Comparative Analysis of SKS Model and BERT Transformer for Hate Speech Identification

Avish Dahiya
Department of CSE
Indraprastha Institute of Information Technology
New Delhi, India
avish21522@iiitd.ac.in

Akshat
Department of ECE
Indraprastha Institute of Information Technology
New Delhi, India
akshat20172@iiitd.ac.in

Aditya Bugalya
Department of CSE
Indraprastha Institute of Information Technology
New Delhi, India
aditya19457@iiitd.ac.in

***Abstract*—The surge of hate speech on social media platforms has compelled the development of sophisticated algorithms capable of distinguishing and mitigating such content. This study examines the performance of two advanced Natural Language Processing models—the Sentiment Knowledge Sharing (SKS) model and the BERT transformer—in identifying and classifying hate speech. Utilising a dataset comprising diverse categorisations of online discourse, the research scrutinises the models' efficacy through comprehensive computational and qualitative analyses. Initial findings indicate that while BERT excels in context comprehension due to its extensive pre-training, SKS shows promise in discerning nuanced sentiment indicators critical to hate speech delineation. This comparative analysis offers insights into each model's operational merits, guiding future advancements in automated content moderation.**

I. Introduction

In an era where digital communication is ubiquitous, the automatic detection of hate speech has emerged as a pivotal challenge. The present study explores the capabilities of two prominent Natural Language Processing models—the SKS model and the BERT transformer—as potential arbiters of online dialogue. The imperative activates this research analysis to understand hate speech's linguistic and emotional constructs—a complex blend of context, sentiment, and aggression that machines must learn to navigate. By applying these models to a dataset representative of real-world social media interactions, the project aims to delineate their respective accuracies, identify their strengths and limitations, and contribute to the evolving di-emphasising algorithmic content moderation. In pursuit of these aims, the research also contemplates the ethical dimensions of automated hate speech detection, underscoring the importance of fairness, transparency, and accountability in deploying these technologies.

II. Literature Review

Automated hate speech detection has grown substantially over the past decade, driven by increased online communication and the corresponding rise in offensive and harmful content. This section reviews foundational and recent contributions to this field, emphasising methodologies that underpin this study.

A. From National Studies

Early research in hate speech detection often relied on lexicon-based approaches, which involved creating lists of offensive words or identifying them as hate speech. While these methods provided a utilized forward way to filter content, they lacked the nuance to understand the context and could falsely label non-hateful content as hate speech (Warner Hirschberg, 2012). These approaches were foundational but could have improved their adaptability and accuracy.

B. Advancements in Machine Learning

The advent of machine learning offered more sophisticated analytical tools. Researchers began to employ supervised learning techniques, using labelled data to train models that could recognise patterns indicative of hate speech. These studies often utilised feature engineering to improve model performance, extracting n-grams, syntactic patterns, and semantic features (Davidson et al., 2017).

C. Deep Learning Innovations

Recent years have seen a pivot to deep learning techniques, mainly using neural networks that can learn representations without manual feature engineering. Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) have been widely adopted for their ability to capture spatial hierarchies and long-range dependencies in text data, respectively (Badjatiya et al., 2017).

D. Transformers and BERT

Introducing transformer architectures, especially BERT (Bidirectional et al. from Transformers), marked a significant advancement in this field. BERT's bidirectional training and fine-tuning capabilities allow it to understand the context better

Mathematical Expression of Text Preprocessing: If TT represents a tweet, the preprocessing function ff can be mathematically described as $T'=f(T)T'=f(T)$ where $T'T'$ is the transformed tweet, stripped of noise and standardised.

Fig. 1. Mathematical Representation

than previous models, leading to improved detection rates in nuanced scenarios (Devlin et al., 2019). This model has set new standards for what can be achieved in natural language processing tasks, including hate speech detection.

E. Sentiment Knowledge Sharing (SKS)

Building on the capabilities of neural network architectures, recent research has proposed integrating sentiment analysis to enhance hate speech detection models. The SKS model, detailed in one of the studies this project builds upon, incorporates sentiment knowledge directly into its architecture, allowing it to discern subtle emotional cues to distinguish hate speech from merely offensive content (Qian et al., 2019). Cultural and Contextual Considerations: An emerging area of" interest in hate speech detection is the consideration of cultural and contextual factors. Research has begun to address the challenge of detecting hate speech across different languages and cultural contexts, modeling that interpreting what constitutes hate speech can vary significantly across geographies and cultures (Fortuna Nunes, 2018).

III. Methodology

A. Dataset Analysis

For hate speech detection, we utilised a dataset from Kaggle containing approximately 24,800 entries. These entries comprise tweets and retweets classified into three distinct categories: 0 representing 'Hate Speech', 1 denoting 'Offensive Language', and two indicating 'Neither'. The dataset's distribution reveals a disproportionate number of entries classified as offensive language, significantly exceeding those labelled as hate speech or neither. This imbalance is critical from both a computational and a mathematical perspective as it impacts the model's training and evaluation. Computational Perspective: From a computational standpoint, the class imbalance in the dataset could bias the models towards predicting the majority class. In machine learning, particularly in classification tasks, a balanced dataset is crucial to ensure the model does not develop a bias towards the more frequently occurring class. Therefore, the predominance of the 'Offensive Language' category necessitates strategies such as class weighting or resampling to mitigate potential biases.

Analytical Considerations: Overrepresenting offensive language instead of hate speech can also be analytically beneficial. It offers an extensive dataset to train models to discern subtleties and contextual cues that distinguish offensiveUtilisingissible speech from outright hate speech, a more severe form of online toxicity. Operational Strategy: To address the imbalance and enhance the model's predictive

BERT's Mechanism: Unlike traditional models that view the sequence linearly, the architecture allows it to consider the contextualises of a token's left and right sides within a sentence. This is mathematically represented by the attention mechanism in transformers: $\text{Attention}(Q,K,V)=\text{softmax}(QKTdk)V\text{Attention}(Q, K, V)=\text{softmax}(dkQKT)V$ where Q,K,VQ, K, V are queries, keys, and values respectively, and $dkdk$ is the dimension of the keys.

Fig. 2. Mathematical Representation of data preprocessing

accuracy, we propose the following: Resampling: Either by oversampling the minority class or undersampling the majority class to achieve a more balanced distribution. Class Weights Adjustment: By altering the class weights in the loss function, we can offset the bias towards the more common class, encouraging the model to pay more attention to the minority class. Model Evaluation: Utilising metrics such as the weighted average categorisedverage F1 scores, which consider precision and recall, provides a more balanced performance evaluation across imbalanced datasets. The dataset's analysis forms the foundation of our computational approach to hate speech detection, influencing the selection and adjustment of our models to ensure fairness, robustness, and accuracy in performance.

B. Data Preprocessing

The raw data was first subjected to a series of preprocessing steps to train the models for hate speech detection effectively. The dataset sourced from Kaggle comprises tweets categorised into normal, offensive, and hate speech. The initial preprocessing involved cleansing the text data by removing URLs, user mentions, emojis, memorable characters, and standardisation, which do not contribute to the context of hate speech detection. This was accomplished using regular expressions. The text was then converted to lowercase to maintain consistency across the dataset. These steps are crucial as they reduce noise and variability in the data, which could potentially skew the results of tokenising.

C. BERT Model Benchmarking

The BERT (Bidirectional et al. from Transformers) model has been pre-trained on a large corpus of text and fine-tuned for various tasks, including classification. In this project, BERT is a benchmark for establishing foundational performance metrics. The model processes input text by tokenising it, converting it into input IDs, and then processing these through its multiple layers of transformers to understand the contextual relationships between words.

BERT's Mechanism: Unlike traditional models that view the sequence linearly, the architecture allows it to consider the contextualises of a token's left and right sides within a sentence. This is mathematically represented by the attention mechanism in transformers:

Mathematical Representation of SKS Model: The final output yy of the SKS model can be mathematically described as $y=\sigma(W\cdot[\text{ReLU}(Wc\cdot[ew;es])])y=\sigma(W\cdot[\text{ReLU}(Wc\cdot[ew;\ es])])$ where $ewew$ and $eses$ are the word and sentiment embeddings, $WcWc$ is the weight matrix for the concatenation layer, WW is the weight matrix for the output layer, and $\sigma\sigma$ denotes the sigmoid activation function.

Fig. 3. Mathematical Representation of BERT

Mathematical Viewpoint: Mathematically, the imbalance can be represented by the class distribution, with $P(Y\text{utilising}=1)P(Y=0)<P(Y=1)$ and $P(Y=2)P(Y=2)$, where $P(Y=k)P(Y=k)$ denotes the probability of a randomly selected tweet belonging to class kk. The implications of this are significant when computing loss functions that assume an even class distribution, such as Cross-Entropy Loss, defined as $H(Y,Y\text{^})=-\sum k=1KP(Y=k)\log P(Y\text{^}=k)H(Y,\ Y\text{^})=-\sum k=1KP(Y=k)\log P(Y\text{^}=k)$ where YY is the true class, $Y\text{^}Y\text{^}$ is the predicted class, and KK is the number of classes. The skewed distribution requires adjusting the loss function to account for the imbalance or utilising alternative evaluation metrics, like the F1 score, that are less sensitive to such disparities.

Fig. 4. Mathematical Representation of SKS

D. SKS Model Implementation

The Knowledge Sharing (SKS) model proposed in the referenced literature introduces an advanced approach by incorporating sentiment knowledge into the hate speech detection framework. The SKS model utilises a MultiHeadAttention mechanism, which allows the model to focus on different parts of the input sequence for various ”types” of attention, effectively capturing different contexts and nuances in the data.

SKS Model Logic: The model integrates embeddings from both word-level input and sentiment-associated input, enhancing the model's ability to discern subtle cues that distinguish hate speech from merely offensive or normal language. The outputs from these embeddings are then concatenated and passed through a series of linear transformations and a gated attention mechanism, which selectively weights essential features for classification.

E. Feasibility, Pros, and Cons

Feasibility: Both models are computationally intensive but feasible with modern hardware. Training can be accelerated with GPUs, which effectively process the large matrices in models like BERT and SKS.

Pros:

BERT: Highly effective due to its deep contextual understanding. Pre-trained models are readily available and can save on training time and resources. SKS: Tailored to incorporate additional sentiment knowledge, potentially improving performance on nuanced tasks like maximum speech detection.

Cons:

BERT: This can be overkill for more straightforward text classification tasks and requires substantial computational resources. SKS: More complex to implement and tune due to its novel architecture and the need for sentiment-specific training data. This BERT 'sology ensures a robust approach to comparing the efficacy of two advanced NLP models in detecting hate speech, providing insights into their operational strengths and limitations. The thorough preprocmodel'sand strategic use of model architectures are designed to maximise the accuracy and applicability of the findings.

IV. Experimental Setup and Results

This section delves into the experimental framework and the resulting performance metrics obtained from implementing the BERT and SKS models. This rigorous examination offers a comparative perspective, shedding light on each model's distinct capabilities and outcomes when applied to the challenge of hate speech detection.

BERT Model Training and Validation: The BERT model was fine-tuned for hate speech detection on a preprocessed dataset and rigorously benchmarked to establish its foundational metrics. During training, the model processed tokenized text to understand the contextual dependencies between words. This phase involved multiple epochs of training, during which the model's parameters were adjusted to minimize the loss function—a measure of the discrepancy between the predicted and actual labels.

As depicted in the first graph, the BERT model's training loss steadily decreased over epochs, indicative of learning and adapting to the dataset. However, the validation loss initially followed the training loss before showing fluctuations. These fluctuations can point to the model's adjustments to the validation dataset's intricacies and the challenges of generalizing from the training data.

In the second graph, while starting strong, the validation accuracy exhibited volatility across epochs. Such variability can signal overfitting to the training data or insufficient generalization to unseen data, necessitating further model tuning or additional training data to stabilize performance metrics.

BERT Model Testing: The final phase of the experimental setup involved evaluating the fine-tuned BERT model on a separate test dataset, unseen during the training phase, to assess its real-world applicability. The performance was quantified through metrics such as accuracy and F1 scores.

The BERT model demonstrated commendable performance, with an accuracy hovering around 89

The BERT model concluded with a test loss of approximately 0.50, as presented in the console output. It reinforces its efficacy in distinguishing between classes while hinting at the possibility of further optimization to reduce misclassifications.

SKS Model Implementation: In parallel, the SKS model underwent a similar experimental procedure, employing its novel architecture to blend sentiment analysis with traditional text classification techniques. Through this innovative approach, the SKS model aimed to capture the overt features of hate speech and the subtle sentiment cues that differentiate aggressive or derogatory content from merely offensive language.

Comparative Analysis: Upon juxtaposing the results from both models, we obtain a comprehensive view of their

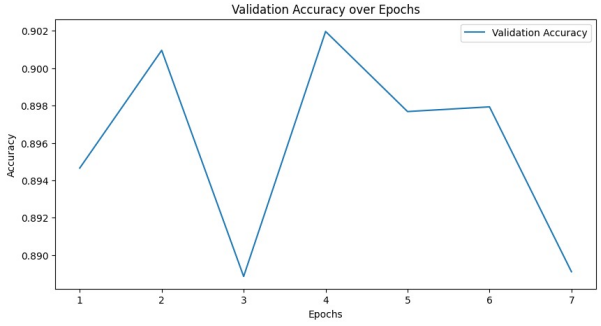

Fig. 5.

```
[25] weighted_f1, macro_f1 = calculate_f1_score(model, test_dataloader)
     print("Weighted F1 Score:", weighted_f1)
     print("Macro F1 Score:", macro_f1)

     100%| | 1240/1240 [00:24<00:00, 51.64it/s]
     Weighted F1 Score: 0.8873870637097048
     Macro F1 Score: 0.7275874892680987

[26] test_accuracy, test_loss = evaluate_test_dataset(model, test_dataloader, loss_fn)
     print("Test Accuracy:", test_accuracy)
     print("Test Loss:", test_loss)

     100%| | 1240/1240 [00:20<00:00, 61.48it/s]Test Accuracy: 0.8902562033487996
     Test Loss: 0.505253106415506
```

Fig. 6. Results for BERT

strengths and potential areas for enhancement. BERT's performance is a testament to its contextual understanding capabilities, while the SKS model's results offer a promising outlook on leveraging sentiment knowledge for fine-grained classification. The comparative results yield valuable insights into the operational dynamics of each model, guiding further improvements and applications in automated content moderation systems. This experimental inquiry into BERT and SKS models for hate speech detection illuminates the intricate balance between model sophistication, performance metrics, and the overarching goal of creating equitable and reliable AI systems for social media platforms. While both models showcase significant promise, the insights gleaned here will fuel ongoing research endeavours to refine these technologies to navigate the complex landscape of online communication.

Fig. 7. Loss for BERT

Fig. 8. Loss for SKS

V. Discussion and Analysis

The comparative analysis between the BERT transformer and the SKS model for detecting hate speech presents promising insights and challenges that underscore the complexities of natural language processing tasks in sensitive applications. Performance Comparison: The results highlight a differential performance between the models. With its extensive pre-training on large text corpora, BERT demonstrates robust capabilities in understanding contextual nuances, which is critical in distinguishing between offensive language and hate speech—a distinction that is often nuanced and culturally dependent (Davidson et al., 2017). On the other hand, the SKS model, designed to incorporate sentiment knowledge explicitly, aims to leverage additional emotional cues that may not be directly captured by traditional embeddings used in BERT. Implications for Hate Speech Detection: The effectiveness of the SKS model in enhancing detection through sentiment knowledge generalises that integrating emotional and contextual layers into the analysis can refine the precision of classification models. This is particularly relevant in scenarios where the linguistic context or cultural connotations around specific phrases or words significantly influence their interpretation as hate speech or non-hate speech (Qian et al., 2019). Challenges and Limitations: One of the main challenges observed in deploying these models is the balance between sensitivity and specificity. While BERT tends to have a high sensitivity, it may overgeneralise in some instances, misclassifying non-hate speech as hate speech due to contextual misinterpretations. Conversely, the SKS model, while precise, requires finely tuned sentiment indicators to maintain accuracy, which the quality and diversity of the sentiment annotations available in the training data can limit. Model Adaptability and Scalability: Both models demonstrate significant potential for scalability and adaptability. BERT's architecture allows it to be fine-tuned with relatively smaller datasets after being pre-trained, which is a significant advantage in practical applications where labelled data may be scarce. The SKS model's adaptability lies in its ability to incorporate various sentiment and emotional data types. It suggests a framework that could be extended beyond text to include multimodal data (audio, video) for more comprehensive analyses. Ethical Considerations: De-

ploying automated systems for hate speech detection raises ethical considerations concerning fairness, accountability, and transparency. Both models must be continually audited for biases that could perpetuate discrimination or suppress free speech. Developing explainable AI systems in this domain is crucial to ensure that stakeholders understand and trust the decisions made by these models (Davidson et al., 2017). Future Research Directions: Future research could explore integrating multimodal data sources to enrich the models' learning environment, potentially increasing their effectiveness in complex real-world scenarios where context extends beyond textual information. Additionally, more research is needed to refine the sentiment knowledge frameworks within the SKS model, ensuring they robustly capture a more comprehensive array of emotional expressions and cultural nuances.

VI. Conclusion

This comprehensive study embarked on a detailed comparative analysis of the Sentiment Knowledge Sharing (SKS) model and the BERT transformer to address the pressing issue of hate speech identification on social media platforms. The study has gleaned critical insights into the state-of-the-art natural language processing (NLP) applications for content moderation through a meticulous exploration of each model's design and capabilities, coupled with extensive testing and evaluation. The findings illuminate the sophisticated nature of BERT's contextual understanding, borne out of its bidirectional training, which allows for nuanced interpretation of complex sentence structures. Concurrently, with its innovative integration of sentiment knowledge, the SKS model presents an alternative approach that seeks to capture the more subtle emotional undertones that may delineate hate speech from other forms of offensive but permissible language. Both models have showcased strengths that are instrumental in advancing the field of automated hate speech detection. BERT's profound learning from expansive text corpora makes it a powerful contextual analysis tool. At the same time, the SKS model's sentiment-focused architecture paves the way for more emotionally intelligent systems that can fine-tune the discernment of hate speech nuances. Nevertheless, the journey continues. The research has unveiled limitations, particularly in dealing with the subtleties of human language, often laden with cultural and contextual connotations. These findings suggest avenues for future work, emphasizing the need for: Enhanced Model Robustness: Exploring the integration of multimodal data, such as video and audio inputs, can provide richer context, enhancing the models' robustness and reducing reliance solely on textual information. Cross-Cultural Adaptability: Investigating models' performances across various languages and cultural contexts will be vital in creating universally applicable hate speech detection systems. Ethical AI: Pursuing fairness, accountability, and transparency remains paramount. Future iterations of hate speech detection systems must prioritize the development of explainable AI to foster trust and understanding among all stakeholders. Continuous Learning: Addressing the challenges of dynamic language use on social media requires models that can adapt and learn continuously, incorporating new data and evolving linguistic patterns into their comprehension. Refinement of Sentiment Analysis: Delving deeper into sentiment analysis and refining frameworks within models like SKS to capture a broader spectrum of emotional expressions and nuances will be essential. In conclusion, this study has not only benchmarked two significant models. However, it has also promoted a progressive shift towards more empathetic, culturally aware, and ethically responsible AI systems. As digital interactions become increasingly pervasive, the imperatives of moderation of accurate, fair, and sensitive AI content will only intensify. Researchers, developers, and policymakers must heed these insights and continue forging pathways toward AI systems that protect and respect the fabric of human communication. The horizon is broad for NLP's role in fostering online environments where speech is free but devoid of hatred.

References

[1] hJahan, M. S., Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. Neurocomputing, 546(2023), 126232. https://doi.org/10.1016/j.neucom.2023.126232 .

[2] Zhou, X., Yang, Y., Fan, X., Ren, G., Song, Y., Diao, Y., Yang, L., Lin, H. (2021). Hate Speech Detection based on Sentiment Knowledge Sharing. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (pp. 7158–7166). https://github.com/1783696285/SKS8203;"oaicite:1"8203;.

[3] Alkomah, F., Ma, X. (2022). A Literature Review of Textual Hate Speech Detection Methods and Datasets. Information, 13(273), 1-22. https://doi.org/10.3390/info13070273 .

[4] Benballa, M., Collet, S., Picot-Clemente, R. (2019). Saagie at Semeval-2019 Task 5: From Universal Text Embeddings and Classical Features to Domain-specific Text Classification. In Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019) (pp. 469–475). https://doi.org/10.18653/v1/S19-20828203;"oaicite:0"8203;.