Sentiment Analysis for the African Language Twi: Translation-based vs. End-to-End Approaches

Linrui Zhang, Belinda Copus
University of Central Missouri
W.C. Morris 222, Warrensburg, Missouri
{lzhang, copus}@ucmo.edu

Abstract

This paper presents our approach to sentiment analysis for Twi, a low-resource African language. We developed two types of systems: a translation-based system and an end-to-end system. These systems leverage various popular large-scale language models (LLMs), such as BERT and its multilingual variants, and their performances were compared. Our evaluation focused on the accuracy and robustness of these systems in identifying sentiments within Twi text. We also explored the challenges associated with low-resource languages, including limited annotated datasets and the need for effective cross-lingual transfer. The results highlight the potential of end-to-end multilingual LLMs for low-resource languages while emphasizing the importance of translation quality in translation-based approaches to sentiment analysis tasks. Additionally, we provide insights into the practical implications of our findings for future research.

Task Description and Motivation

SemEval-2023 Task 12 (Muhammad et al. 2023) presents a sentiment analysis task for low-resource African languages. In this task, tweets in 14 different African languages are provided, and participants are required to identify the sentiment of the tweets as positive, negative, or neutral. For example, a tweet in Twi language, "nse kofi taylor s3 ne gyimie nu 3ka s3n (translation: Someone tell Kofi Taylor to stop his foolishness.)" is classified as negative.

When handling low-resource languages, most participating teams (Akrah and Pedersen 2023) (Wang et al. 2023) (Benlahbib and Boumhidi 2023) (El Mahdaouy et al. 2023) opted to use pre-trained multilingual LLMs, such as mBERT (Devlin et al. 2018), AfriBERTa (Ogueji, Zhu, and Lin 2021), and Afro-xlmr (Alabi et al. 2022). However, many of these multilingual models lack extensive training corpora or sufficient digital resources compared to English language models. For example, TwiBERT (Akrah and Pedersen 2023), a pre-trained language model specifically developed for the Twi language, was trained on a relatively small dataset of approximately 5 MB, sourced exclusively from the Bible domain. In contrast, the training corpus for the basic English BERT model (Devlin et al. 2018) consisted of around 16 GB of data from a variety of domains. As data is the cornerstone of deep learning, these multilingual models, with fewer training resources, are prone to delivering suboptimal results for resource-scarce languages.

Given that the size, quality, and diversity of training data significantly influence a model's ability to understand and generate language—and with English being the most widely accessible language globally—our team proposed a novel approach. This method involves first translating low-resource languages into English using OpenAI (ChatGPT) API (Brown et al. 2020) and then processing them using pre-trained English LLMs. By leveraging the extensive and diverse English corpora available, pre-trained English LLMs have the potential to achieve superior results. However, this approach presents a notable challenge: the performance of the English models depends heavily on the accuracy of the translation, especially for sentiment analysis tasks. For example, certain sentiments or nuances unique to the original language may lose their meaning when translated literally into another language.

This study investigates the effectiveness of two research approaches for sentiment analysis in low-resource languages: end-to-end multilingual models trained on limited data versus well-trained English models that risk semantic loss during translation. The focus is on Twi, a widely spoken language in Ghana, West Africa. We conducted experiments using various existing multilingual and English large-scale language models to assess their ability to accurately identify sentiment in Twi text. The findings provide valuable insights into the influence of translation quality on the performance of English LLMs and the ability of multilingual LLMs to generalize across languages.

Methodology

Current state-of-the-art approaches to sentiment classification rely on pre-trained large language models such as BERT (Devlin et al. 2018) and RoBERTa (Liu et al. 2019). When working with languages other than English, their multilingual variants are typically utilized.

In this paper, we experimented with seven multilingual LLMs, including TwiBERT (Akrah and Pedersen 2023), mBERT (Devlin et al. 2018), AfriBERTa (Ogueji, Zhu,

English LLMs	F1	Multilingual LLMs	F1
BERT	57.38%	TwiBERT	62.17%
RoBERTa	58.24%	mBERT	64.81%
Deberta	57.39%	AriBERTa-base(large)	62.07%(62.8%)
BERTweet	59.09%	Afro-XLMR-base(large)	62.59%(65.33%)
XLNet	56.74%	XLM-R-base(large)	65.65%(62.47%)
Average	57.73%	Average	64.35%

Table 1: The performance of the English LLMs and multilingual LLMs in the sentiment analysis task for the Twi language

and Lin 2021), AfroXLMR-base (large), and XLM-R-base (large) (Alabi et al. 2022), applying them directly to the Twi language. For the translation-based approach, we utilized the OpenAI (ChatGPT) API (Brown et al. 2020) to translate Twi into English and processed the translations using five English LLMs: BERT-base (Devlin et al. 2018), RoBERTa (Liu et al. 2019), DeBERTa-v3-base (He, Gao, and Chen), XLNet (Yang 2019), and BERTweet (Nguyen, Vu, and Nguyen 2020). Here, we selected two representative LLMs and briefly introduced their architectures and training procedures.

RoBERTa is a transformer model developed by Facebook AI in 2019. It is based on BERT but improves upon its pretraining methodology to enhance performance. The model consists of 12 transformer layers with a total of 125 million parameters. It was trained on 160 GB of English text—10 times more than BERT—enabling it to capture a much richer and broader linguistic understanding.

AfriBERT is a multilingual variant of BERT with approximately 111 million parameters. It is pre-trained on 11 African languages, including Amharic and Hausa (note: Twi is not one of the 11 languages), using data from the news articles domain. The model is trained on a relatively small dataset of less than 1 GB in size.

We adopted the implementation of the LLMs from the HuggingFace Model Hub (Wolf 2019). We fine-tuned them for a maximum of 10 epochs, employing an early stopping callback. The experiments were conducted using a Google Colab T4 GPU.

Experimental Results and Error Analysis

Table 1 illustrates the performance of the two research approaches in sentiment analysis for Twi. The evaluation metric used is the macro-averaged F1 score, which was the original assessment method in SemEval-2023 Task 12. From the results, we observe that, on average, the multilingual LLMs outperform the English LLMs by approximately 6% in F1 score. Below are our interpretations and analyses:

1. The experimental results confirm our initial concern that sentiments can be lost or even changed when translating into a different language. For instance, the following two sentences could both be translations of the same original sentence: (1) He takes his time to carefully analyze every situation before acting. (2) He hesitates too much, wasting time overthinking every situation. These two sentences convey the same literal meaning but express completely different sentiments.

2. Twi is a resource-scarce language with very limited linguistic resources. The OpenAI translation model is data-driven, and the lack of training resources significantly reduces its translation accuracy. To test this, we conducted experiments with Kinyarwanda, another language provided by SemEval 2023 Task 12, which is a more widely spoken African language. We found that the translation quality of OpenAI API improved significantly, and the English models achieved results very comparable to the multilingual models.

3. Due to the protection mechanism, ChatGPT is unable to provide translations if a Twi tweet is highly offensive or unethical. However, such tweets often contain strong sentiments. Removing them from the training and testing corpora can significantly impact the performance of English models.

4. In many multilingual models, such as AfriBERTa and mBERT, although Twi is not included in their training corpora, its linguistic cousins are used during pre-training. This allows the multilingual model to partially capture the linguistic understanding of Twi. Furthermore, according to the research of (Pfeiffer et al. 2020), multilingual LLMs are capable of delivering superior performance even when working with languages not encountered during pre-training. This reasoning further supports the superior performance of multilingual models compared to English models.

Conclusion and Future Work

In this paper, we developed sentiment analysis systems for the low-resource African language Twi using two research approaches and compared their performances. Experimental results demonstrate that end-to-end multilingual approach consistently outperforms the translation-based approach, even though they are trained on significantly less data and cover fewer domains. However, we believe that the translation-based approach holds substantial potential if translation quality can be improved. Specifically, we are interested in investigating whether using gold-standard translations, such as manually translated Twi sentences, could enhance the performance of the English models. Furthermore, we plan to extend our research in the future to explore other low-resource African languages included in SemEval-2023 Task 12.

References

Akrah, S., and Pedersen, T. 2023. Duluthnlp at semeval-2023 task 12: Afrisenti-semeval: Sentiment analysis for low-resource african languages using twitter dataset. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 1697–1701.

Alabi, J. O.; Adelani, D. I.; Mosbach, M.; and Klakow, D. 2022. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, 4336–4349.

Benlahbib, A., and Boumhidi, A. 2023. Nlp-lisac at semeval-2023 task 12: Sentiment analysis for tweets expressed in african languages via transformer-based models. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 199–204.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

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

El Mahdaouy, A.; Alami, H.; Lamsiyah, S.; and Berrada, I. 2023. Um6p at semeval-2023 task 12: Out-of-distribution generalization method for african languages sentiment analysis. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

He, P.; Gao, J.; and Chen, W. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR* abs/1907.11692.

Muhammad, S. H.; Abdulmumin, I.; Yimam, S. M.; Adelani, D. I.; Ahmad, I. S.; Ousidhoum, N.; Ayele, A. A.; Mohammad, S.; Beloucif, M.; and Ruder, S. 2023. Semeval-2023 task 12: Sentiment analysis for african languages (afrisenti-semeval). In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Nguyen, D. Q.; Vu, T.; and Nguyen, A.-T. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9–14.

Ogueji, K.; Zhu, Y.; and Lin, J. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, 116–126.

Pfeiffer, J.; Vulić, I.; Gurevych, I.; and Ruder, S. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7654–7673.

Wang, M.; Adel, H.; Lange, L.; Strötgen, J.; and Schütze, H. 2023. Nlnde at semeval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 488–497.

Wolf, T. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yang, Z. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.