Explainable AI for Truthful Assessment in Low-Resource Languages

Arno JOOSTE[1]
[1] *University of Pretoria, 21A Bernard Road, Roodepoort, 2040, South Africa*
Tel: +27 726009123, Email: u21457451@tuks.co.za

Abstract: Misinformation can spread rapidly in low-resource languages such as Afrikaans due to a lack of reliable information sources. This study investigates the effectiveness of Explainable Artificial Intelligence (XAI) in analysing a misinformation detection model for Afrikaans misinformation translated from English. The research uses a four-step methodology, starting with translating misinformation texts from English to Afrikaans and evaluating their quality. Text encodings are subsequently generated using BERT, which are then fed into classification models such as Random Forest, Logistic Regression, and SVM. Finally, the LIME Explainable AI (XAI) model is applied to the best-performing model, providing explanations for predictions, which are further refined for human readability. Our findings suggest that while LIME explanations offer valuable insights, they need further refinement to enhance interpretability in low-resource contexts. The study not only establishes the foundation for XAI in detecting misinformation in Afrikaans but also offers opportunities for future research, such as domain adaptation techniques (e.g. DANN) or improved text preprocessing for better feature extraction.

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. 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].

Misinformation poses a significant threat to society, particularly to speakers of low-resource languages such as Afrikaans, who may have limited access to reliable fact-checking sources. Low-resource languages (LRLs) are characterised by fewer speakers, limited digital resources, and less widespread computerisation [2]–[4]. Consequently, they receive less attention in education and research. Therefore, the ability to detect and combat misinformation in these language is crucial in ensuring the integrity of public discourse and protecting individuals from harm.

This research focuses on addressing the challenges of misinformation detection for low-resource languages using Explainable Artificial Intelligence (XAI). XAI is a branch of AI that aims to make complex AI models more understandable to humans, thereby increasing trust and transparency [5].

While existing misinformation detection models often cater to well-resourced languages such as English, there is a pressing need to develop effective solutions for low-resource languages.

Furthermore, the growing interest in explainability highlights the importance of understanding how models make decisions, ensuring that their predictions can be interpreted by human users.

Previous research by Joshi et al. [6] has demonstrated the potential of integrating domain adaptation and XAI for misinformation detection. Their approach, which combines a Domain Adversarial Neural Network (DANN) with Local Interpretable Model-Agnostic Explanations (LIME), offers a promising framework for combating misinformation effectively.

However, their study also highlighted the limitations of LIME in providing explanations that are accessible to non-experts. This raises the question of whether misinformation detection can be effectively applied in low-resource languages such as Afrikaans, and whether XAI techniques can accurately assess the truthfulness of Afrikaans misinformation.

This research aims to address these challenges by exploring the integration of machine translation, classification algorithms, and XAI techniques to provide a framework for misinformation detection in Afrikaans. The primary goal is to develop a system that not only accurately detects misinformation but also provides clear and understandable explanations to users.

By understanding the underlying reasoning behind the model's predictions, users can evaluate the credibility of information and make informed decisions. This is particularly important in low-resource settings where access to reliable information may be limited.

2 Methodology

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

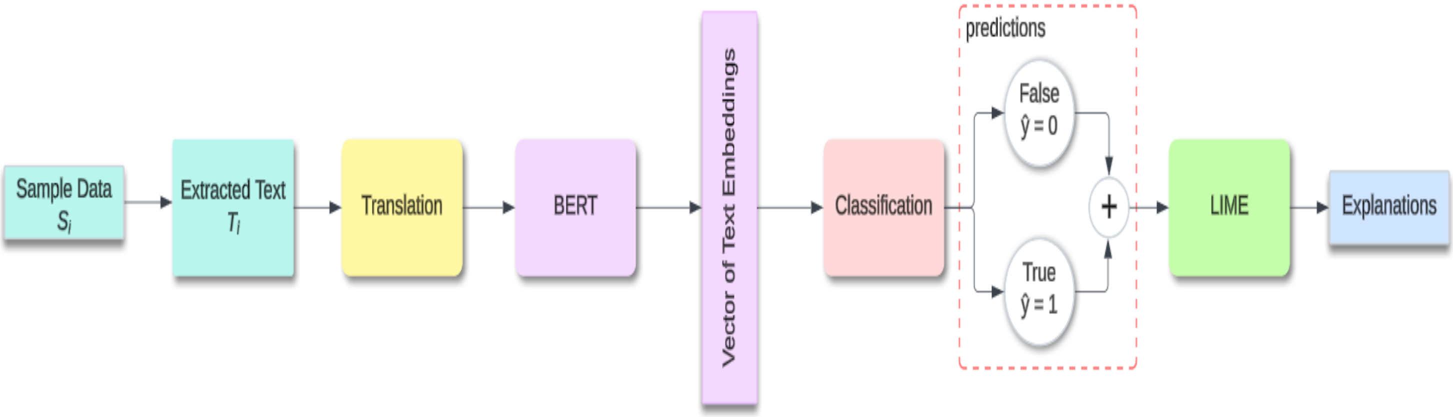


Figure 1: Overview of the system pipeline

The LIAR dataset [7] was employed to provide a foundation for misinformation analysis. Translations from English to Afrikaans was assessed using a similarity score, which measures the similarity between the English and Afrikaans texts. The translations from the helsinki-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 Figure 2.

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 Forest, Logistic Regression, and Support Vector Machines

Table 1: Translation Quality comparison

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

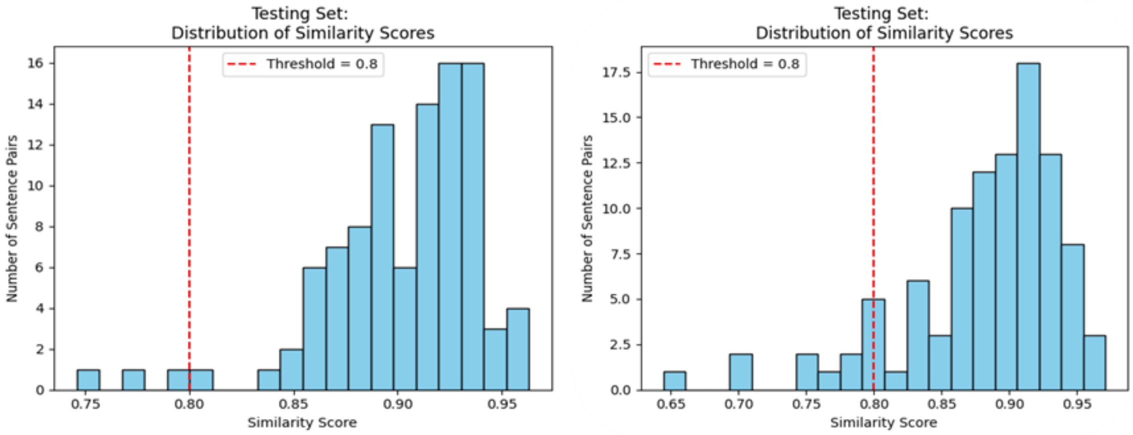


Figure 2: Translation Quality: Gemini (left) vs. Helsinki-opus-mt-en-af (right)

(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 conducted a series of different binary classification experiments, reducing the dataset's labels to only "true" and "false." They simplified the labels in five different ways:

1. All labels except "true" were labelled as "false."

2. Labels "pants-fire" and "false" were labelled as "false," while the rest were labelled as "true."

3. All labels except "pants-fire" were labelled as "true."

4. Labels were split from the middle into "true" and "false."

5. Labels "true" and "mostly-true" were labelled as "true," while the rest were labelled as "false."

They used these labelling methods to determine when non-binary classification works best for the LIAR dataset and which labels make the language the most differentiable. 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 5.

3 Technology Description

- Virtual environment:
  - Environment provided by Kaggle
  - Accelerator: GPU T4 (x2)
- Local machine:
  - Processor: Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz 1.50 GHz
  - Installed RAM: 16.0 GB (15.8 GB usable)
  - System Type: 64-bit operating system, x64-based processor

4 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.

5 Results

The 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 contributed to these misclassifications, particularly in edge cases where subtle differences of misinformation were misrepresented.

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.

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

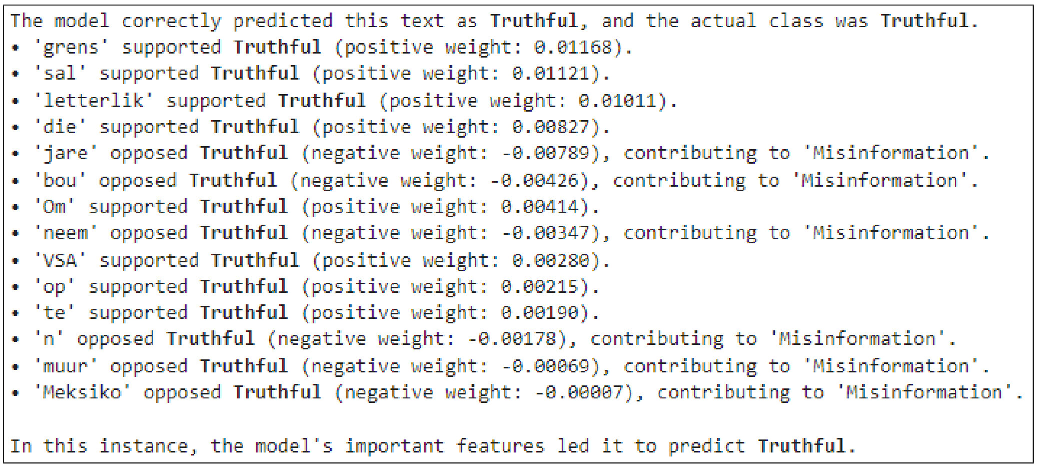
Class	precision	recall	f1-score	support
0	0.07	0.30	0.12	92
1	0.93	0.69	0.79	1170
accuracy			0.66	1262
macro avg	0.50	0.50	0.46	1262
weighted avg	0.86	0.66	0.74	1262

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.

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 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 3 and Figure 4. 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 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.

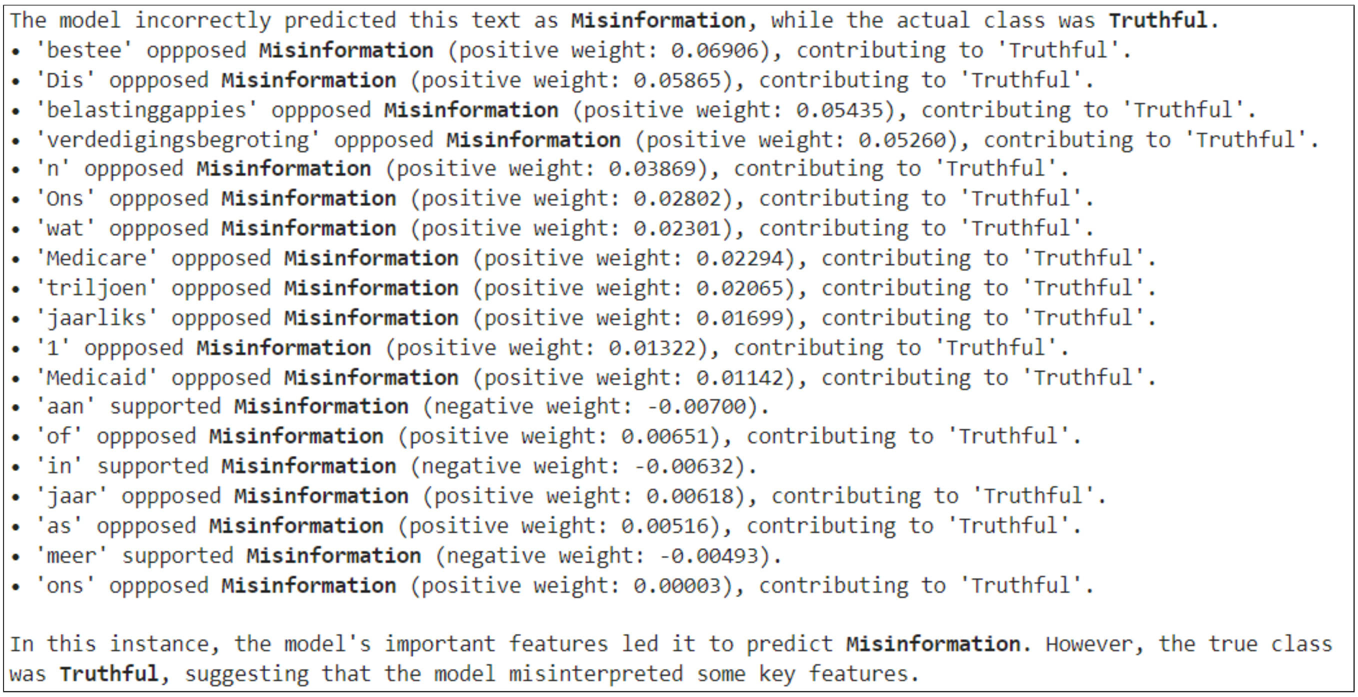


```
The model correctly predicted this text as Truthful, and the actual class was Truthful.
• 'grens' supported Truthful (positive weight: 0.01168).
• 'sal' supported Truthful (positive weight: 0.01121).
• 'letterlik' supported Truthful (positive weight: 0.01011).
• 'die' supported Truthful (positive 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 3: Sentence 0: Human-readable Explanation generated from LIME model's output

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.

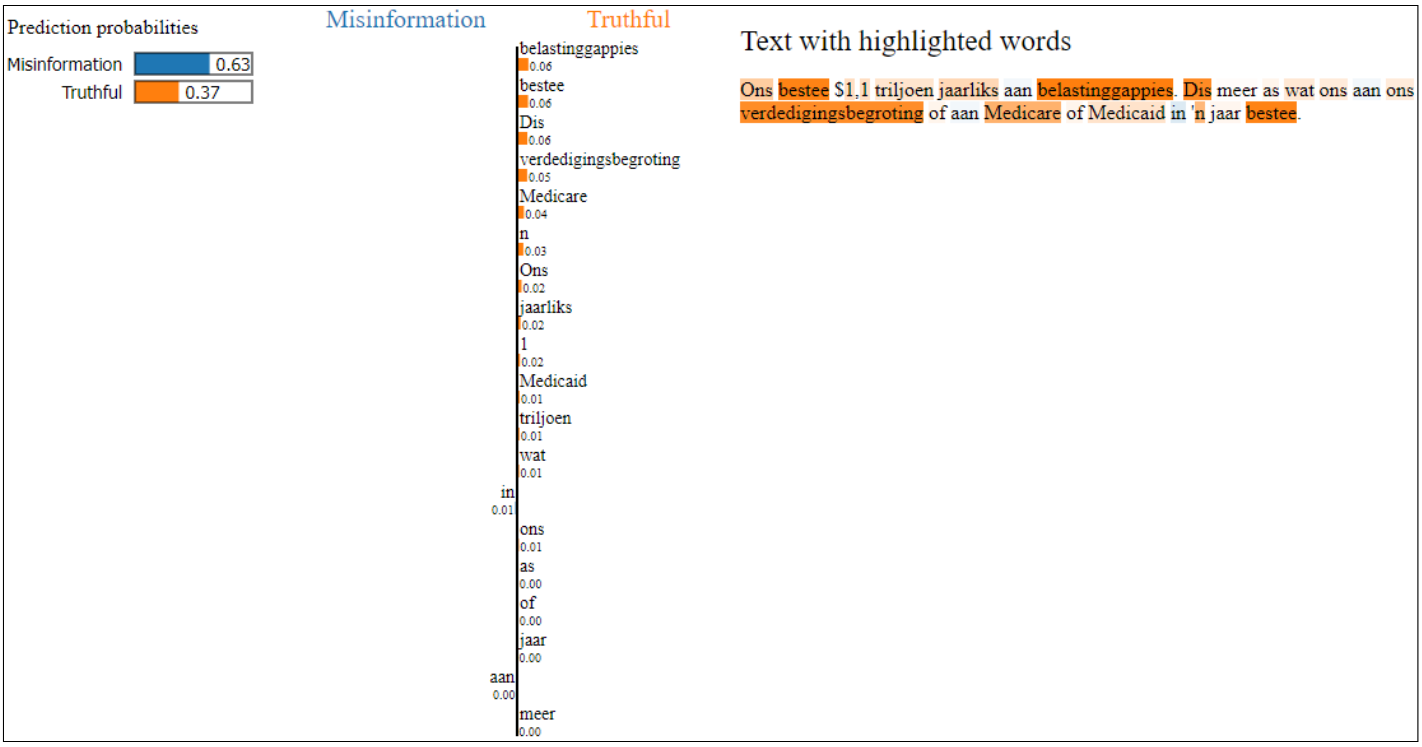


```
The model incorrectly predicted this text as Misinformation, while the actual class was Truthful.
• 'bestee' oppposed Misinformation (positive weight: 0.06906), contributing to 'Truthful'.
• 'Dis' oppposed Misinformation (positive weight: 0.05865), contributing to 'Truthful'.
• 'belastinggappies' oppposed Misinformation (positive weight: 0.05435), contributing to 'Truthful'.
• 'verdedigingsbegroting' oppposed Misinformation (positive weight: 0.05260), contributing to 'Truthful'.
• 'n' oppposed Misinformation (positive weight: 0.03869), contributing to 'Truthful'.
• 'Ons' oppposed Misinformation (positive weight: 0.02802), contributing to 'Truthful'.
• 'wat' oppposed Misinformation (positive weight: 0.02301), contributing to 'Truthful'.
• 'Medicare' oppposed Misinformation (positive weight: 0.02294), contributing to 'Truthful'.
• 'triljoen' oppposed Misinformation (positive weight: 0.02065), contributing to 'Truthful'.
• 'jaarliks' oppposed Misinformation (positive weight: 0.01699), contributing to 'Truthful'.
• '1' oppposed Misinformation (positive weight: 0.01322), contributing to 'Truthful'.
• 'Medicaid' oppposed Misinformation (positive weight: 0.01142), contributing to 'Truthful'.
• 'aan' supported Misinformation (negative weight: -0.00700).
• 'of' oppposed Misinformation (positive weight: 0.00651), contributing to 'Truthful'.
• 'in' supported Misinformation (negative weight: -0.00632).
• 'jaar' oppposed Misinformation (positive weight: 0.00618), contributing to 'Truthful'.
• 'as' oppposed Misinformation (positive weight: 0.00516), contributing to 'Truthful'.
• 'meer' supported Misinformation (negative weight: -0.00493).
• 'ons' oppposed Misinformation (positive weight: 0.00003), contributing to 'Truthful'.

In this instance, the model's important features led it to predict Misinformation. However, the true class
was Truthful, suggesting that the model misinterpreted some key features.
```

Figure 4: 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. 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.



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.

Figure 5: LIME Explanation for Afrikaans Sentence (Test Set Index 170)

6 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) [6], 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.

7 Acknowledgements

I want to express my heartfelt gratitude to my supervisors, [1]Ms. Seani Rananga and Mr. Thapelo Sindane, for their significant support, critical feedback, and continuous support during my research. I am also grateful to my co-supervisor, [2]Prof. Vukosi Marivate, for his helpful recommendations and encouragement. I also wish to thank the [3]Data Science For Social Impact (DSFSI) team for the opportunity to be a part of such a meaningful research community.

Furthermore, I acknowledge the use of Google Gemini for generating Afrikaans sentences by translating English text, which supported the translation tasks involved in this study.

References

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

[2] A. Magueresse, V. Carles, and E. Heetderks, *Low-resource languages: A review of past work and future challenges*, 2020. arXiv: 2006.07264 [cs.CL].

[3] 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, *et al.*, Eds., Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 4543–4549. [Online]. Available: https://aclanthology.org/L16-1720.

[4] 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.

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

[6] G. Joshi, A. Srivastava, B. Yagnik, *et al.*, "Explainable misinformation detection across multiple social media platforms," *IEEE Access*, vol. 11, pp. 23 634–23 646, 2023. DOI: 10.1109/ACCESS.2023.3251892.

[7] W. Y. Wang, *"liar, liar pants on fire": A new benchmark dataset for fake news detection*, 2017. arXiv: 1705.00648 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1705.00648.

[8] J. Tiedemann, M. Aulamo, D. Bakshandaeva, *et al.*, "Democratizing neural machine translation with OPUS-MT," *Language Resources and Evaluation*, no. 58, pp. 713–755, 2023, ISSN: 1574-0218. DOI: 10.1007/s10579-023-09704-w.

[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. arXiv: 2007.01852 [cs.CL].

[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. arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.

[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, ISSN: 1999-5903. DOI: 10.3390/fi16090322. [Online]. Available: https://www.mdpi.com/1999-5903/16/9/322.

[1] Ms. Seani Rananga [2] Prof. Vukosi Marivate [3] Data Science For Social Impact (DSFSI)