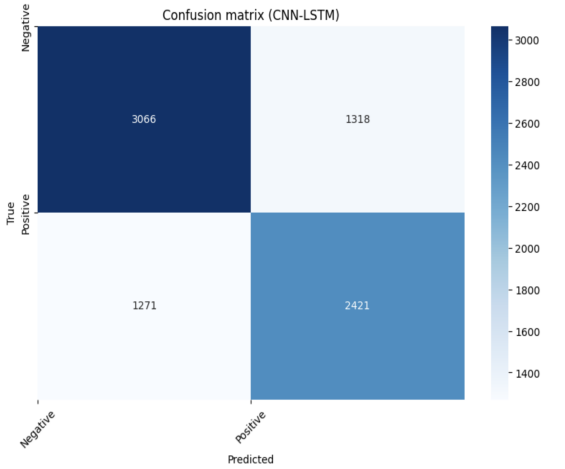
	Linear Regression	**Random Forest**	**CNN-LSTM**	**ALBERT**
Training Runtime	4:59	00:11	00:27	2:00:00
Accuracy (%)	54.049	53.219	68.164	80.01
Precision	.48755	.4861	.65428	.80316
Recall	.10076	.40845	.64382	.75804
F1 Score	.16701	.44392	.64901	.77994

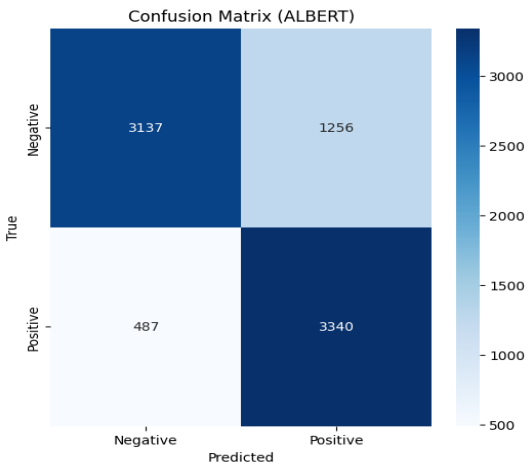




Random Forest

Linear Regression





CNN-LSTM

ALBERT

Ethical Concerns

Model Biases

One of the foremost ethical concerns when deploying pretrained models such as ALBERT for hate speech detection is the potential for model biases. ALBERT, like many other machine learning models, is trained on large datasets that inherently contain biases. These biases can be related to race, gender, religion or other sensitive attributes which hate speech normally targets. Given these biases the model might not be as effective and might even amplify the biases. In many countries where speech and news are limited, social media platforms offer a lifeline for disenfranchised communities. Unjustly banning people from platforms due to hate speech violations can have harmful consequences.

Labeling and Annotation Bias

The process of labeling data can introduce biases. Human annotators may have their own biases. These biases can inadvertently influence the data leading to bias in the models. Moreover, the definition of hate speech itself can be quite subjective. What one annotator considers hate speech may differ from another.

Conclusion

In this study, we undertook the crucial task of hate speech detection in online communications using advanced machine learning techniques. We leveraged the 'Social Media Hate Comments' dataset from Kaggle and implemented a diverse set of models, including Linear Regression, Random Forest, CNN-LSTM, and ALBERT-base-v2, to comprehensively evaluate their performance in this context. Our aim was to understand the effectiveness of ALBERT-base-v2, a transformer-based model, and compare it with traditional models in handling the nuanced and challenging task of hate speech classification.

Our findings indicate that ALBERT-base-v2 outperforms traditional models, achieving an impressive accuracy of 80.01% in hate speech detection. This surpasses the accuracy of Linear Regression (54.05%), Random Forest (53.22%), and even the CNN-LSTM model (68.16%). ALBERT's superior performance underscores its capacity to handle complex language patterns and effectively discern hate speech in diverse online communications.

However, while our results demonstrate the promise of advanced models like ALBERT in addressing the pressing issue of online hate speech, it is important to acknowledge the ethical concerns and limitations associated with such models. Model biases, stemming from biased training data, remain a significant challenge, and there is a need for continuous efforts to mitigate these biases and ensure fair and equitable outcomes. Additionally, the subjectivity of hate speech labeling and the potential for annotation bias highlight the importance of transparency and careful data curation in hate speech detection research.

Our work contributes significantly to the field of natural language processing and the management of online hate speech. It not only showcases ALBERT's potential for large-scale, ethical content moderation but also informs future research and digital policy-making strategies in the ongoing battle against online hate speech. As we move forward, it is imperative that we continue to refine and improve these models while addressing ethical concerns to create safer and more inclusive digital spaces for all.

Works Cited

1. Anti-Defamation League. (2023). "Online Hate and Harassment: The American Experience." [Online]. Available: https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2023

2. Davidson, T., Warmsley, D., Macy, M., and Weber, I. "Automated Hate Speech Detection and the Problem of Offensive Language." 2017. In Proceedings of the 11th International AAAI Conference on Web and Social Media.

3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." [Online]. Available: https://arxiv.org/abs/1810.04805

4. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." [Online]. Available: https://arxiv.org/abs/1909.11942

5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). "Attention is All You Need." [Online]. Available: https://arxiv.org/abs/1706.03762

6. Warner, W. and Hirschberg, J. "Detecting Hate Speech on the World Wide Web." 2012. In Proceedings of the Workshop on Language in Social Media (LSM 2012), pages 19–26, Montréal, Canada.

7. Wullach, T., Adler, A., & Minkov, E. (2021). "Fight Fire with Fire: Fine-tuning Hate Detectors using Large Samples of Generated Hate Speech." In Findings of ACL: EMNLP 2021. [Online]. Available: https://arxiv.org/abs/2109.00591 [cs.CL]

8. Young, T., Hazarika, D., Poria, S., and Cambria, E. "Recent Trends in Deep Learning Based Natural Language Processing." 2018. IEEE Computational Intelligence Magazine.

9. Zeng, Xu, and Wu. "ALBERT for Hate Speech and Offensive Content Identification." HASOC 2021. Available at: https://ceur-ws.org/Vol-3159/T1-26.pdf