Named Entity Recognition in Afrikaans: A Transfer Learning Approach

Shakeel Malagas, Mohammad Zaid Moonsamy, Luca von Mayer
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
{2424161, 2433079, 2427051}@students.wits.ac.za

***Abstract*—Despite recent advances in Natural Language Processing, there remains a significant gap in language models for African languages, including Afrikaans. This paper presents BERTaal, a novel approach to Named Entity Recognition (NER) in Afrikaans using transfer learning from Dutch. Leveraging the linguistic similarities between Dutch and Afrikaans, we fine-tune the Dutch pre-trained model BERTje through a two-stage process: first with masked language modeling on an Afrikaans Wikipedia corpus, followed by task-specific fine-tuning on an Afrikaans NER dataset. Our results demonstrate that BERTaal achieves 94% accuracy and a weighted F1-score of 0.94, outperforming a baseline model fine-tuned solely on the NER task. Through detailed attention map analysis, we uncover and address key patterns in the model's behavior, including the impact of dataset structure on attention mechanisms. This work not only presents a practical solution for Afrikaans NER but also provides insights into effective transfer learning strategies for low-resource languages with linguistic similarities to better-resourced languages.**

***Index Terms*—Afrikaans, LLMs, fine-tuning, NER, AfriBERTa , Dutch, BERTaal, BERTje.**

I. Introduction

The rapid advancement of Natural Language Processing (NLP) has predominantly centered on languages with vast resources, such as English or other European languages, and despite Africa's rich linguistic diversity, there remains a significant scarcity of Large Language Models (LLMs) for these languages, particularly Afrikaans. This lack of representation not only limits access to technology for millions of speakers, including those in South Africa and Namibia where Afrikaans holds unique historical and cultural significance, but also hampers the development of technology that could address local needs such as education, healthcare, and social services [1], [2].

Significant research efforts have been made to address this scarcity, such as initiatives like AfriBERTa and AfriBERT. Designed to function effectively with limited training data, AfriBERTa [1] leverages transfer learning techniques to adapt pre-trained models for various African languages . This model is particularly relevant to our task, demonstrating that effective language models can be developed in resource-constrained environments and paving the way for broader applications in underrepresented languages. Additionally, AfriBERT [2], a monolingual model specifically built for Afrikaans, outperforms the multilingual mBERT model in tasks like part-of-speech tagging, named-entity recognition, and dependency parsing.

Despite these advancements, notable limitations persist regarding the relative scarcity of dedicated Afrikaans models and data, which impacts their performance and usability. Furthermore, Afrikaans exhibits unique linguistic characteristics that distinguish it from many other African languages; while it shares structural and syntactical similarities with European languages, it also incorporates numerous borrowed terms from various African languages. This blend of influences complicates the development of effective NLP models that can accurately understand and generate Afrikaans text, highlighting the necessity for targeted research and model development in this area [3].

To address these limitations, we propose a novel approach utilizing a Dutch BERT model called BERTje [4] model as a backbone. The historical and linguistic connections between Dutch and Afrikaans, rooted in colonial influences, make this an appropriate choice; both languages share significant structural similarities while reflecting unique lexical contributions from African languages. Our methodology involves a two-step fine-tuning process: first, we fine-tune the BERTje model on the entire Afrikaans Wikipedia dataset using masked language modeling (MLM), followed by a second fine-tuning specifically targeting Named Entity Recognition (NER) using a relatively small Afrikaans NER dataset.

This dual-fine-tuning approach is compared to a baseline model consisting of the BERTje fine-tuned only on the same Afrikaans NER dataset. Our results demonstrate improved accuracy on the Afrikaans NER task with our proposed model, indicating that the Dutch backbone enhances its applicability in NLP tasks.

The remainder of this paper is structured to provide a comprehensive understanding of our proposed model. In Section 2, we present the background and related work necessary to understand our approach, focusing on LLMs, the BERT architecture, and existing initiatives such as AfriBERTa. Section 3 outlines our methodology, detailing the fine-tuning processes employed for both masked language modeling and named entity recognition. Finally, in Section 4, we discuss our results, analyzing the implications of our findings and their contributions to the development of Afrikaans language models.

II. Background and Related Work

In this section, we cover the key concepts and research essential to our study. We begin with LLMs, focusing on BERT, which forms the backbone of many multilingual models, including AfriBERTa and AfriBERT. Understanding BERT's architecture is crucial for addressing the scarcity of Afrikaans LLMs and other low resource african languages in general. We also discuss the NER task, central to our work, and review related research, highlighting existing solutions and their limitations in supporting Afrikaans. This background sets the foundation for our approach and identifies the gaps our project aims to fill.

A. BERT

Large language models have dominated NLP since the introduction of the Transformer architecture [5]. Among these models, BERT (Bidirectional Encoder Representations from Transformers) [6] stands as a pivotal innovation, distinguished by its ability to capture deep bidirectional context by considering both left and right contexts of words across all model layers.

BERT's training methodology employs two key objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, the model learns to predict randomly masked tokens using their surrounding context, forcing bidirectional understanding rather than the traditional sequential processing. The NSP objective enhances the model's capability to understand relationships between sentence pairs, particularly beneficial for tasks like question-answering and sentence pair classification [6].

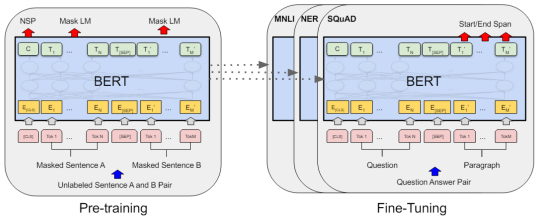

Fig. 1. BERT Architecture (adapted from [6])

BERT's architecture, displayed in Figure 1, employs a stack of transformer layers, each consisting of a multi-head self-attention mechanism and feed-forward neural networks. Each token in the input sequence is represented by a token embedding, a positional encoding (to capture the order of words in a sentence), and segment embeddings (which help differentiate between different input segments, such as sentence pairs in NSP tasks). The self-attention mechanism in each layer allows BERT to compute contextualized embeddings by attending to every token in the sequence in relation to every other token, regardless of their distance. The model uses the [CLS] token at the beginning of the sequence to represent the aggregate sentence-level embedding, and [SEP] tokens to denote sentence boundaries [6]. After pre-training with MLM and NSP objectives, the output embeddings from BERT can be fine-tuned for various downstream tasks by adding task-specific layers, as shown in the fine-tuning section of Figure 1.

Most of the new LLMs are based or inspired from the original BERT model [1], [2], [7]. For instance, mBERT (multilingual BERT) [6] was developed to handle multiple languages simultaneously, training on Wikipedia data across various languages. This multilingual capability has inspired models like AfriBERTa [1], though mBERT's limited coverage of African languages has been noted as a constraint for applications in the African linguistic context [7].

B. Named Entity Recognition

Named entity recognition identifies and classifies text entities into categories like person, organization, and location [8]. While early NER systems relied on manual feature engineering, modern deep learning approaches has achieved state-of-the-art performance through continuous vector representations [8]. NER serves as a common benchmark for evaluating language model performance [7], [1], [2], in applications like question answering and machine translation [7].

C. Related Work

Recent African language models include AfriBERTa [1], trained on 11 low-resource African languages, and InkubaLM [9], the first autoregressive African language model. However, these models, along with MMTAfrica [10], exclude Afrikaans. While [7] introduced comprehensive NER datasets and demonstrated successful mBERT fine-tuning for African languages, Afrikaans remained unstudied. Early Afrikaans AI work by [3] produced Afrikaana and AVRA using the Corpus of Spoken Afrikaans, though these systems had limited domain coverage and linguistic capabilities. While mBERT [6] included Afrikaans in its multilingual training, [2] developed AfriBERT, a monolingual Afrikaans model based on BERT's architecture. AfriBERT significantly outperformed mBERT in multiple tasks, including parts-of-speech tagging, dependency parsing and named-entity recognition [2].

III. Methodology

To address the scarcity of Afrikaans LLMs, we adopt a transfer learning approach from Dutch to Afrikaans. We implemented two models: a baseline and an improved model, BERTaal. For the baseline model, we utilised a pre-trained Dutch language model, BERTje, which was trained on a Dutch corpus, and fine-tuned it on our downstream Afrikaans task of NER. In the BERTaal, we introduced an additional step of Afrikaans-specific domain fine-tuning prior to the downstream task. This involved further fine-tuning the model on an Afrikaans Wikipedia corpus before fine-tuning it on the NER task. We hypothesise that BERTaal will perform better than the baseline model on the NER task.

A. Large Language Model

We leveraged BERTje, a Dutch pre-trained BERT model [4], to develop an Afrikaans language model through transfer

learning. This choice was motivated by BERT's proven effectiveness in semantic understanding and BERTje's strong performance in Dutch NER tasks [4], combined with the linguistic similarities between Dutch and Afrikaans.

BERTje was monolingually pre-trained on a vast corpus of Dutch data, consisting of 12GB (approximately 2.4 billion tokens). It features 110 million parameters and closely mirrors the architecture of the base BERT model [6], with 12 transformer blocks, a context window size, and a sequence length equivalent to BERT. Additionally, BERTje includes its own tokenizer, which uses Byte Pair Encoding (BPE) to preprocess the text for training and downstream tasks.

B. Datasets and Preprocessing

For this study, we employed two distinct datasets to fine-tune our models:

1) Wikipedia: We obtained a corpus of Afrikaans data from Wikipedia dumps, comprising diverse general information across various topics. This unstructured dataset, totaling approximately 254 MB of text, consisted of raw Afrikaans text without labels. We selected this dataset to fine-tune the BERTje model on Afrikaans data, aiming to enhance the model's understanding of the language. The preprocessing phase involved cleaning the Wikipedia dumps by removing extensive metadata, templates, links, and other non-textual elements, as well as eliminating redundant whitespace. The cleaned corpus was then tokenized using BERTje's BPE tokenizer for training.

2) NER Dataset: Given the limited availability of Afrikaans datasets, we focused on Named Entity Recognition (NER) for our downstream application, following research by [8] and [7] that establishes NER as a fundamental NLP task. We utilized the NCHLT (National Centre for Human Language Technology) Afrikaans NER dataset [11], which contains approximately 15,000 tokens across 8,962 sentences, primarily from the technology domain. Unlike the Wikipedia dumps, this dataset includes labeled entities for each word within the sentences, using the BIO (Beginning, Inside, Outside) tagging scheme. The entity labels are presented in Table I. For instance, the entity "Adobe Acrobat," "Adobe" is labeled as B-ORG (beginning of an organization entity) and "Acrobat" as I-ORG (inside an organization entity).

TABLE I
ENTITIES IN THE NCHLT NER DATASET

Label	Entity Name
ORG	Organization
LOC	Location
PERS	Person
MISC	Miscellaneous
OUT	Not an entity

A notable characteristic of this dataset is its class imbalance, with OUT tokens accounting for 85% of all entity tokens, while some labels like I-LOC are underrepresented. The dataset required minimal preprocessing compared to the Wikipedia corpus, mainly involving the removal of sentences lacking labels and redundant whitespace. We employed BERTje's tokenizer for processing the NER sentences, maintaining consistency with our Wikipedia preprocessing approach.

C. Fine-tuning and Evaluation of Models

For our baseline model, we conducted a full fine-tuning of BERTje on the NER dataset, utilizing a 70-15-15 train-validation-test split. Each input consisted of a single sentence, with the model producing corresponding NER labels. The fine-tuning process was conducted over 10 epochs with early stopping to mitigate the risk of overfitting, given the small size of the dataset. We set the maximum input sequence length to 128 to accommodate the shorter NER sentences, using a learning rate of 0.01 and batch size of 16. For the BERTaal model, we first fine-tuned BERTje using MLM on the Afrikaans Wikipedia corpus, with an 80-20 train-validation split. This process ran for 10 epochs using a batch size of 16, gradient accumulation over 8 steps, and a weight decay of 0.01, with a learning rate of 2e-5. The maximum sequence length was set to 512 to accommodate longer Wikipedia sentences. After fine-tuning the model using MLM, we applied the trained model to the NER task, adhering to the baseline approach.

To evaluate our models, we monitored the training and validation loss curves throughout the MLM fine-tuning process. For the downstream NER task, we assessed key evaluation metrics such as accuracy, F1-score, recall, and precision on the validation and test datasets. Additionally, we utilized attention maps to enhance the interpretability of the model's results. An inference function was implemented, taking a sentence as input and returning an attention heatmap, allowing for a more nuanced understanding of the model's predictions.

IV. Results and Discussion

In this section, we present a comprehensive comparison between our two models: the baseline model, which was solely fine-tuned on the NER dataset, and BERTaal, which underwent a two-stage fine-tuning process. We evaluate model performance through both quantitative metrics (classification reports including accuracy and F1-scores) and qualitative analysis of attention maps, which reveal the internal focus patterns of our models during inference.

A. Masked Language Modeling Performance

To validate BERTaal's first fine-tuning phase, we monitored the MLM training process, as illustrated in Figure 2. The training and validation losses exhibit clear convergence, with the decline in validation loss indicating that the model has effectively learned Afrikaans. Notably, the validation loss remains lower than the training loss throughout the entire training process, a phenomenon that warrants discussion. This behavior can be attributed to several factors:

- The relatively small batch size of 16 used during training, which can introduce higher variance in training updates
- The effect of dropout and other regularization techniques active during training but disabled during validation
- The possibility that validation set contains more frequently occurring patterns easier for the model to learn

Fig. 2. Training and validation loss curves during BERTaal's pre-training phase on the Afrikaans Wikipedia corpus, showing consistent convergence without overfitting

B. Named Entity Recognition Performance

As hypothesised, the classification reports (Figures 3 and 4) reveal that the baseline model was outperformed by BERTaal with accuracies of 0.92 and 0.94, respectively. Due to class imbalance from the OUT class, the macro f1-score average (0.74-0.77) differs substantially from the weighted average (0.93-0.94). BERTaal matched or exceeded the baseline's f1-scores across all entity categories, demonstrating that MLM fine-tuning enhanced downstream NER performance. While our model's f1-score of 0.94 appears to surpass the 0.85 reported by [2], their ambiguous averaging method makes direct comparison uncertain.

Final Test Set Performance:

	precision	recall	f1-score	support
b-loc	0.83	0.86	0.84	946
b-misc	0.65	0.65	0.65	2627
b-org	0.59	0.68	0.63	1645
b-pers	0.78	0.71	0.75	942
i-loc	0.52	0.69	0.60	149
i-misc	0.75	0.70	0.72	1893
i-org	0.64	0.73	0.68	1162
i-pers	0.80	0.85	0.83	832
out	0.97	0.97	0.97	51430
accuracy			0.92	61626
macro avg	0.73	0.76	0.74	61626
weighted avg	0.93	0.92	0.93	61626

Fig. 3. Classification report for the baseline model, showing performance metrics across all entity categories

Final Test Set Performance:

	precision	recall	f1-score	support
b-loc	0.83	0.89	0.86	946
b-misc	0.71	0.70	0.70	2627
b-org	0.74	0.63	0.68	1645
b-pers	0.80	0.76	0.78	942
i-loc	0.58	0.59	0.59	149
i-misc	0.76	0.75	0.75	1893
i-org	0.77	0.69	0.73	1162
i-pers	0.81	0.87	0.84	832
out	0.97	0.98	0.97	51430
accuracy			0.94	61626
macro avg	0.77	0.76	0.77	61626
weighted avg	0.94	0.94	0.94	61626

Fig. 4. Classification report for the BERTaal model, demonstrating improved performance across entities

C. Attention Map Analysis

1) Initial Observations and Baseline Behavior: Our initial attention map visualization for the baseline model revealed unexpected patterns in token attention distribution (Figure 5), (for details on the sentence used to investigate the attention maps, please refer to Table II in the supplementary material section). We observed that both the full stop token and the [SEP] token received disproportionate attention from all other tokens in the sequence. Particularly noteworthy was the comparatively minimal attention paid by the "Adobe Acrobat Reader" entity to the [SEP] token, visible as a distinctly light block in the attention map. This observation led us to hypothesize that the model might be overly reliant on capitalization for entity detection.

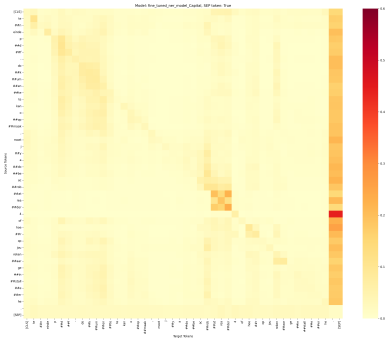

Fig. 5. Initial attention map for baseline model showing strong attention on full stop and [SEP] tokens

To test this hypothesis, we retrained the model using a lowercase version of the NER dataset while maintaining all other training parameters constant. The resulting attention map (Figure 6) showed markedly different patterns.

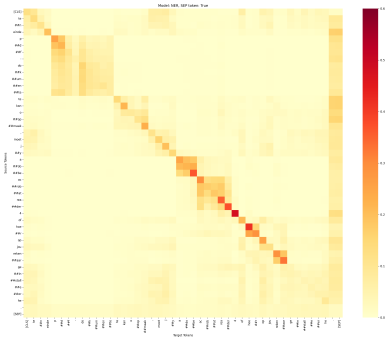

Fig. 6. Improved baseline model attention map after lowercase training, showing more distributed attention patterns

The lowercase-trained baseline model demonstrated more intuitive attention patterns, with clear attention blocks over multi-token words, indicating successful learning of word-level representations. More importantly, it showed enhanced attention to named entities, aligning with the model's NER training objective. This observation is notable, as the presence of block diagonals suggests highly salient attention heads—despite their scarcity—making their appearance, even subtly, in Fig 6 significant.

2) BERTaal Attention Patterns: The attention map for BERTaal (Figure 7) revealed some unexpected characteristics. While it successfully identified attention blocks for multi-token words, it exhibited unusually high attention on the [SEP] token from almost all other tokens in the sequence. Though the baseline model also showed some attention to

the [SEP] token, this attention perfectly matched the full stop token pattern and was significantly less pronounced than in BERTaal, which notably showed minimal attention to the full stop. BERTaal appears to diminish its focus on specific tokens, such as “Adobe Acrobat,” as indicated by the lighter block in Figure 7, suggesting a loss of previously acquired attention in certain heads. This result was unexpected, as we anticipated the opposite outcome.

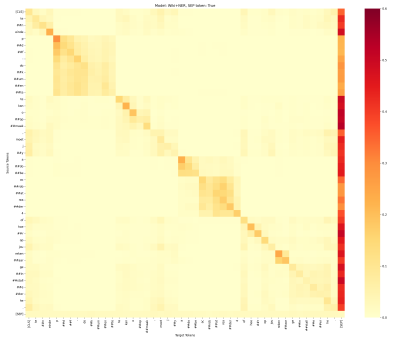

Fig. 7. BERTaal attention map showing strong [SEP] token attention and varying intensity patterns

D. Investigation of Attention Patterns

To address the challenges of mechanistic interpretability in these LLMs and to understand this unexpected attention behavior, we developed and tested three theories:

1) Theory 1: MLM Training Impact: We hypothesized that the unusual attention patterns might be a consequence of the MLM training process. Given BERT’s bidirectional nature, we speculated that the model might be using the [SEP] token as a structural anchor for understanding sentence structure as entities are often at similar places in a sentence such as the the subject or object positions. To test this theory, we analyzed attention patterns across multiple different sentences. The results showed consistently high attention on the [SEP] token regardless of entity presence (Figure 8), effectively disproving this hypothesis.

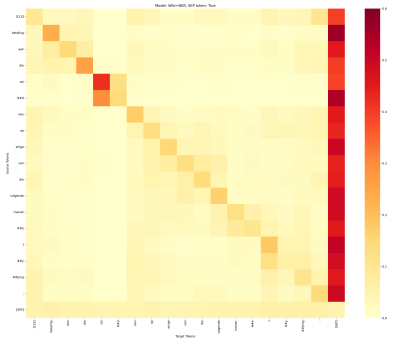

Fig. 8. Attention analysis showing consistent [SEP] token attention patterns on all entities (Sentence Two - Table III in supplementary material)

2) Theory 2: Word Frequency and Language Origin Effects: Our second hypothesis focused on the relationship between attention patterns and word characteristics. We postulated that less common words in the training data or words from different languages are encountered less frequently during training might show different attention patterns, with more attention focused on the words themselves rather than the [SEP] token. We tested this by systematically replacing words in our test sentences with rare and common Afrikaans and English words, examining both named entities and non-entity words. We selected words like “headphones,” “typing,” “oorfone,” and “spieel,” which were notably rare in the Wikipedia dataset—with “headphones” and “spieel” each appearing only once. The attention map results, shown in Figures 9 and 10, were inconclusive, offering both supporting and contradictory evidence for the English and Afrikaans words alike. For more detailed example results, please refer to the supplementary material section.

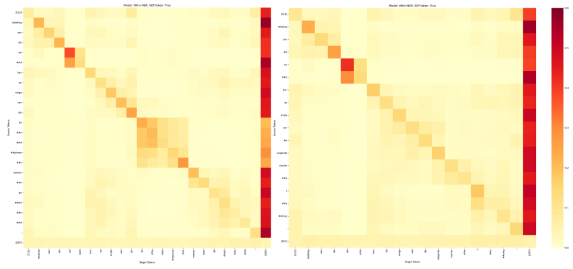

Fig. 9. Attention maps of English words: “headphones” (Left - supports) and “typing” (Right - contradicts) Theory 2

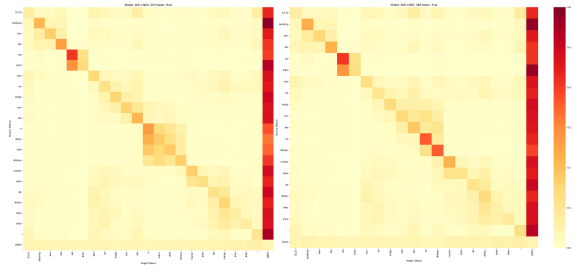

Fig. 10. Attention maps of Afrikaans words: “oorfone” (Left - supports) and “spieel” (Right - contradicts) Theory 2

3) Theory 3: Dataset Structure Impact: By analyzing attention maps at different training stages in Figure 11, we discovered that the extreme [SEP] token attention emerged after MLM training. Combined with our previous findings, this led to our final theory: the Wikipedia dataset’s sentence structure, particularly the distribution of [SEP] tokens or newline characters, might be causing the model to develop an over-reliance on these tokens for decision-making.

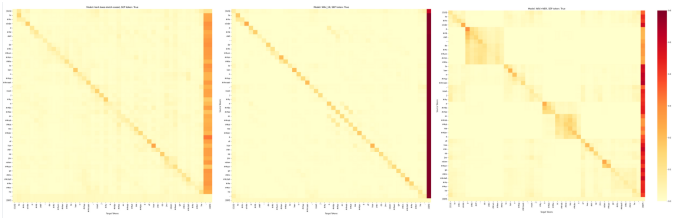

Fig. 11. Attention map progression across training stages: (Left) BERTje, (Center) BERTje fine-tuned on Wikipedia, and (Right) BERTaal

To test this hypothesis, we reformatted the Wikipedia dataset by placing all text on a single line and splitting it into 250-word segments. After retraining the model with this modified data, we observed significantly different attention patterns (Figure 12).

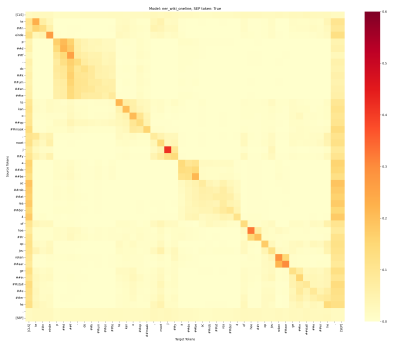

Fig. 12. Attention map for model trained on reformatted data showing more balanced attention distribution

The resulting attention map showed patterns more similar to the baseline model, with significantly reduced [SEP] token attention and matching patterns between [SEP] and full stop tokens. Interestingly, the [CLS] token attention also aligned with the [SEP] token patterns, though we observed that the clear attention blocks for longer words were less pronounced. While the exact mechanism behind this behavior remains unclear, the model achieved comparable performance with 94% accuracy and 93% weighted F1-score, with a notable improvement in I-LOC F1-score from 59% to 68% (Figure 13). For further additional investigations, please refer to the supplementary section.

Final Test Set Performance:

	precision	recall	f1-score	support
b-loc	0.79	0.93	0.86	946
b-misc	0.70	0.67	0.69	2627
b-org	0.71	0.64	0.67	1645
b-pers	0.81	0.75	0.78	942
i-loc	0.60	0.78	0.68	149
i-misc	0.75	0.76	0.75	1893
i-org	0.80	0.65	0.72	1162
i-pers	0.81	0.87	0.84	832
out	0.97	0.98	0.97	51430
accuracy			0.94	61626
macro avg	0.77	0.78	0.77	61626
weighted avg	0.93	0.94	0.93	61626

Fig. 13. Classification report for BERTaal trained on reformatted data showing improved I-LOC performance

The results reveal that BERTaal likely engaged in shortcut learning, primarily relying on [SEP] tokens as a decision-making aid. This behavior is likely influenced by two factors: an imbalanced distribution of OUT tokens in the NER dataset and irregular patterns of [SEP] tokens in the MLM dataset. We propose that the model could be treating the [SEP] token as a bias variable, which may lead to a tendency to default to the dominant OUT class in its predictions.

We suggest that the attention mechanism functions relative to this OUT bias, requiring other attention heads to compensate when identifying less frequent entity types, such as I-LOC. This may create a hierarchical decision boundary system, where each attention head attempts to identify specific entities by adjusting the baseline OUT bias. The model's inclination to use [SEP] tokens as bias variables might arise from their high frequency in the training data, representing an efficient, yet potentially problematic, strategy for minimizing loss rapidly (seen in Figure 2). This pattern may exemplify how the winner-take-all nature of softmax attention can lead to attention diffusion, whereby the dominance of the OUT class suppresses the model's ability to recognize other classes, given that all tokens compete for limited attention resources.

E. Limitations

While our research demonstrates promising results, several limitations should be acknowledged:

- **Attention Map Interpretability**: Despite extensive analysis, some patterns in our attention maps remain inconclusive, particularly regarding the model's handling of rare words and cross-language tokens
- **Computational Constraints**: Due to limited computational resources, we were unable to run the model for the optimal number of training epochs
- **Dataset Limitations**: The NER dataset's technology domain focus and class imbalance may affect the model's generalizability and performance on underrepresented entity types
- **Transfer Learning Boundaries**: While Dutch-to-Afrikaans transfer learning proved effective, the preservation of unique Afrikaans linguistic features requires further investigation

V. Conclusion and Future Work

Our research demonstrates the effectiveness of leveraging Dutch language resources for Afrikaans NLP through transfer learning. The BERTaal model, with its two-stage fine-tuning approach, achieved superior performance with 94% accuracy and a 0.94 weighted F1-score on the NER task. Through detailed attention map analysis, we uncovered how dataset structure and token distribution significantly influence model behavior, particularly regarding [SEP] tokens. The improved performance on the I-LOC class (from 59% to 68% F1-score) after dataset restructuring demonstrates that addressing these structural issues leads to more balanced model learning.

Future directions include training BERTaal for more than 10 epochs using MLM, implementing temperature scaling to address class imbalance issues (particularly with the OUT class), investigating the phenomenon of "grokking", which is training for significantly more epochs after convergence for potential breakthrough points in learning on the NER task. Overall, these findings contribute to NLP for low-resource languages, providing a framework for leveraging linguistic relationships between resource-rich and resource-scarce languages.

Finally, developing an Afrikaans LLM could enhance digital accessibility and education for Afrikaans speakers. However, given the language's historical role in colonialism and apartheid, such development must be weighed against the risk of perpetuating linguistic inequalities. To promote true linguistic justice in South Africa, any Afrikaans LLM initiative should be paired with comparable investments in indigenous language models for isiZulu, isiXhosa, Sesotho, and other historically marginalized languages.

REFERENCES

[1] K. Ogueji, Y. Zhu, and J. Lin, "Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages," in *Proc. 1st Workshop on Multilingual Representation Learning*, 2021.

[2] S. Ralethe, "Adaptation of deep bidirectional transformers for Afrikaans language," in *Proc. 12th Lang. Resources and Eval. Conf.*, 2020, pp. 2475-2478.

[3] B. Abu Shawar and E. Atwell, "Using the Corpus of Spoken Afrikaans to generate an ALICE chatbot," in *Southern African Linguistics and Applied Language Studies*, vol. 21, pp. 283-294, 2003.

[4] W. De Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim, "Bertje: A Dutch BERT model," *arXiv preprint arXiv:1912.09582*, 2019.

[5] A. Vaswani, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 1, Minneapolis, MN, USA, Jun. 2019, pp. 4171-4186.

[7] D. I. Adelani, J. Abbott, G. Neubig, D. D'souza, J. Kreutzer, C. Lignos, C. Palen-Michel, H. Buzaaba, S. Rijhwani, S. Ruder, and S. Mayhew, "MasakhaNER: Named entity recognition for African languages," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 1116-1131, 2021.

[8] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50-70, 2020.

[9] A. L. Tonja *et al.*, "InkubaLM: A small language model for low-resource African languages," *arXiv preprint arXiv:2408.17024*, 2024.

[10] C. C. Emezue and B. F. Dossou, "MMTAfrica: Multilingual machine translation for African languages," *arXiv preprint arXiv:2204.04306*, 2022.

[11] M. Puttkammer, M. Schlemmer, and R. Bekker, "Afrikaans NCHLT Annotated Text Corpora," South African Language Resource Management Agency, Potchefstroom, 1.0, ISLRN 139-586-400-050-9, 2014.

APPENDIX

NEURIPS PAPER CHECKLIST

1) **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: Yes

Justification: All the topics and results mentioned in the abstract were accurately covered in the paper, these claims are supported in the methodology and results section.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2) **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: Yes

Justification: The limitations are discussed in the limitations sections near the end of the paper.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3) **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: NA

Justification: No theoretical results or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper

should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4) **Experimental Result Reproducibility**
Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
Answer: Yes
Justification: The methodology is robust and covers all the steps needed to reproduce the results.
Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5) **Open access to data and code**
Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
Answer: Yes
Justification: The datasets and base Dutch model are openly available and our code will be submitted along with the paper.
Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is

permitted.

6) **Experimental Setting/Details**
Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
Answer: Yes
Justification: All relevant details are available in either the paper or can be seen by directly accessing the supplied code.
Guidelines:
- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7) **Experiment Statistical Significance**
Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
Answer: No
Justification: Error bars were not used due to compute limitations, but results were validated with classification reports.
Guidelines:
- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8) **Experiments Compute Resources**
Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
Answer: No
Justification: The experiments did not rely on any specialized hardware and thus was not explicitly stated.
Guidelines:
- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9) **Code Of Ethics**
Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?
Answer: NA
Justification: We have not reviewed the NeurIPS Code of Ethics.
Guidelines:
- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10) **Broader Impacts**
Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
Answer: Yes
Justification: Societal and broader impacts are briefly discussed in the introduction and conclusion.
Guidelines:
- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11) **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: NA

Justification: NER does not pose any risks as it does not generate any new data that could be harmful.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12) **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: Yes

Justification: Credit was given to the authors of base models and datasets used in our paper, everything used in our paper was cited and freely available.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13) **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: NA

Justification: We are not releasing any assets in this paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14) **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: NA

Justification: No crowdsourcing or research with human subjects was done.

Guidelines:

- The answer NA means that the paper does not

involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15) **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**
Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
Answer: NA
Justification: No crowdsourcing or research with human subjects was done.
Guidelines:
- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

CONTRIBUTION STATEMENT

Mohammad Zaid Moonsamy led the technical implementation of this project, contributing the much of the codebase, including the initial fine-tuning of the dutch-bert model using masked language modeling techniques. His work encompassed extensive preprocessing of both the Wikipedia dataset and the Afrikaans Named Entity Recognition dataset. Luca von Mayer provided the strategic vision for the Afrikaans LLM solution and along with Shakeel Malagas implemented the NER fine-tuning and the attention map inference function, while also collaborating with Zaid on preprocessing the Afrikaans Wikipedia dataset. All team members played active roles in sourcing and evaluating appropriate datasets and models for the project. The collaborative effort extended to conducting experiments, analyzing and interpreting results, and collectively authoring the research paper.

SUPPLEMENTARY MATERIAL

Tables

TABLE II
SENTENCE ONE

Entity	Tag
Ten	OUT
einde	OUT
PDF-dokumente	B-MISC
te	OUT
kan	OUT
oopmaak	OUT
,	OUT
moet	OUT
jy	OUT
Adobe	B-ORG
Acrobat	I-ORG
Reader	B-MISC
4	I-MISC
of	OUT
hoër	OUT
op	OUT
jou	OUT
rekenaar	OUT
geïnstalleer	OUT
hê	OUT
.	OUT

TABLE III
SENTENCE TWO

Entity	Tag
Betaling	OUT
aan	OUT
die	OUT
SAID	B-ORG
kan	OUT
op	OUT
enige	OUT
van	OUT
die	OUT
volgende	OUT
maniere	OUT
typing	OUT
:	OUT

TABLE IV
SENTENCE THREE

Entity	Tag
Wanneer	OUT
jy	OUT
Internet	B-MISC
Explorer	I-MISC
as	OUT
jou	OUT
blaaier	OUT
gebruik	OUT
:	OUT

Figures

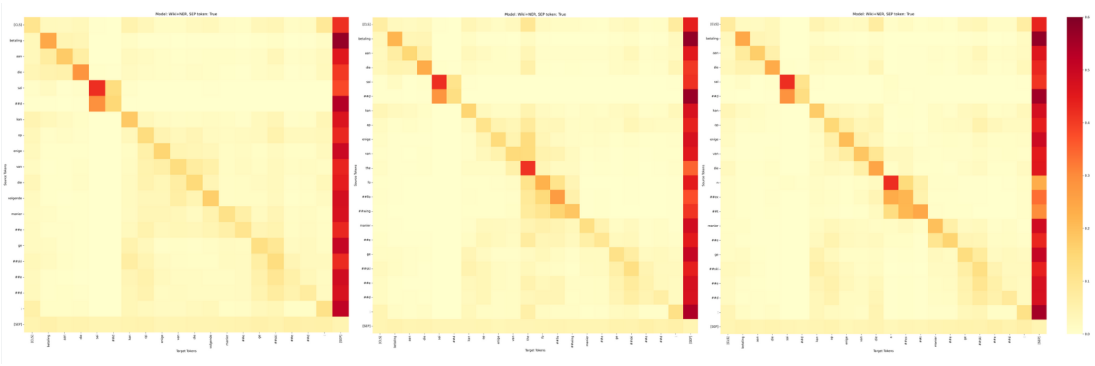

Fig. 14. Attention maps for Sentence 2 with different translation words: “Volgende” (Left), “Following” Center), and “Next” (Right)

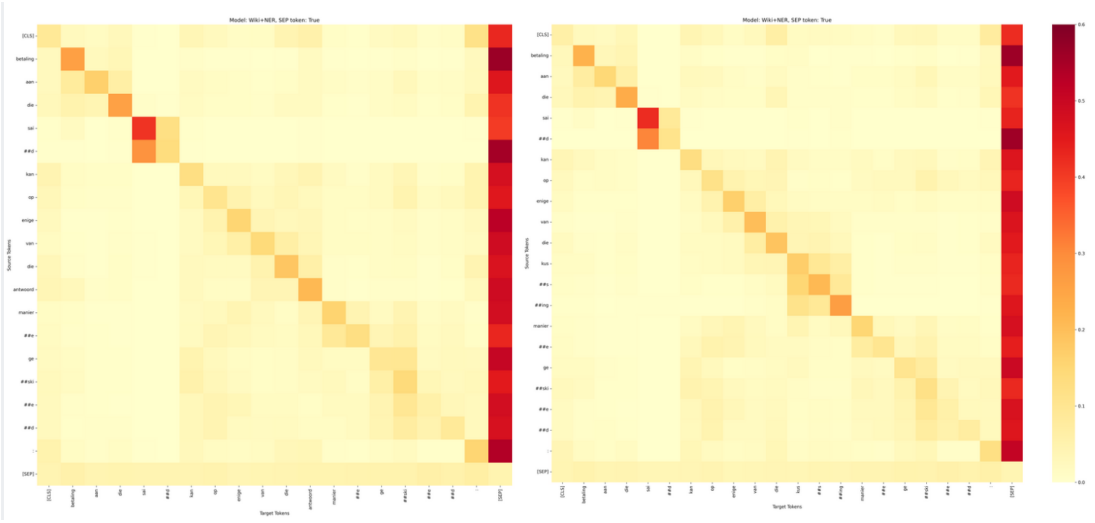

Fig. 15. Attention maps for Sentence 2 with different Afrikaans words: “Antwoord” (Left), and “Kussing” (Right)

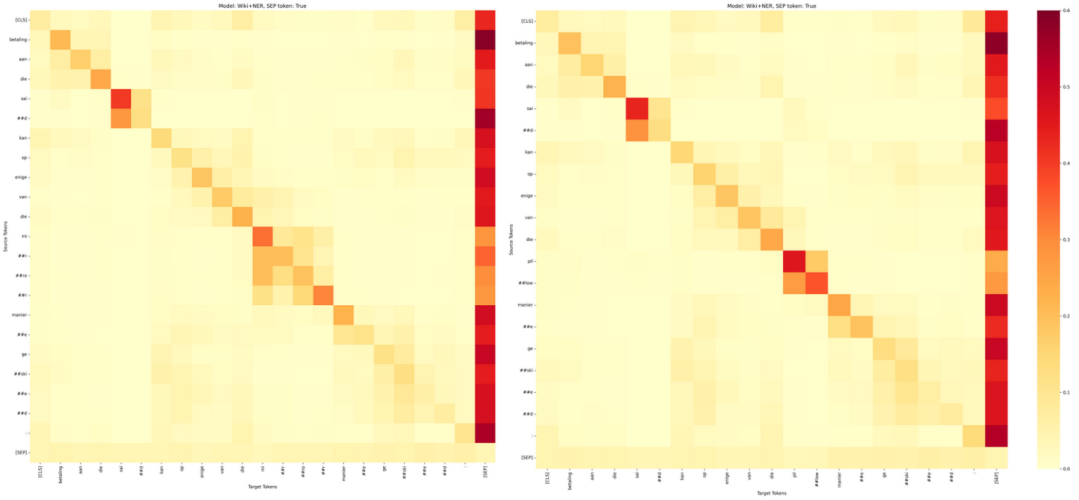

Fig. 16. Attention maps for Sentence 2 with different English words: “Mirror” (Left) and “Pillow” (Right)

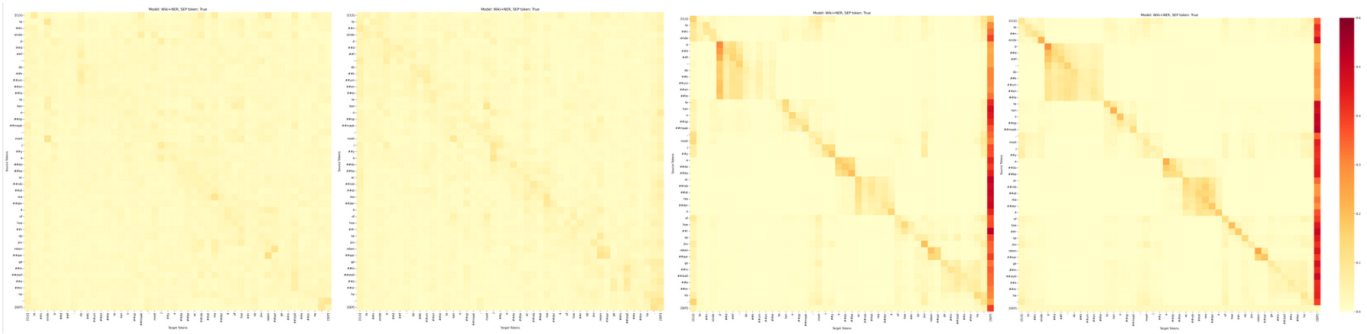

Fig. 17. Attention analysis across various layers of the BERT model: First layer (Left), Second layer (Middle Left), Second-to-Last layer (Middle Right), and Last layer (Right)

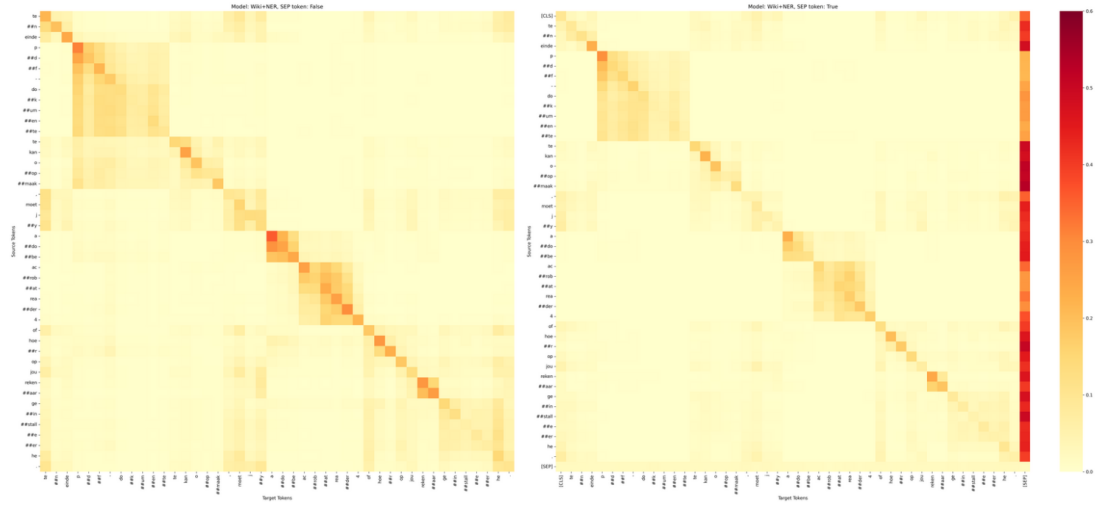

Fig. 18. Attention maps of Sentence One - Table II with (RIGHT) and without (LEFT) special tokens for inference

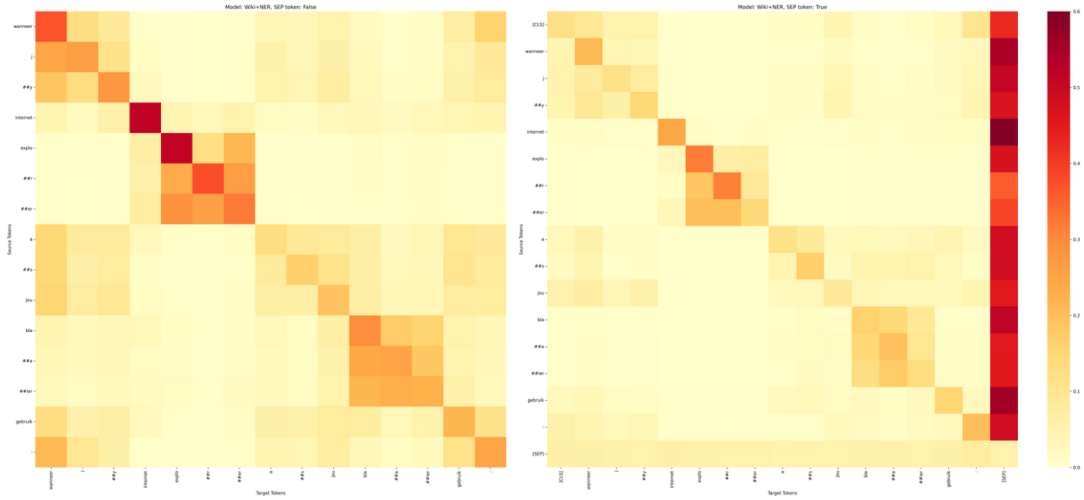

Fig. 19. Attention maps of Sentence Three - Table IV with (RIGHT) and without (LEFT) special tokens for inference

Additional Results for Experiments on Theory 2

To analyze how different words in an input sentence impact attention maps, we conducted multiple experiments. We tested several variations, and here we present results for specific words in Sentence Two from Table III. This sentence contains the Afrikaans word "Volgende," which translates to "following" or "next" in English. We explored these translations and their corresponding attention maps, shown in Figure 14.

Our findings indicate that using English words generally decreases attention toward the [SEP] token. For instance, the word "following" shows increased attention near the word itself, highlighted by a darker square in the center map of Figure 14. Notably, "following" attracts more attention than "Volgende" (Left). The attention on "next" is even more pronounced, with a distinct block-like region in the last map, suggesting that "next" captures higher attention than both "Volgende" and "following." Similar patterns appear in Figure 16, where English words like "mirror" and "pillow" attract more direct attention than the [SEP] token. These words were rare in the Wikipedia dataset, providing insights into model behavior with uncommon tokens. It is noteworthy that these words replaced the "Volgende" word in Sentence Two. However, exceptions also arose; in Figure 9, the word "typing" almost entirely lacked influence on attention.

In our experiments with Afrikaans words (Figure 15), we tested variations like "kussing" and "antwoord" in the place of "volgende" and observed only minor shifts in attention. For "antwoord," there was minimal dark intensity around the word, while "kussing" showed slightly more, but the attention toward the [SEP] token remained dominant. Notably, in Figure 10, another exception emerged with the Afrikaans word "öorfone," where expected attention patterns were not observed.

These observations suggest that further investigation is necessary to better interpret these attention behaviors before making conclusive claims.

Deeper Look into the Attention Maps for the first and last layers

Upon analyzing the initial attention map of Sentence One (II) for BERTaal shown in Figure 7, we observed unexpected patterns where the SEP token received disproportionately high attention compared to other tokens. To better understand this phenomenon, we conducted a detailed investigation of the attention maps across different model layers, specifically examining the first and last two layers of the model head as illustrated in Figure 17. The analysis revealed distinct patterns across layers. In the initial layers, we observed no discernible or meaningful attention patterns across tokens. However, by the second layer, a diagonal pattern began to emerge in the attention head, notably with minimal attention on the SEP token. In contrast, the final two layers exhibited markedly different characteristics: they showed intense attention on the SEP token and more pronounced diagonal attention patterns. These last two layers demonstrated similar attention distributions, with both maintaining high intensity around the SEP token and consistent diagonal patterns, as evident in Figure 17.

BERTaal Attention Map Analysis: Impact of Special Tokens

We investigated how the presence or absence of special tokens, particularly the [SEP] token, affected the model's attention patterns during inference. Our analysis revealed notable differences in attention distribution across multiple test cases. For the first test sentence (Sentence One, Figure 18), the attention maps without the [SEP] token showed heightened attention intensities on both entities and other sentence tokens. When the [SEP] token was reintroduced, we observed a redistribution of attention: the intensity along the diagonal and on key entities diminished as attention shifted predominantly toward the [SEP] token. This pattern was further corroborated in our analysis of the third sentence in IV (Figure 19). Without special tokens, the model demonstrated strong attention (indicated by dark red intensity) on the entity "Internet Explorer" and maintained high attention levels across other tokens along the diagonal. However, upon reintroducing the special tokens, this focused attention pattern dissipated, with attention again redirecting notably toward the [SEP] token. This consistent shift in attention distribution suggests a significant impact of special tokens on the model's attention mechanisms.