UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

COS700 Research Report

Explainable AI For Truthful Assessment in Low-Resource Languages

Student number: u21457451

Supervisor(s):
Ms. Seani Rananga
Mr. Thapelo Sindane

Arno Jooste

October 2024

Abstract

Misinformation can spread rapidly in low-resource languages such as Afrikaans due to a lack of reliable information sources. This research investigates the effectiveness of Explainable Artificial Intelligence (XAI) in analysing the behaviour of a misinformation detection model that analyses Afrikaans misinformation translated from English. While existing models primarily target high-resource languages, this study addresses the gap in methods for misinformation detection in Afrikaans by employing a four-step methodology. First, misinformation texts from the LIAR dataset are translated from English to Afrikaans and evaluated for translation quality. Next, BERT is used to generate text encodings, which are then fed into classification models, including Random Forest, Logistic Regression, and SVM. The LIME Explainable AI (XAI) model is subsequently applied to the best-performing classification model, to provide explanations for the predictions, which are further refined for human readability. The primary objective is to assess the reliability of LIME explanations in clarifying the model's predictions. Our findings suggest that, while LIME explanations provide helpful insights, they require further refinement to increase interpretability in low-resource contexts. This study not only establishes the groundwork for utilising XAI to detect misinformation in Afrikaans but also offers opportunities for future research, such as incorporating domain adaptation techniques like DANN or improving text preprocessing for better feature extraction. Finally, the study contributes to developing more powerful and interpretable AI systems for combating misinformation in languages with limited resources.

Keywords:

Explainable Artificial Intelligence (XAI), Misinformation, Low-resource Languages

1 Introduction

The rapid rise of mobile technologies and social media has fundamentally transformed communication and news sharing [1]. However, this newfound ease of information dissemination has also facilitated the spread of misinformation, often referred to as "fake news." To combat this growing problem, researchers have explored various strategies, including models for fact-checking and misinformation detection [1, 2].

While these models show promise, a crucial aspect remains: understanding how they arrive at their predictions [3]. This has led to the integration of explainability methods within various models, applicable to both specific architectures such as Convolutional Neural Networks (CNNs) [4] and more general models [5]. Explainability methods can even be used to generate human-readable text explanations [6]. However, the accuracy of these explanations themselves can be a concern.

The authors of [2] made significant strides in generating explanations for model predictions, demonstrating their effectiveness in conveying the underlying reasoning. This offers an exciting prospect: integrating their model with a misinformation detection model to gain insights into why the model classifies information as true or false. However, a critical question emerges: what if the intended user lacks familiarity with the technical language often used in these explanations?

This limitation raises further questions: Can misinformation detection be effectively applied in languages with limited resources, such as Afrikaans? Can Explainable Artificial Intelligence (XAI) techniques accurately assess the truthfulness of Afrikaans misinformation? To address these gaps, this study proposes leveraging Explainable Artificial Intelligence (XAI) techniques, specifically LIME, to generate text explanations that illuminate a misinformation detection model's reasoning process. We then assess the reliability of these explanations.

Historically, research on Explainable Artificial Intelligence (XAI) for misinformation detection has primarily focused on well-resourced languages, such as English. This limits the accessibility and effectiveness of Explainable Artificial Intelligence (XAI) in combating misinformation across diverse communities with low-resource languages. Our research aims to bridge this gap by exploring the potential of Explainable Artificial Intelligence (XAI) to explain the workings of a misinformation detection model specifically de-

signed for Afrikaans. By examining Afrikaans, we can assess the usefulness of Explainable Artificial Intelligence (XAI) for spotting misinformation in such contexts. Furthermore, this research seeks to establish a robust methodological framework for applying Explainable Artificial Intelligence (XAI) in low-resource language settings. This framework leverages techniques such as machine translation (MT), misinformation detection and the explanations generated by the LIME model. Ultimately, this research lays the groundwork for future studies exploring Explainable Artificial Intelligence (XAI)'s potential to combat misinformation in a broader range of languages.

2 Problem Statement

The growing spread of misinformation online poses a significant threat, particularly for speakers of low-resource languages, such as Afrikaans. Limited access to reliable information sources in these communities can intensify this issue. If left unchecked, misinformation can have severe consequences, ranging from shaping public opinion on critical matters to inciting violence.

However, there is a critical deficiency in research on employing Explainable Artificial Intelligence (XAI) techniques to evaluate the veracity of misinformation in low-resource languages. Present Explainable Artificial Intelligence (XAI) applications primarily concentrate on well-resourced languages, such as English, restricting their accessibility and effectiveness in combating misinformation across diverse communities.

This study aims to bridge this gap by examining the potential of Explainable Artificial Intelligence (XAI) to explain the decision-making process of a misinformation detection model specifically designed for Afrikaans. We explore the effectiveness of Explainable Artificial Intelligence (XAI) techniques, specifically LIME, in generating human-readable explanations for the model's predictions. By evaluating the reliability of these explanations, we can assess the overall usefulness of Explainable Artificial Intelligence (XAI) in this context.

3 Methodology

This section outlines the final core methodology of our research and does not include a discussion of any models or comparisons done during the translation phase. More on these extra comparisons can be found in Section 6.3.

A general structure of the framework can be seen in Figure 1.

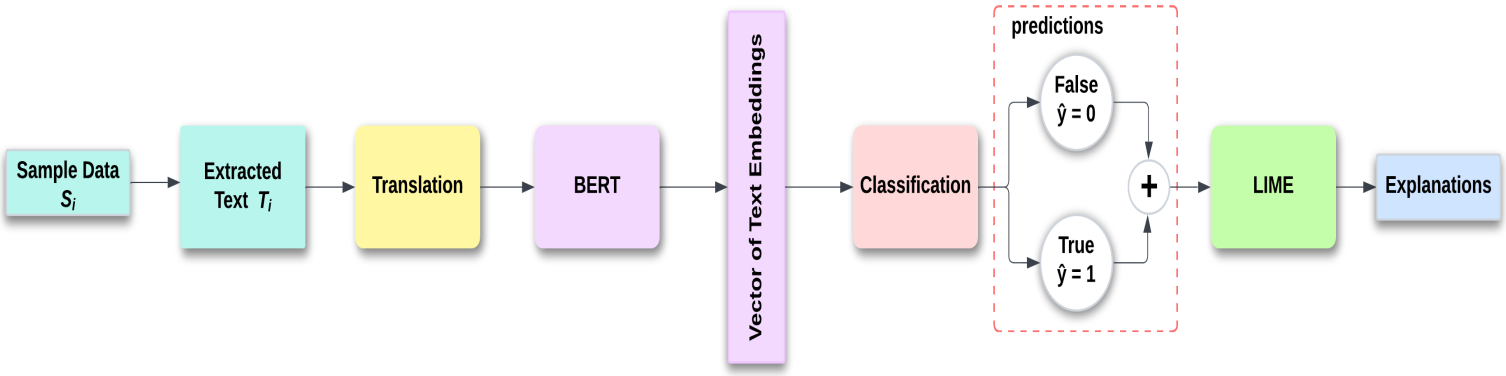


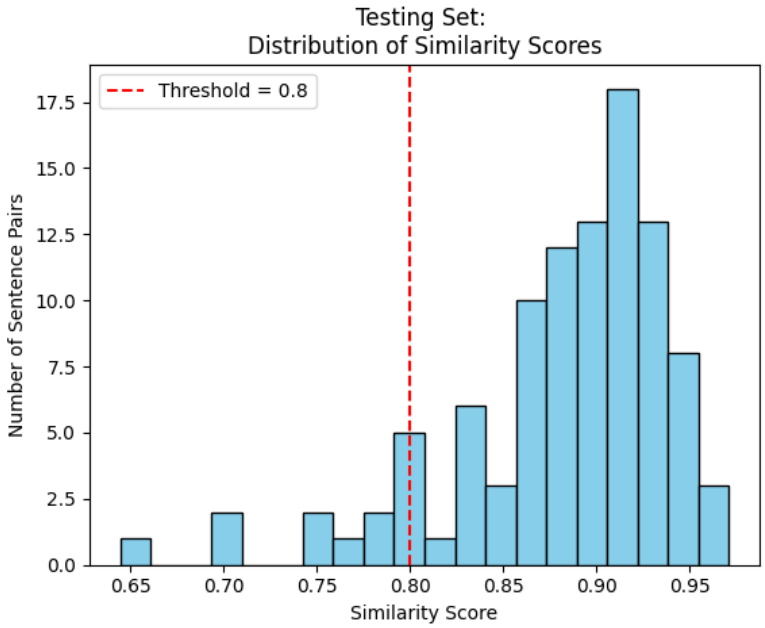
Figure 1: Overview of the system pipeline

The LIAR dataset [7] was employed to provide a foundation for misinformation analysis. Translations from English to Afrikaans were assessed using a similarity score, which measures the similarity between the English and Afrikaans texts. The translations from the *Helsinki-NLP/opus-mt-en-af* model [8, 9] were compared to Gemini's translations, with Gemini demonstrating better translation accuracy, since it had fewer similarity scores below a threshold of 80% similarity. This step ensured that the input data for the classification models was of higher quality. The original English and translated Afrikaans sentences, resulting from the models' translations, were encoded to calculate similarity scores using the Language-agnostic BERT Sentence Encoder (LaBSE) [10]. The embeddings produced from this step were then processed to calculate similarity scores using L2 normalisation. The results are shown in Table 1 and Section 3.

Table 1: Translation Quality comparison

Model	Max	Min
Gemini	0.9631	0.7459
Helsinki-NLP/opus-mt-en-af	0.9713	0.6446

Subsequently, the Google multilingual BERT (BERT-base-multilingual-cased) model [11] was utilised to generate embeddings for the classification algorithms. These embeddings served as input for three classifiers: Random



(a) Helsinki-NLP/opus-mt-en-af Translation Similarity Scores

(b) Gemini Translation Similarity Scores

Figure 2: Translation Quality: Helsinki-NLP/opus-mt-en-af (a) vs. Gemini (b)

Forest, Logistic Regression, and Support Vector Machines (SVM). Initially, all three models yielded accuracy scores around 20%, which posed potential problems for the LIME model. Fortunately, the authors of [12] encountered the same issue and provided a solution. They completed a number of experiments using binary classification by reducing the dataset's classes to simply "true" and "false." They streamlined the labels in five unique ways: (1) All labels but "true" were assigned as "false", (2) The labels "pants-fire" and "false" were assigned as "false," while the others were assigned as "true", (3) All labels were assigned as "true", except "pants-fire", (4) Labels were separated from the centre into "true" and "false", (5) All the labels were assigned as "false", except those of "true" and "mostly-true".

They used various labelling methods to identify when non-binary classification is most effective for the LIAR dataset and which labels make the language the most distinguishable. Ultimately, the third option produced the best accuracy score of 91%. Therefore, by using this approach, our classification results also improved dramatically, with Random Forest yielding the highest classification accuracy (approximately 92%), making it the preferred model for further explainability analysis.

Finally, an Explainable Artificial Intelligence (XAI) technique, specifically LIME, was applied to the Random Forest model to generate explanations for its predictions. This approach allowed us to understand the features the model used when identifying Afrikaans misinformation and provided a way to communicate these results to end-users by generating human-readable text explanations. A detailed explanation for the results is reported in Section 6.3.

4 Background

This study addresses a challenge at the intersection of three fields: Explainable Artificial Intelligence (XAI), misinformation detection, and their application to low-resource languages. The goal is to develop a system that can identify and explain the accuracy of claims in Afrikaans, a low-resource language.

4.1 Explainable AI

Explainable Artificial Intelligence (XAI) is a branch of AI that aims to clarify complex AI models, making them understandable to humans [13]. Numerous techniques have been proposed in the domain of explainability, including

simplification-based [5], perturbation-based [14], and gradient-based methods [15, 16]. Augenstein et al. [2] conducted a diagnostic study of these techniques across three text classification tasks and three model architectures (Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Transformer). Their research highlighted that gradient-based methods performed best across these tasks and architectures, using methods such as Saliency, InputXGradient, and Guided Backpropagation. While these methods show promise, it is crucial to investigate whether they are suitable for low-resource languages such as Afrikaans.

4.2 Misinformation Detection

In the context of misinformation detection, Joshi et al. [17] explored the integration of domain adaptation and XAI for misinformation detection. They utilised a Domain Adversarial Neural Network (DANN) to detect misinformation across various social media platforms, handling unseen data from similar sources. However, DANN lacks the ability to explain its classifications. To address this, they applied Local Interpretable Model-Agnostic Explanations (LIME), which makes predictions understandable. Their case study on COVID-19 misinformation demonstrated that DANN significantly improved accuracy and worked well with diverse data sources. This suggests that combining DANN with LIME offers a promising approach to combating misinformation effectively by providing both accurate detection and clear explanations. These explanations, however, might not always be accessible to non-experts, which highlights the need for constructing human-readable explanations from LIME outputs.

4.3 Low-Resource Languages

Low-resource languages (LRLs) are characterised by fewer speakers, limited digital resources, and less widespread computerisation [18, 19, 20]. Consequently, they receive less attention in education and research. Transformer models currently lead in machine translation (MT) for high-resource languages, achieving impressive quality [21]. Machine translation, as defined by [22], is the automated process of converting written text from one natural language to another, a field now dominated by Neural Machine Translation (NMT), which uses a single neural network for text translation. Lankford et al. [21] explored enhancing NMT for LRLs, specifically focusing on Irish and Marathi. They addressed the scarcity of parallel datasets by developing the

first bilingual collection of health data for Irish and introduced adaptNMT and adaptMLLM, open-source tools for NMT model creation, fine-tuning, and deployment. Their findings showed improved MT performance in low-resource scenarios by using subword models.

Similarly, Zoph et al. [23] introduced a transfer learning approach that significantly improved BLEU scores (a measure of translation quality) for various low-resource languages. Martinus et al. [24] focused on Neural Machine Translation (NMT) for African languages, including Afrikaans. They utilised the Autshumato corpora, which are aligned South African government documents specifically designed for training machine translation systems [25]. In their study, Martinus et al. [24] evaluated two NMT models: Convolutional Sequence-to-Sequence (ConvS2S) models, introduced by Gehring et al. [26] as a type of NMT model that utilises a convolutional neural network architecture for sequence-to-sequence translation, and Transformer models. Their findings showed that Transformer models achieved higher accuracy based on BLEU scores and qualitative back-translations by native speakers [24]. These findings suggest that the methodologies explored by Martinus et al. [24] can be valuable for translating English misinformation to Afrikaans.

The approaches presented by [2], [17], and [21] provide a robust foundation for future research. We propose a framework by combining pretrained misinformation detection models with machine translation models to translate misinformation data from English to Afrikaans.

Overall, the model's performance was evaluated on a held-out test set to assess its accuracy in both identifying the truthfulness of claims and generating understandable explanations about the Afrikaans text. By leveraging the findings from previous research, we developed a system that not only detects misinformation in Afrikaans but also provides clear, human-readable explanations for its predictions.

5 Related Work

De Jager et al. investigated multimodal approaches to detecting misinformation in social media, specifically for the South African context [27]. They explored a multimodal approach using the Fakeddit and a new South African misinformation dataset, namely Real411, datasets. The authors emphasised the limitations of applying models trained on non-local data (e.g., American) to South African social media due to cultural and linguistic differences.

To address this, their research introduced a South African misinformation dataset sourced from Real411, a fact-checking initiative. The paper proposed the Multimodal Misinformation-in-Context (MMiC) model, leveraging pre-trained BERT and ResNet for text and image features respectively. Their model demonstrated competitive performance compared to unimodal baselines on the Fakeddit dataset. The authors investigated the impact of using local and non-local data for training, highlighting the potential benefit of incorporating local data for improved performance in specific contexts. While their research focused on multimodal approaches, our study primarily focused on text data, particularly Afrikaans text translated from English. De Jager et al.'s work inspired our methodological framework, providing a valuable foundation for our research.

Another study done by Aslan et al. aimed to explore the automatic detection of false news using machine learning and natural language processing techniques [12]. The authors compared the effectiveness of various computational methods and analysed the differences between human-generated and automatically-generated false news. They reviewed previous research on false news detection, including studies that used linguistic analysis, user detection, and multi-source fusion to identify different degrees of falsehood in news articles. In their research they made experimented with various false news datasets, including the LIAR dataset. The LIAR dataset includes six classes to label an English statement. After achieving only 20% classification accuracy, using classification models such as Random Forest, and Support Vector Machine (SVM), Aslan et al. conducted experiments by manipulating the this six-way classification problem into a binary classification problem. From their experiments they concluded that one of their methods achieved a highest accuracy score of 91%. The experiments conducted by Aslan et al. allowed us to make use of the same classification models, while also maintaining high accuracy scores, which was needed for the LIME model.

6 Discussion

6.1 Developments

This study developed a methodological framework for understanding misinformation detection models in low-resource languages. Furthermore, we investigated the application of a XAI technique, specifically LIME, to explain the predictions of the best classification model, selected from several classification models (Random Forest, Logistic Regression, and SVM), on

Afrikaans misinformation text. We aimed to provide a reliable and transparent strategy for misinformation detection that can be extended to other low-resource languages by comparing the performance of various models.

6.2 Technologies

The technologies used in this project include both a virtual environment and a local machine. The virtual environment was provided by Kaggle, which offered access to two T4 GPUs to accelerate computations. On the local machine, the hardware consisted of an Intel(R) Core(TM) i7-1065G7 CPU operating at 1.30 GHz with a boost to 1.50 GHz, along with 16 GB of installed RAM (15.8 GB usable). The system is a 64-bit operating system, running on an x64-based processor. These resources were crucial for handling the computational demands of the project.

6.3 Results

6.3.1 Evaluation Metrics

To evaluate the performance of the machine translation models and the subsequent misinformation detection task, we employed a combination of standard metrics:

Machine Translation Evaluation

- **BLEU (BiLingual Evaluation Understudy):**

$$BLEU = BP \exp\left(\sum_{n=1}^{N} w_n \ln(p_n)\right)$$

where:

- BP is the Brevity Penalty.
- w_n represents the importance or weight assigned to n-gram precision of order n.
- p_n represents the n-gram modified precision score of order n.
- N represents the maximum n-gram order which should be considered.

- **METEOR (Metric for Evaluation of Translation with Explicit ORdering):**

$METEOR = F_{mean} \times (1 - Penalty)$
where:

- F_{mean} represents the harmonic mean of unigram recall and precision
- $(1 - Penalty)$ represents the word order penalty

- **Precision, Recall, and F1-Score:**

  Precision: $Precision = \frac{TP}{TP+FP}$

  Recall: $Recall = \frac{TP}{TP+FN}$

  F1-Score: $F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$
  where TP is true positive, FP is false positive, and FN is false negative.

Misinformation Detection Evaluation

- Accuracy: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- Precision: $Precision = \frac{TP}{TP+FP}$
- Recall: $Recall = \frac{TP}{TP+FN}$
- F1-Score: $F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$

Similarity Score Calculation

- LaBSE (Language-agnostic BERT Sentence Encoder): This pre-trained BERT-based model generates sentence embeddings for 109 languages [10].
- L2 Normalisation is applied using the sentence embeddings to calculate the similarity scores.

6.3.2 Results discussion

As mentioned in Section 3 our final core methodology only focuses on the final outline containing the best performing models and approaches, in the context of this study. However, in this section we include a detailed discussion to ellaborate on how we reached our conclusions on the best performing models and approaches.

The first two translation models used for comparison, were the *facebook/mbart-large-50-many-to-many-mmt* [28] and *Helsinki-NLP/opus-mt-en-af* translation models. We first compared these models using two parallel corpora:

Autshumato and NLLB. The Autshumato Afrikaans-English bilingual corpus is an aligned parallel corpora for the Afrikaans-English language pair. It includes aligned text segments obtained from items for which the Centre for Text Technology has distribution rights, such as government documents, publications, bulletins, and policies [25, 29]. We used precision, recall, and F1-scores to compare the models on each corpus across various sample sizes. Additionally, we measured the performance using sentence-by-sentence and word-by-word translations. These results are illustrated in Figures 3 to 8. From these results we concluded that the NLLB corpus was the better corpus to use. Both models achieved better overall translation performance when working on the NLLB corpus. Therefore, for our next comparison we made use of the OPUS collection's [30] NLLB corpus, and included another one of its corpora: CCMatrix.

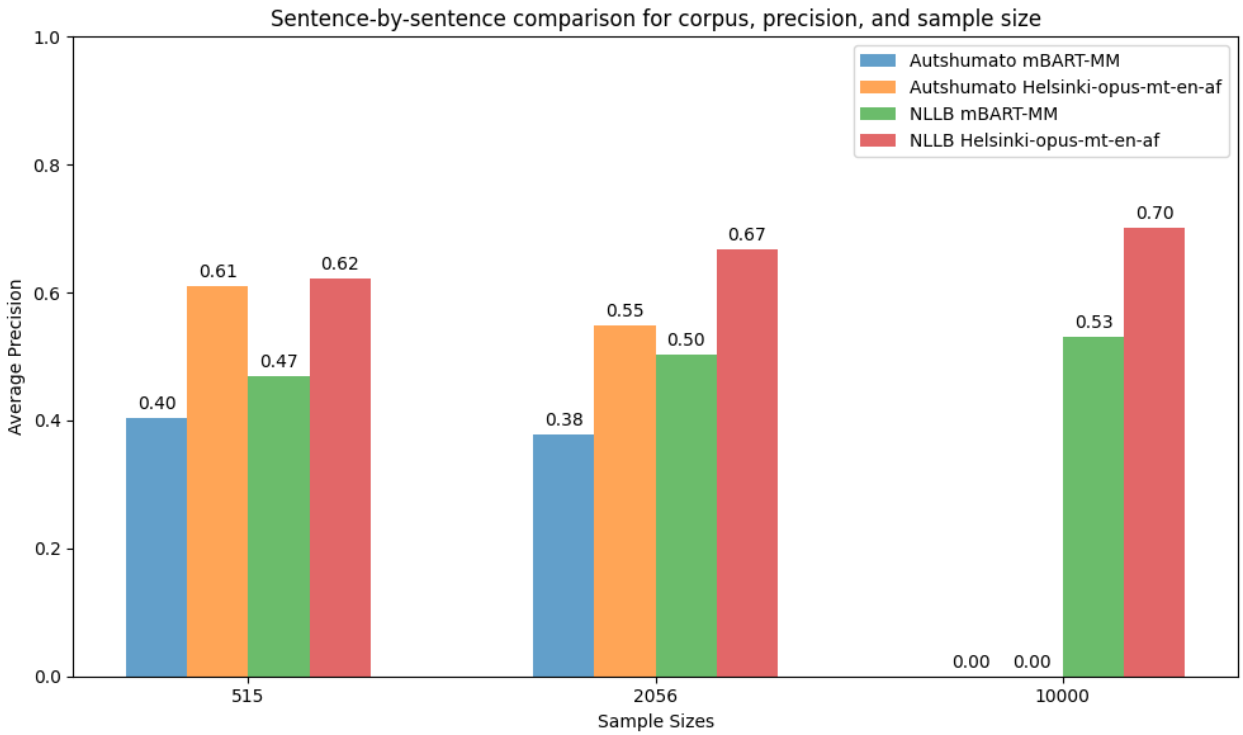


Figure 3: Sentence-by-sentence comparison of Precision metric

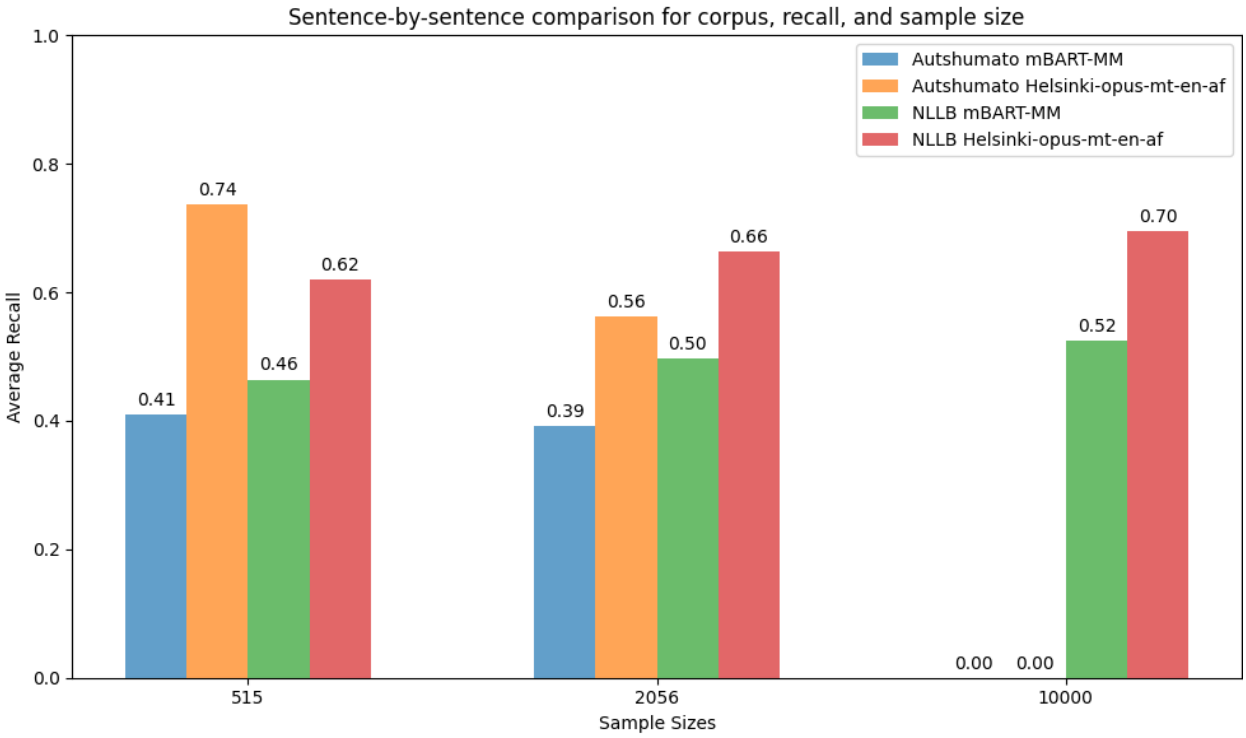


Figure 4: Sentence-by-sentence comparison of Recall metric

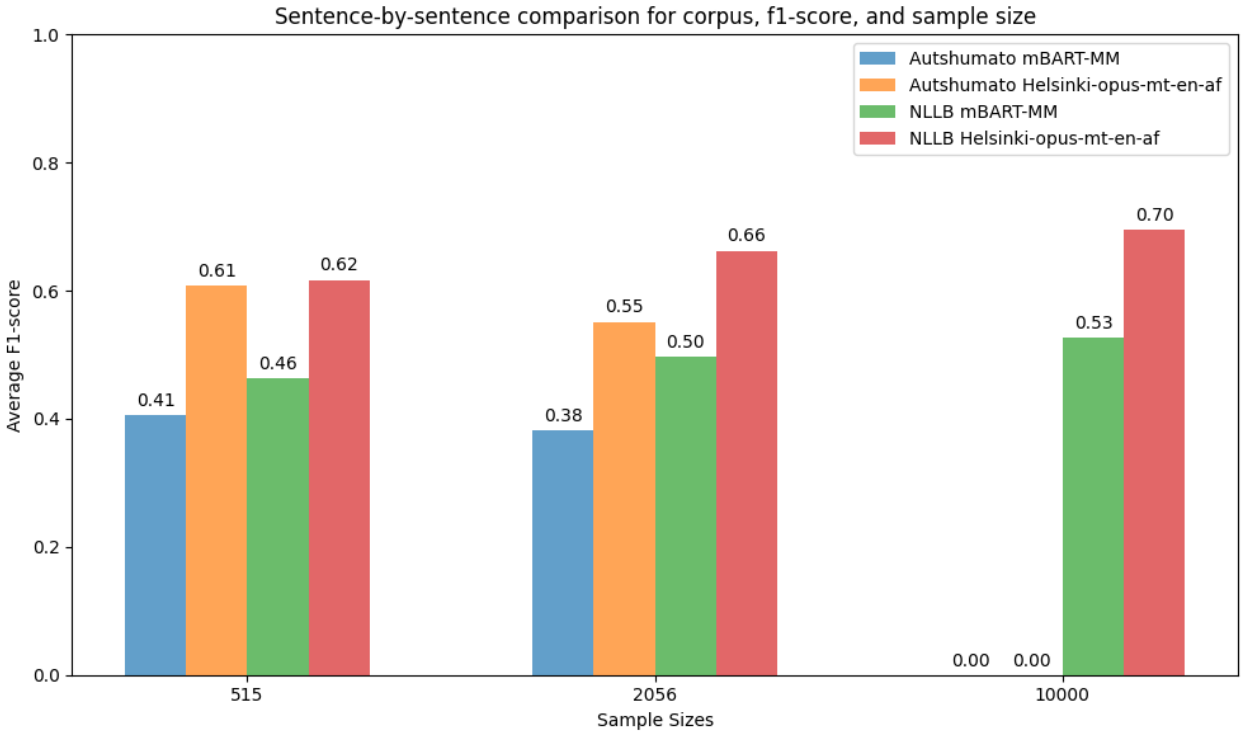


Figure 5: Sentence-by-sentence comparison of F1-score metric

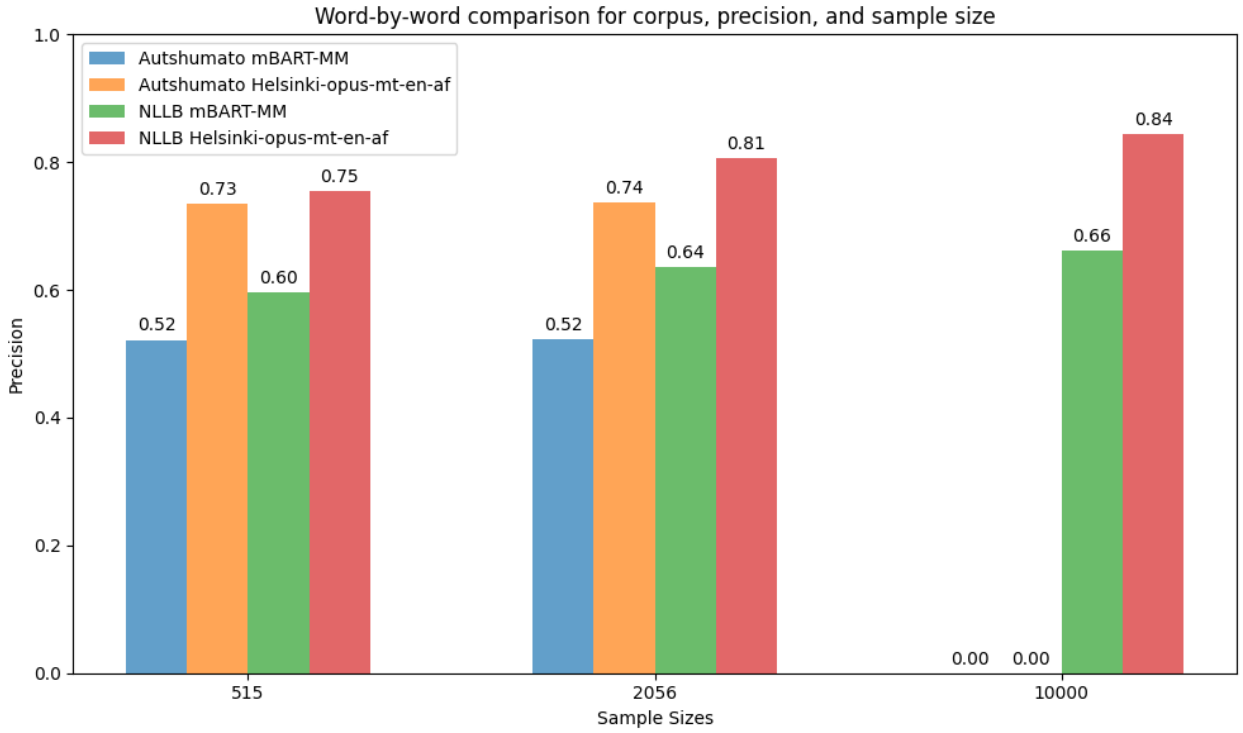


Figure 6: Word-by-word comparison of Precision metric

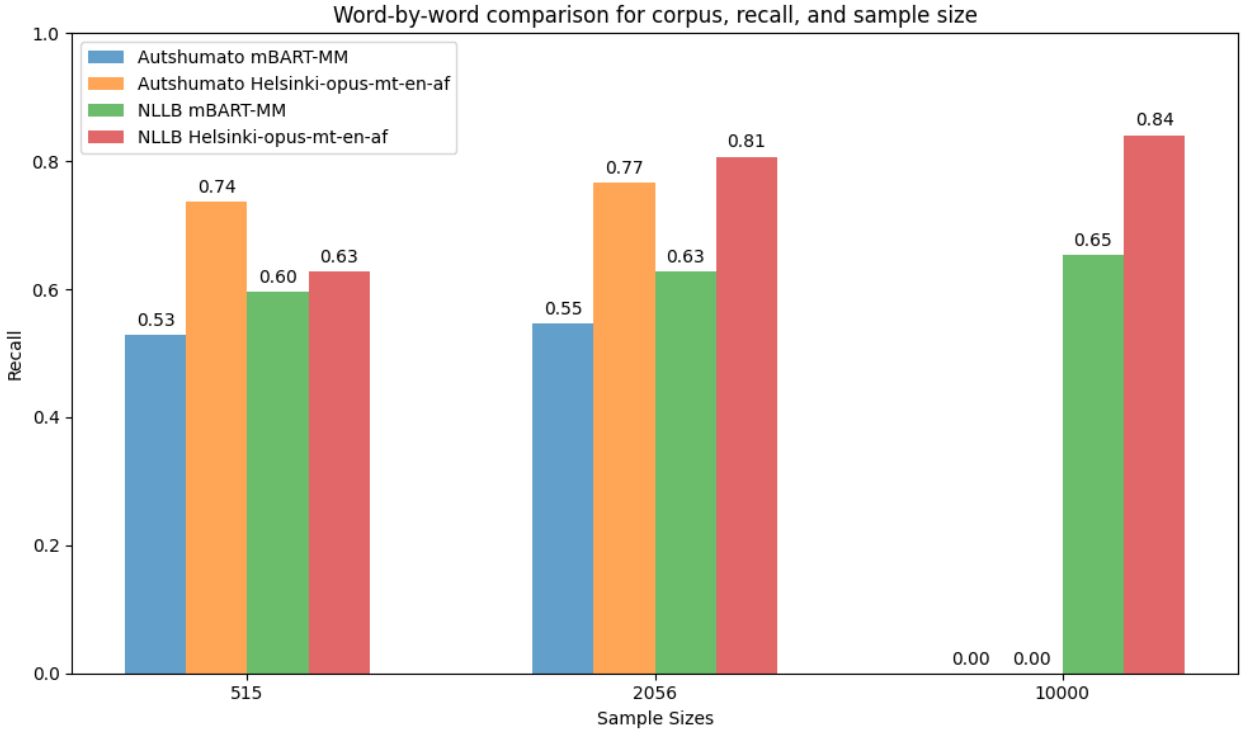


Figure 7: Word-by-word comparison of Recall metric

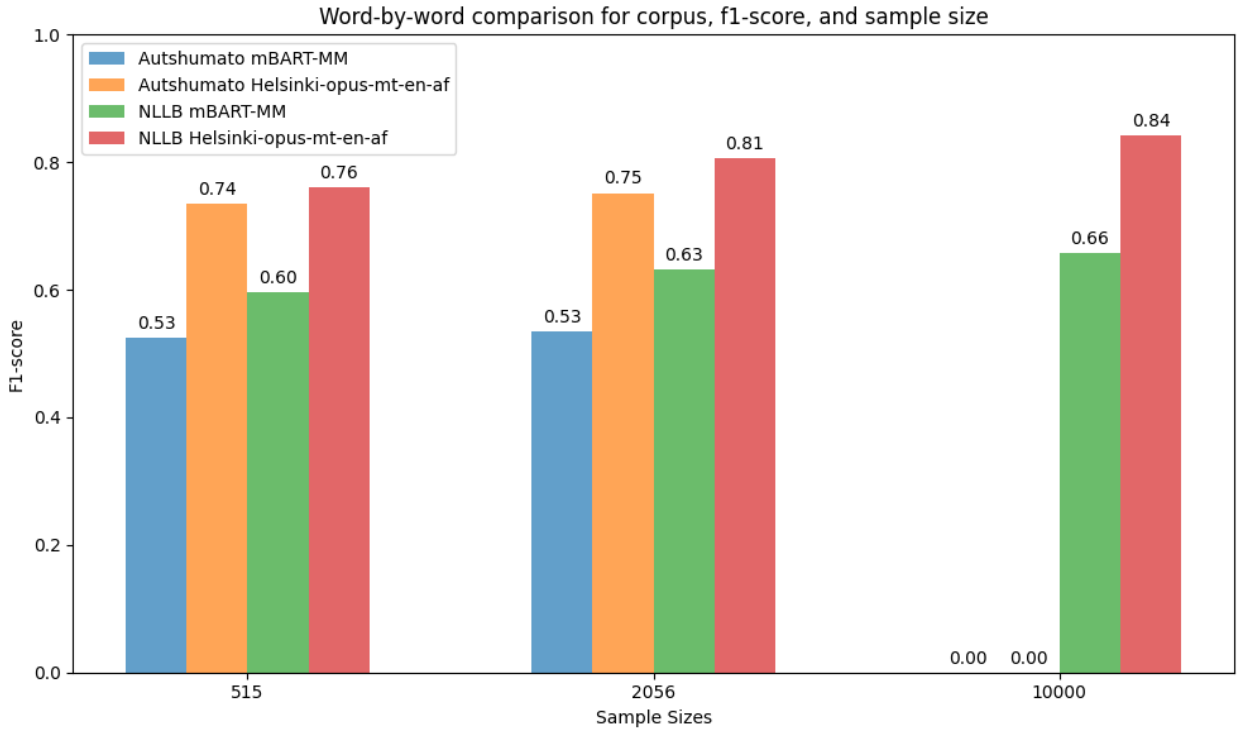


Figure 8: Word-by-word comparison of F1-score metric

We used BLEU and METEOR metrics to evaluate the performance of both models across the CCMatrix and NLLB corpora. From the results shown in Figures 9 and 10 we concluded that the *Helsinki-NLP/opus-mt-en-af* model achieved the better translation results using the NLLB corpus. It is important to note that even though the NLLB corpus proved to be a better parallel corpus to use, this corpus did not focus its content on misinformation. However, this corpus helped us in determining the better translation model. Before selecting the Helsinki model as the chosen model for translating a misinformation dataset, LIAR, from English text to Afrikaans, we wanted to compare its performance with a large language model (LLM), specifically Google's Gemini.

We evaluated the final two translation models, *Helsinki-NLP/opus-mt-en-af* and Gemini, using the LIAR dataset and similarity scores between the English and translated sentences. The performance results were already shown in Table 1 and section 3. Overall, Gemini achieved better performance, indicating that Gemini produced more accurate and fluent translations. Once we selected Gemini as the preferred translation model, we used its translations for the classification tasks.

The classification models used for our study were Random Forest, SVM,

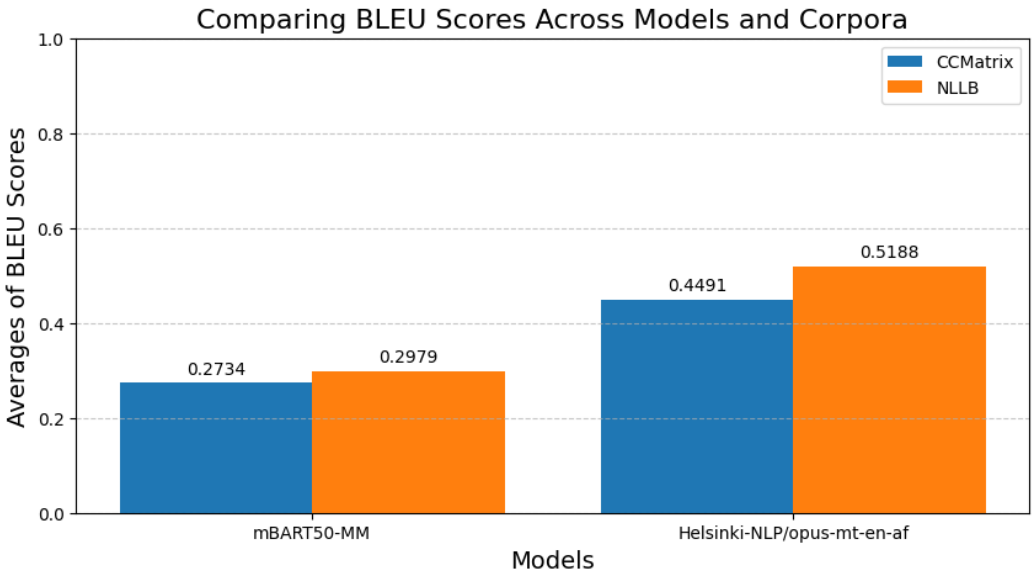


Figure 9: BLEU Score comparison between two translation models across two corpora

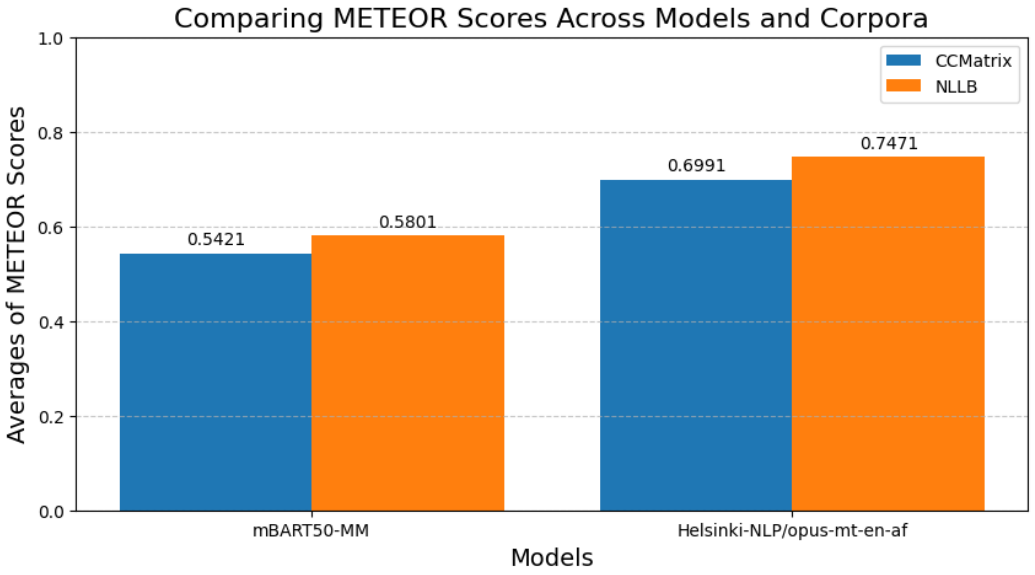


Figure 10: METEOR Score comparison between two translation models across two corpora

and Logistic Regression. These models were evaluated using accuracy, precision, recall, and F1-score metrics. As mentioned in Section 3 the classification models achieved, at best, around 20% classification accuracy. This is a common problem when working with the LIAR dataset [12]. Hence, we made use of Aslan et al.'s approach by relabelling the classes such that all were considered 'true', except for the class 'pants-fire'.

The new classification results, reported in Table 2, Table 3, and Table 4, show promising performance for the class labeled as 'True'/'Truthful' (encoded as '1'). However, the models struggled significantly when classifying the 'False'/'Misinformation' class, performing poorly for this category. This mismatch is likely caused by the relabelling approach employed during preprocessing, which introduced bias and impacted the model's ability to accurately classify Afrikaans text as misinformation. Even though the training data was balanced, this relabelling approach contributed to these misclassifications, particularly in edge cases where subtle differences of misinformation were misrepresented.

Table 2: Testing Set Classification Report using Random Forest

Class	precision	recall	f1-score	support
0	0.08	0.01	0.02	92
1	0.93	0.99	0.96	1170
accuracy			0.92	1262
macro avg	0.51	0.50	0.49	1262
weighted avg	0.87	0.92	0.89	1262

Table 3: Testing Set Classification Report using Logistic Regression

Class	precision	recall	f1-score	support
0	0.08	0.30	0.12	92
1	0.93	0.71	0.80	1170
accuracy			0.68	1262
macro avg	0.50	0.51	0.46	1262
weighted avg	0.87	0.68	0.75	1262

Table 4: Testing Set Classification Report using SVM

Class	precision	recall	f1-score	support
0	0.07	0.30	0.12	92
1	0.93	0.69	0.79	1170
accuracy			0.66	1262
macro avg	0.50	0.50	0.46	1262
weighted avg	0.86	0.66	0.74	1262

To address this challenge, future research could explore more multi-class classification models that would eliminate the need for relabelling, particularly on the LIAR dataset. Additionally, a better approach might be to merge more Afrikaans-translated fake news datasets. This could provide a broader, more varied resource pool for training, validation, and testing. Expanding the dataset in this manner would enhance the models' capacity to generalise across different contexts and misinformation patterns.

Furthermore, using BERT for sentence encoding increased the complexity of the models by capturing deeper contextual relationships between words. This level of complexity may be beneficial, particularly for a low-resource language such as Afrikaans, where capturing these deeper relationships is critical to improving classification accuracy. However, this complexity may have been too complex for the Explainable AI (XAI) model, LIME, which focusses on word-level implications. Simpler models, such as TF-IDF, that focus on essential word-level features, may provide an improved solution for low-resource languages, as it focuses on the importance of words based on their frequency within a sentence. Future research should investigate whether TF-IDF might improve pattern recognition in misinformation detection, especially for low-resource languages.

Finally, we used LIME to explain the model's predictions. LIME provides interpretable explanations by highlighting the most important features that contributed to the model's decision. This helped us understand how the model was making its predictions and identify potential biases or limitations.

The LIME model was applied to explain the Random Forest classifier's decisions, which were then transformed into human-readable interpretations using the weights assigned to each word. Examples of these explanations are provided in Figure 11 and Figure 12. Interestingly, LIME assigned low weights to individual words. This could be attributed to two challenges that were hinted before. The first is the model's confidence. The Random Forest

Table 5: LIME explanations for the test sentences (Random-Forest-based classifier).

Test Set Index	English Sentence	Afrikaans Sentence	Probability 'Misinformation'	Probability 'Truthful'
0	Building a wall on the U.S.-Mexico border will take literally years.	Om 'n muur op die VSA.-Meksiko-grens te bou, sal letterlik jare neem.	0.22	0.78
170	We spend in tax loopholes annually $1.1 trillion. Thats more than we spend on our defense budget in a year, on Medicare or Medicaid in a year.	Ons bestee $1,1 triljoen jaarliks aan belastinggappies. Dis meer as wat ons aan ons verdedigingsbegroting of aan Medicare of Medicaid in 'n jaar bestee.	0.63	0.37

The model correctly predicted this text as **Truthful**, and the actual class was **Truthful**.
- 'grens' supported **Truthful** (positive weight: 0.01168).
- 'sal' supported **Truthful** (positive weight: 0.01121).
- 'letterlik' supported **Truthful** (positive weight: 0.01011).
- 'die' supported **Truthful** (positive weight: 0.00827).
- 'jare' opposed **Truthful** (negative weight: -0.00789), contributing to 'Misinformation'.
- 'bou' opposed **Truthful** (negative weight: -0.00426), contributing to 'Misinformation'.
- 'Om' supported **Truthful** (positive weight: 0.00414).
- 'neem' opposed **Truthful** (negative weight: -0.00347), contributing to 'Misinformation'.
- 'VSA' supported **Truthful** (positive weight: 0.00280).
- 'op' supported **Truthful** (positive weight: 0.00215).
- 'te' supported **Truthful** (positive weight: 0.00190).
- 'n' opposed **Truthful** (negative weight: -0.00178), contributing to 'Misinformation'.
- 'muur' opposed **Truthful** (negative weight: -0.00069), contributing to 'Misinformation'.
- 'Meksiko' opposed **Truthful** (negative weight: -0.00007), contributing to 'Misinformation'.

In this instance, the model's important features led it to predict **Truthful**.

Figure 11: Sentence 0: Human-readable Explanation generated from LIME model's output

classifier had a high bias towards identifying texts as 'Truthful,' therefore the model had already made a prediction based on more significant attributes, decreasing the influence of individual words.

Second, BERT has resulted in a high dimensional text representation. BERT's capacity to capture complex contextual relationships may reduce the significance of individual words, contributing to the low LIME weights. The complexity of BERT's embeddings may have presented difficulties in this low-resource context, where a simpler model such as TF-IDF may have performed better by focussing on less complex word-level features.

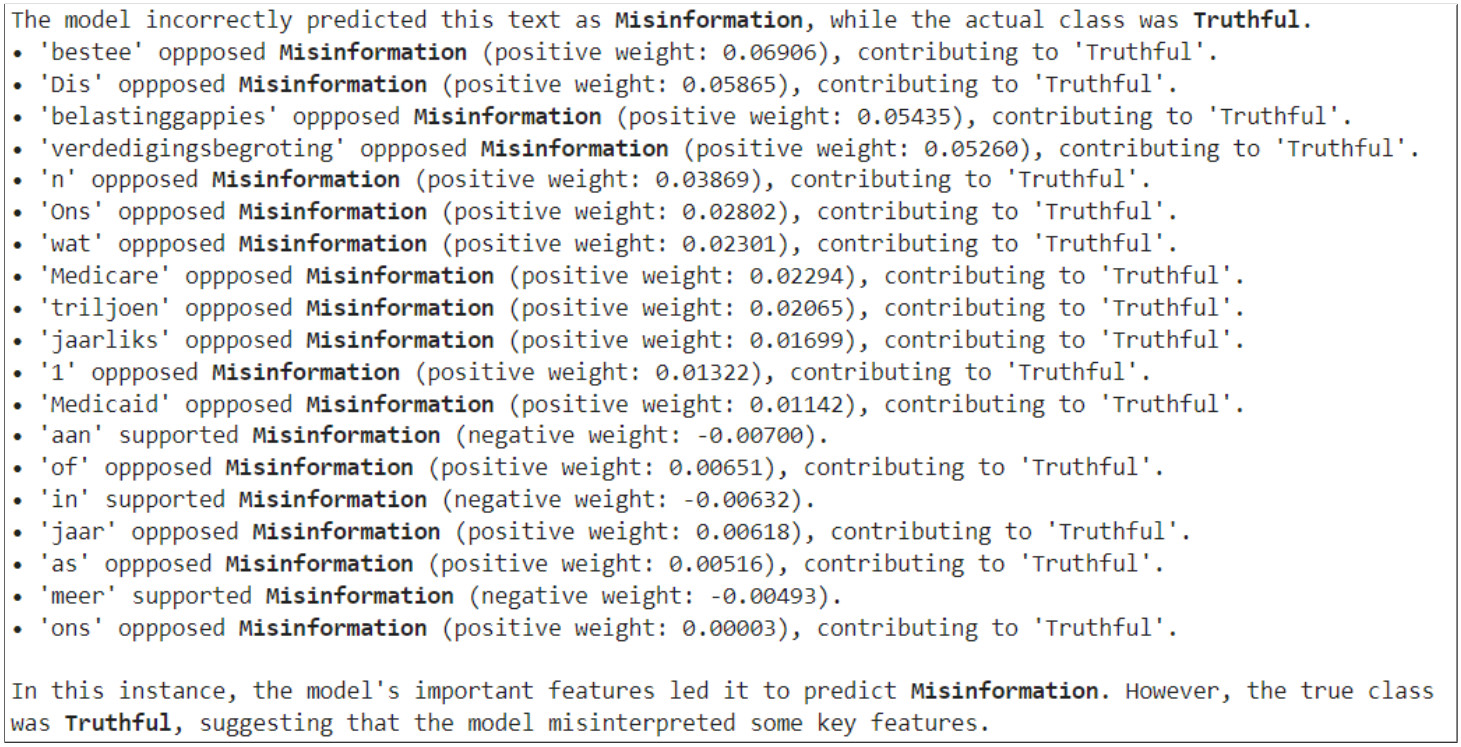


Figure 12: Sentence 170: Human-readable Explanation generated from LIME model's output

While the LIME explanations provided useful insights into the model's decision-making process, there were several misunderstandings, particularly when words were grouped as contributing to 'Truthful' when the correct class was 'Misinformation.' One example of such a misunderstanding is the second Afrikaans sentence, provided in Table 5. This limitation, which is worse in low-resource languages, emphasises the importance of being cautious when analysing word-level explanations in translated texts, as wider contextual differences are more difficult to capture.

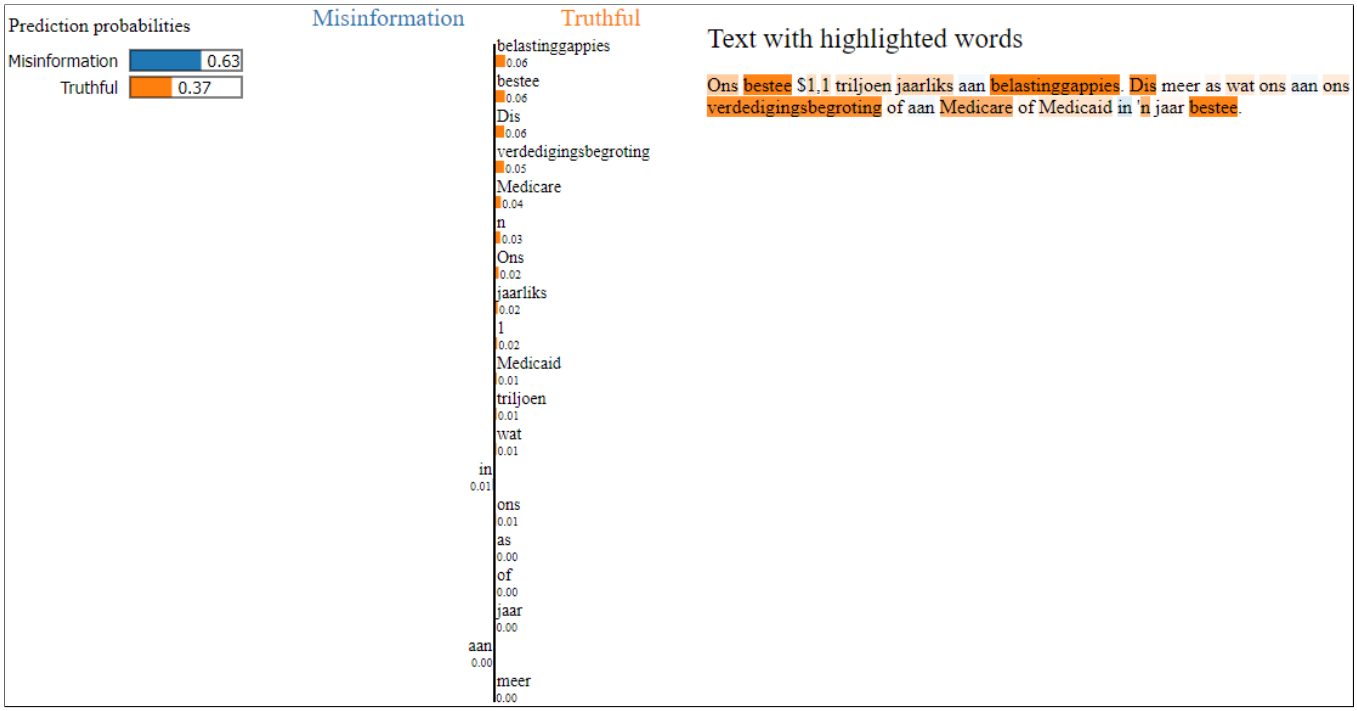


Figure 13: LIME Explanation for Afrikaans Sentence (Test Set Index 170)

Nevertheless, LIME's transparency can increase the faith in understanding the model's predictions and provides opportunities for further development, especially in low-resource languages.

7 Conclusion

This research demonstrated the potential of using Explainable AI (XAI) for truthful content assessment in low-resource languages, specifically Afrikaans. While this study focused on integrating BERT with various classification models and LIME, future research could explore alternative approaches.

One promising avenue is to investigate the combination of domain adaptation techniques, such as Domain Adversarial Neural Networks (DANN) [17], with XAI methods. DANN can potentially help mitigate the challenges posed by limited data in low-resource languages by adapting the model to unseen data from similar sources.

Additionally, exploring alternative tokenizers and text preprocessing techniques optimised for specific languages or dialects could lead to better feature extraction and improved classification accuracy. Furthermore, investigating

other sophisticated multi-class classifiers, like neural networks, could enhance the model's ability to capture complex linguistic patterns, particularly in misinformation detection.

Finally, further improvements in the LIME model's local explanations, such as incorporating contextual embeddings or using simpler word-level features, could provide more nuanced explanations that better reflect the model's decision-making process in low-resource settings.

By addressing these potential limitations, future studies could strengthen the efficacy of AI-based approaches for misinformation detection in low-resource languages, improving both classification accuracy and the interpretability of model decisions.

References

[1] D. WALKER, S. RANANGA, B. ISONG, and V. MARIVATE, "Generalising across domains in video misinformation detection," in *2024 IST-Africa Conference (IST-Africa)*, pp. 1–8, 2024.

[2] I. Augenstein, "Towards explainable fact checking," *arXiv preprint arXiv:2108.10274*, 2021.

[3] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A survey of the state of explainable ai for natural language processing," *arXiv preprint arXiv:2010.00711*, 2020.

[4] J. Wagner, J. M. Kohler, T. Gindele, L. Hetzel, J. T. Wiedemer, and S. Behnke, "Interpretable and fine-grained visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9097–9107, 2019.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[6] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom, "e-snli: Natural language inference with natural language explanations," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[7] W. Y. Wang, ""liar, liar pants on fire": A new benchmark dataset for fake news detection," 2017.

[8] J. Tiedemann, M. Aulamo, D. Bakshandaeva, M. Boggia, S.-A. Grönroos, T. Nieminen, A. Raganato Y. Scherrer, R. Vazquez, and

S. Virpioja, "Democratizing neural machine translation with OPUS-MT," *Language Resources and Evaluation*, no. 58, pp. 713–755, 2023.

[9] J. Tiedemann and S. Thottingal, "OPUS-MT — Building open translation services for the World," in *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, (Lisbon, Portugal), 2020.

[10] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," 2020.

[11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[12] L. Aslan, M. Ptaszynski, and J. Jauhiainen, "Are strong baselines enough? false news detection with machine learning," *Future Internet*, vol. 16, no. 9, 2024.

[13] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, *Explainable AI: interpreting, explaining and visualizing deep learning*, vol. 11700. Springer Nature, 2019.

[14] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), (Cham), pp. 818–833, Springer International Publishing, 2014.

[15] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, PMLR, 06–11 Aug 2017.

[16] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[17] G. Joshi, A. Srivastava, B. Yagnik, M. Hasan, Z. Saiyed, L. A. Gabralla, A. Abraham, R. Walambe, and K. Kotecha, "Explainable misinformation detection across multiple social media platforms," *IEEE Access*, vol. 11, pp. 23634–23646, 2023.

[18] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource languages: A review of past work and future challenges," 2020.

[19] C. Cieri, M. Maxwell, S. Strassel, and J. Tracey, "Selection criteria for low resource language programs," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, eds.), (Portorož, Slovenia), pp. 4543–4549, European Language Resources Association (ELRA), May 2016.

[20] A. K. Singh, "Natural language processing for less privileged languages: Where do we come from? where are we going?," in *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 2008.

[21] S. Lankford, "Enhancing neural machine translation of low-resource languages: Corpus development, human evaluation and explainable ai architectures," *arXiv preprint arXiv:2403.01580*, 2024.

[22] F. Stahlberg, "Neural machine translation: A review," *Journal of Artificial Intelligence Research*, vol. 69, p. 343–418, Oct. 2020.

[23] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," 2016.

[24] L. Martinus and J. Z. Abbott, "A focus on neural machine translation for african languages," *arXiv preprint arXiv:1906.05685*, 2019.

[25] H. J. Groenewald and W. Fourie, "Introducing the autshumato integrated translation environment," in *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, 2009.

[26] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 1243–1252, PMLR, 06–11 Aug 2017.

[27] A. De Jager, V. Marivate, and A. Modupe, "Multimodal misinformation detection in a south african social media environment," in *Artificial Intelligence Research* (A. Pillay, E. Jembere, and A. J. Gerber, eds.), (Cham), pp. 285–299, Springer Nature Switzerland, 2023.

[28] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," 2020.

[29] CTexT® (Centre for Text Technology, North-West University), South Africa and SADiLaR (South African Centre for Digital Language Resources), South Africa and Department of Sport, Arts and Culture, South Africa, "Autshumato Parallel Corpus: English-Afrikaans." `http://humanities.nwu.ac.za/ctext`, `https://sadilar.org`, `http://www.dac.gov.za`, 2024. Accessed: 2024-10-22.

[30] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, eds.), (Istanbul, Turkey), pp. 2214–2218, European Language Resources Association (ELRA), May 2012.