Sentiment Analysis and Language Models for Kwanyama

Ndapa Nakashole[1,2]
[1]University of California, San Diego
[2]Okalai AI
ndapa@ucsd.edu

Abstract

Kwanyama is related to Swahili, Zulu, and the more than 300 other languages in the Bantu family. Yet, unlike its better-known relatives, it remains almost entirely absent from modern Natural Language Processing (NLP). We bring Kwanyama into the LLM era of NLP through two key contributions. First, we introduce OKASENTIMENT, the first sentiment-labeled dataset for Kwanyama, Unlike prior African sentiment corpora that rely primarily on social media, OKASENTIMENT is grounded in an offline, culturally relevant domain: reviews of domestic labor relationships. The dataset is annotated by over 40 native speakers, under expert supervision, with careful quality control. Second, we present OKALM, the first language models for Kwanyama (1B, 3B, and 8B parameters), obtained by continued pretraining of LLaMA-3 checkpoints on a curated Kwanyama corpus. Together, OKASENTIMENT and OKALM bring a left-behind language into the landscape of modern NLP, providing its first benchmark and language models.

Keywords: low-resource sentiment analysis, language resources, African NLP, Kwanyama, Bantu languages, language models

1. Introduction

In their study of 2,485 languages in which they assess resource inclusion in NLP, Joshi et al. (2020) found that over 88% (2,191) were"Left-Behind": suffering from both labeled and unlabeled data scarcity. The authors concluded that it would be a "monumentous, probably impossible effort to lift them up in the digital space." Yet, with the optimism of initiatives such as No Language Left Behind (NLLB) (Goyal et al., 2022), we take a concrete step toward that goal by bringing Kwanyama into modern NLP.

Kwanyama is a Bantu language related to Swahili, Zulu, and Lingala (Maho, 1999; Guthrie, 1967, 1935). It is the language of Namibia's largest ethnic group (Census, 2023), with a significant community of speakers in southern Angola (Figure 1), totaling roughly 1.5 million people. Despite this large speaker base, Kwanyama lacks even the most basic NLP infrastructure—no pretrained embeddings (Mikolov et al., 2013; Grave et al., 2018), no contextualized language models (Vaswani et al., 2017; Peters et al., 2018; Devlin et al., 2019), and no representation in multilingual LLM benchmarks (Conneau et al., 2020; Xue et al., 2021; Muennighoff et al., 2022). Simply put, there are no labeled or unlabeled corpora available for computational use.

We take the first steps toward bridging this gap through two new resources. First, we introduce OKASENTIMENT, the first sentiment analysis dataset in Kwanyama, grounded in the domain of domestic labor relationships. While sentiment is among NLP's most widely studied tasks, existing datasets, whether in English or African languages, typically cover reviews of movies, restaurants, or online shopping products. In contrast, OKASENTIMENT reflects a socially central and culturally grounded context: everyday judgments of trust, reliability, and fairness in rural employment relationships, such as between families and domestic workers (e.g., housekeepers, goat herders). This focus contrasts with prior African sentiment corpora, which rely primarily on social media data, especially Twitter (now X.COM) (El Abdouli et al., 2017; Moudjari et al., 2020; Yimam et al., 2020; Martin et al., 2021; Muhammad et al., 2022, 2023).

Second, we develop OKALM, the first open language models for Kwanyama, with 1B, 3B,

Figure 1: Geographic distribution of Kwanyama speakers across northern Namibia and southern Angola. The map also highlights Ndonga, a closely related dialect frequently used in code-switching and reflected in portions of the OKASENTIMENT dataset.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3031–3043
11–16 May 2026.

and 8B parameters. These models are obtained through continued pretraining of the LLaMA-3 family (Dubey et al., 2024) on a curated Kwanyama corpus.

The prefix *oka-* in our resource names is a diminutive in Kwanyama, chosen to reflect both the modest scale of these initial contributions and the early stage of digital infrastructure for the language.

Contributions. Our work makes the following contributions:

1. **OKALM**: the first open language models for Kwanyama (1B, 3B, and 8B parameters; see §2), released publicly to support continued research and development in this underrepresented language.

2. **OKASENTIMENT**: the first sentiment dataset for Kwanyama, focused on domestic labor, annotated by over 40 native speakers under expert supervision, validated via i)perplexity analysis, ii) lexical statistics, and iii) expert evaluation (§3, §4).

3. **A benchmark of 14 models**: spanning both open and proprietary systems, evaluated on OKASENTIMENT, highlighting the challenges of low-resource sentiment classification and the promise of targeted adaptation (§5).

The resources in this work are released openly at https://github.com/okalai-ai/okaResource.

2. Language Models for Kwanyama

Language models underpin nearly every modern NLP capability, from computing perplexity to generating text and fine-tuning downstream classifiers on contextualized representations. To enable these capabilities for Kwanyama, we develop OKALM, the first open language models for the language.

We continue pretraining from LLaMA-3 checkpoints (Dubey et al., 2024) to produce three open-weight Kwanyama language models (1B, 3B, 8B), trained on a 6.58M-token corpus derived primarily from religious and government texts. Our training follows a one-epoch regime, following insights from scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022). Interestingly, scaling laws for repeated data (Muennighoff et al., 2023) did not hold in our setting: training beyond one epoch led to overfitting, likely due to the limited corpus size.

We optimize the standard autoregressive language modeling loss: $\mathcal{L}_{\text{next}} = -\log p(t_{i+1} \mid t_1, t_2, \ldots, t_i)$ where t_i represents the i-th token in the input sequence.

Language Modeling Corpus. Our language modeling corpus comprises 6.58 million tokens. As with many low-resource languages, religious texts form a substantial portion of the data. The majority of our data was crawled from *jw.org*, a publicly accessible website maintained by a religious society. This source has also underpinned prior multilingual corpora such as JW300 (Agić and Vulić, 2019), and in part, GloT-500 (Imani et al., 2023). While explicitly religious, the texts span diverse themes, including family, community, health, and everyday life, and thus offers a lexically and thematically diverse sample of the language.

The articles are mainly translations from a source in English. In our case, over 70% of the data originates from this domain, with the remainder consisting of government documents (e.g., legal texts such as the constitution) and a small amount of news content.

To ensure data quality, we applied preprocessing steps aligned with best practices for training LLMs (Soldaini et al., 2024). These included language filtering to remove non-Kwanyama content, normalization of encoding and exclusion of low-quality or noisy text.

For language filtering, we made an exception for some English content where it offered linguistic value. For instance, we retained parallel government documents in English and Kwanyama. We hypothesize that this curated English content may aid cross-lingual alignment and facilitate transfer of semantic knowledge to Kwanyama.

We use the default LLaMA-3 tokenizer; however, subword tokenizers trained on multilingual data often fragment low-resource languages disproportionately, leading to higher tokenization rates and degraded representation quality (Ahia et al., 2023; Zhang et al., 2022; Rust et al., 2021; Muller et al., 2021). Future work may explore customized tokenization for Kwanyama.

Training Setup. Training was conducted using mixed-precision (fp16) on 4×A40 GPUs with a peak learning rate of 1e-5, using a cosine learning rate schedule and 2500 warmup steps. We observed stable convergence across all model sizes, as shown in Figure 2. For all training and experimentation, we used the HuggingFace Transformers library (Wolf et al., 2020), allowing reproducibility and easy deployment of our models.

Carbon Emissions. Despite the small size of our dataset, training language models still incurs environmental costs. Thus, in line with emerging norms around transparency, we report GPU hours and estimated CO_2 emissions in Table 1. We applied a carbon intensity of $0.2\,\mathrm{kgCO_2e/kWh}$, representative of the server region's electricity grid emissions.

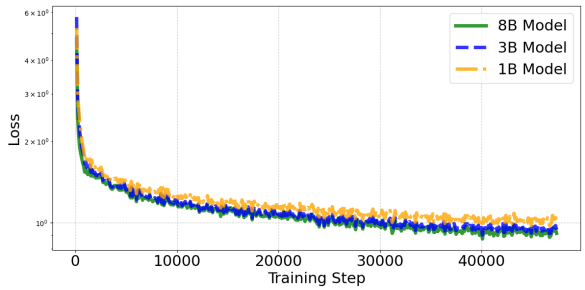


Figure 2: Training loss curves for the three OKALM models. All sizes show stable convergence despite limited training data.

Model Size	GPU Time (Hours)	CO_2 Emissions (kg)
OKALM-1B	2.68	1.07
OKALM-3B	5.77	2.31
OKALM-8B	11.79	4.74
Total	20.24	8.12

Table 1: Estimated carbon emissions from training the OKALM models.

3. OKASENTIMENT Dataset

We now turn to the creation of OKASENTIMENT, the first sentiment-labeled dataset in Kwanyama. Sentiment analysis as an NLP task is technically mature and socially resonant. From product reviews to political speech, it helps machines grasp not just what was said, but how it was felt (Turney, 2002; Pang et al., 2002; Maas et al., 2011; Socher et al., 2013; Liu, 2015; Barbieri et al., 2020; Muhammad et al., 2023).

In the context of Kwanyama, sentiment analysis can help surface trust, reputation, and experience in domestic work relationships. Domestic workers such as goat herders, nannies, and housekeepers often live in their employers' homes. These relationships can go well or poorly. Yet there are no formal systems for accountability or reputation-sharing. Hiring still relies on word-of-mouth or ad-hoc radio adverts. The absence of structured feedback mechanisms can lead to missed opportunities, or worse, exploitation. Our work may support future tools for feedback, hiring, or trust-building in domestic labor networks.

To that end, we introduce OKASENTIMENT, the first sentiment-labeled dataset in Kwanyama. Unlike prior African sentiment resources that focus on social media, OKASENTIMENT is grounded in an offline, real-world domain: reviews of domestic labor relationships.

3.1. Design of the Reviewing Process

We recruited student participants from the two largest public universities in Namibia, leveraging their bilingual fluency and community ties. Initial attempts at collecting open-ended reviews led to low lexical diversity and a strong skew toward positive sentiment. To mitigate this, we started over and adopted a *priming-based annotation strategy*, inspired by NLP work that uses structured prompts to elicit more varied data, such as HotpotQA (Yang et al., 2018). Prompting participants with a sentiment label, polarity, and an aspect—similar to aspect-based sentiment analysis (Pontiki et al., 2014), but grounded in the socioeconomic context of domestic work—encouraged more specific and realistic review generation.

Furthermore, participants were also prompted to assign a name and village to the person being reviewed. While we did not explicitly ask them to draw from personal experience, many likely had familiarity with domestic work relationships.

Annotation Pipeline. Overall, the annotation process involved the following: **Participant Screening.** 65 university students completed a short test assessing Kwanyama writing proficiency. 46 participants passed the screening by a coordinator fluent in Kwanyama, and were approved for full participation.
Quality Control. All submitted reviews were manually inspected by the coordinator. Only one participant's submissions were excluded due to poor quality containing only a few words per review. While participants were explicitly instructed to use standard Kwanyama, some used the Ndonga dialect, which is closely related and mutually intelligible with Kwanyama. Rather than excluding these reviews, we retained them to reflect natural dialectal variation and code-switching practices common in the region.
Compensation. Data collection was conducted over one month. All contributors were compensated fairly, and the total cost was approximately USD 2,500.

3.2. Dataset Statistics

The final OKASENTIMENT dataset contains: 1,127 reviews with a total of 19,927 tokens, with an average review length of 18 words and a maximum of 80 (Figure 3a).

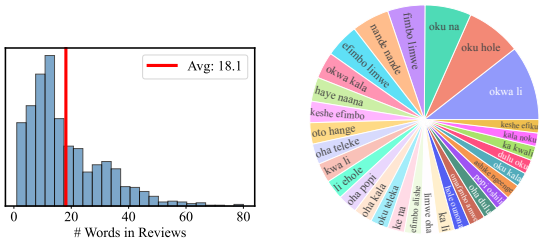
Detailed statistics are shown in Table 2. The 11 annotated aspects are:

honesty, decency, cleanliness, handling of children, cooking, punctuality, respectfulness, payment on time, communication, trustworthiness, and overall recommendation.

Despite our efforts at balance, a mild skew remains due to an initial false start and a budget-constrained early stop in data collection. Still, the label distribution is close to balanced, allowing for

Split	Examples	Tokens	Neg / Pos / Neutral
Train	827	14,885	313 / 313 / 201
Test	300	5,042	100 / 100 / 100
Total	1,127	19,927	413 / 413 / 301

Table 2: Statistics of the OKASENTIMENT dataset.



(a) (b)

Figure 3: **(a)**: review lengths in words. **(b)**: Top 20 most frequent bi-grams in the dataset.

meaningful analysis and effective model training.

4. Review Language Quality

Before evaluating model performance on OKASENTIMENT, we assess the quality of the language itself: is it natural, fluent, and recognizable to native speakers? Many low-resource datasets suffer from data quality issues, establishing linguistic validity is critical to ensuring that models trained on the data will generalize to real-world use cases.

4.1. Perplexity Scores

Perplexity is a standard metric for estimating how well a language model predicts a sample of text, with lower values indicating greater fluency and predictability of a sequence. We compute perplexity using LLaMA-3 8B, BLOOMZ (*bloomz-7b1*) (Muennighoff et al., 2022), a multilingual LLM trained on 46 languages—including African languages related to Kwanyama, such as Swahili—and our three OKALM variants (1B, 3B, 8B).

We evaluate perplexity on four domains: 1) A Kwanyama novel (fiction) 2) News Articles (journalistic nonfiction) 3) The Bible (translated religious text) 4) The OKASENTIMENT reviews (our dataset). The first three domains are included to provide a baseline for comparison, as they contain high quality text in Kwanyama. The news articles are from a local newspaper published after the cutoff date of the training data. The Bible is likely part of the pretraining corpus for many LLMs, including LLaMA-3 and BLOOMZ.

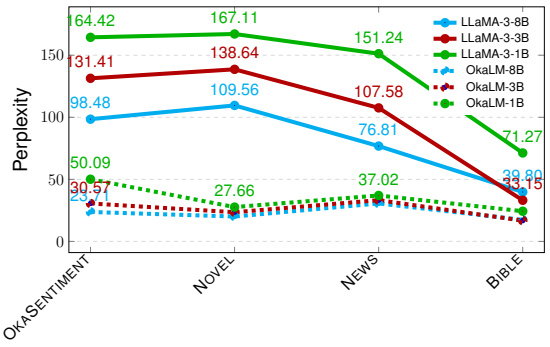


Figure 4: Perplexity (lower is better) of various language models across four Kwanyama domains. OKALM models consistently outperform the untuned LLaMA-3 baselines.

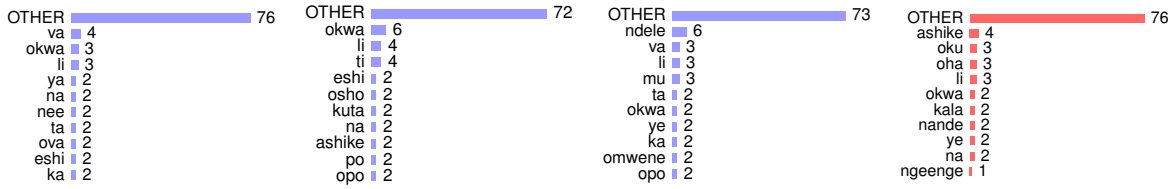
As shown in Figure 4, the OKALM models significantly outperform the untuned LLaMA-3 baseline across all domains. While OKALM-8B achieves the lowest overall perplexity, even the smallest model (OKALM-1B) delivers strong gains over the base models. The Bible shows the lowest baseline perplexity even on the untuned LLaMA-3s, likely due to its presence in the pretraining corpus of most LLMs. BLOOMZ, despite its multilingual training, yielded perplexity scores exceeding 300 on all domains except the Bible. This is likely due to the fact that, while Bantu languages share a common ancestor, their modern forms have diverged in both lexicon and structure (Maho, 1999). For figure legibility, BLOOMZ's numbers are omitted from Figure 4.

Overall, perplexity reductions in our dataset suggest that the reviews are fluent and natural, and that the OKALM models effectively capture the Kwanyama language.

4.2. Lexical Composition

As another dimension of our language quality analysis, we compare the lexical composition of the OKASENTIMENT reviews to that of the three reference corpora mentioned above. We compute unigram frequency distributions for each dataset and visualize the results in Figure 5, which shows the token frequency histograms across all four datasets. The distributions are broadly similar, with frequent use of common Kwanyama function words such as *kwa* (was), *na* (have), and *li* (be). Across all four datasets, the top 10 most frequent tokens account for 24-28% of the dataset, while the remaining tokens fall into the "OTHER" category.

The realistic frequency skew, similar to the reference corpora, is another indicator that the language in OKASENTIMENT is representative of natural Kwanyama usage.



(a) Novel (b) News (c) Bible (d) OKASENTIMENT

Figure 5: Unigram distributions across four Kwanyama corpora. Each bar shows the percentange frequency of each of the top 10 tokens, with remaining tokens grouped as "OTHER." All distributions display typical function word dominance and similar frequency skews, supporting the linguistic naturalness of the OKASENTIMENT reviews.

Review	Sentiment	Target	Aspect
ndeshihafela ounona oku vahole mbela osheshi adalwa kolukwiyu, yee ena nee vali eendenge dihapu, ounona okwe va ikilila unene. ngeenge okwe ku lele okaana koye ohaka kala ashike kafa kaye. okuna ohokwe younona lela.	positive	employee	Children
efimbo limwe ohandi kala nduudite ndahala ndi va shunifile oimaliwa yavo, omolwaashi oilonga yomuumbo haame ashike handiyi longo, ohatu longo ngoo atushe novaneumbo vakwetu ashike aame ngoo handi futwa. ofuto oyali lela yawana no handishi pandula. omnunhu twa kala tuna omalufo okuviyauka pepata? ihashi monika luvali eshi.	positive	employer	Overall

Table 3: Sample reviews selected to be of high quality by a human expert from the OKASENTIMENT dataset, with their sentiment, target, and aspect labels.

4.3. Human Expert Assessment

To complement our automatic evaluations of language quality, we engaged a Kwanyama linguist, an author of a grammar of the language, to assess the OKASENTIMENT test set (§3).

The expert was asked to assign sentiment labels to all 300 test reviews, and to provide qualitative feedback on the fluency and expressiveness of the language used in the reviews. They achieved 90.33% agreement with the gold labels, demonstrating that the task is feasible for fluent native speakers. Among the 29 disagreements, 22 were cases where neutral reviews were misclassified as either positive or negative. Upon inspection, these were often due to the difficulty of writing a truly neutral review.

Beyond labeling, the expert praised the overall quality of the reviews. They highlighted the use of idioms and proverbs with regionally grounded wisdom, and in one case, a biblical reference was used. In total, 38 reviews were marked by the expert as notable either linguistically or socially. Sample reviews highlighted by the expert are shown in Table 3.

To further assess labeling consistency, a second human annotator independently labeled a subset of the reviews. Inter-annotator agreement, measured by Cohen's κ, was 0.73, indicating substantial agreement (Cohen, 1960), showing that the sentiment labels are reliable and consistent across different annotators.

4.4. Summary of Language Quality

Together, the low perplexity, realistic lexical distributions, and expert validation confirm that the OKASENTIMENT reviews are representative of natural, contemporary Kwanyama.

5. Experimental Study and Analysis

How well do modern NLP models perform on sentiment analysis in Kwanyama? We evaluate a diverse set of models on OKASENTIMENT, spanning small-scale supervised learners, multilingual pre-trained encoder-only transformers, proprietary LLMs, and a simple discrete feature baseline, summarized in Table 4, as follows:

1. **Supervised Fine-Tuning:** We fine-tune seven transformer models using the training portion of OKASENTIMENT. These include: Our three custom-built OKALM variants (1B, 3B, 8B);

Model	Size	Description
Prompted LLMs (Zero-Shot via API)		
DeepSeek-V3	671B	DeepSeek-AI (2024)
Gemini 2.0 Flash	-	Google (Anil et al., 2023)
GPT-3.5-turbo	175B	OpenAI (Brown et al., 2020)
GPT-4o	-	OpenAI (OpenAI, 2023)
Mistral-Large	123B	Mistral AI (Mistral-AI, 2024)
Claude-3 Sonnet	-	(Anthropic, 2025)
Fine-Tuned LLMs (Supervised)		
OkaLM-1B	1B	This work
OkaLM-3B	3B	This work
OkaLM-8B	8B	This work
AfroXLM-R	550M	Alabi et al. (2022)
AfriBERTa	550M	Ogueji et al. (2021)
XLM-R Large	550M	Conneau et al. (2020)
mBERT	110M	Devlin et al. (2019)
Discrete Feature Model		
Logistic Regression	–	(Cox, 1958)

Table 4: Models evaluated on OkaSentiment. Prompted LLMs were accessed via paid APIs.

and four multilingual encoder-only transformers: **AfroXLM-R** (Alabi et al., 2022), **AfriBERTa** (Ogueji et al., 2021), **XLM-R Large** (Conneau et al., 2020), and **mBERT** (Devlin et al., 2019), all commonly used in African NLP benchmarks.

2. **Prompted LLMs:** Commercial language models, available through (paid) APIs have been widely claimed to have multilingual capabilities (Ahuja et al., 2023; Kasai et al., 2023; Lai et al., 2023). ChatGPT, for example, is said to cover over 90 languages(Ahuja et al., 2023). We therefore evaluate six commercial LLMs: **Gemini 2.0** (Anil et al., 2023), **Claude 3** (Anthropic, 2025), **DeepSeek-V3** (DeepSeek-AI, 2024), **GPT-4o** (OpenAI, 2023), **GPT-3.5** (Brown et al., 2020), and **Mistral-Large** (Mistral-AI, 2024). We prompted in a zero-shot setting, as preliminary experiments with 3- and 5-shot examples using balanced sentiment classes yielded no measurable improvements, therefore we focus on zero-shot evaluation.

3. **Discrete Baseline:** As a non-neural benchmark, and for interpretability, we include a logistic regression classifier (Cox, 1958) trained on TF-IDF features over word n-grams (n=1–3). This is a standard baseline in early sentiment classification studies.

5.1. Main Sentiment Classification Results

Figure 6 presents overall sentiment classification accuracy across all models on the held-out test set. We make the following key observations:

Supervised models outperform prompted LLMs. Fine-tuned transformer models significantly outperform proprietary LLMs accessed via prompting. Both **AfroXLM-R** and **AfriBERTa** reach the highest performance at **76.67%**, consistent with prior evidence of their strength on African languages (Dione et al., 2023). This underscores the value of supervised adaptation, even on modest datasets, when the model has prior exposure to typologically or lexically similar languages.

Discrete features are competitive. Surprisingly, the logistic regression model trained on n-grams achieves **74%** accuracy, nearly matching the best fine-tuned transformers. This suggests that many reviews contain identifiable polarity markers. However, this model is less robust on longer reviews with more subtle sentiment indicators such as negation or sarcasm, which are better captured by deep contextual representations.

OkaLM scaling effects. Among our models, OkaLM-8B reaches **70%**, outperforming OkaLM-3B (**62%**) and OkaLM-1B (**55%**). These results align with our perplexity trends (§4) and mirror patterns from prior work on domain-adapted LLaMA variants (Etxaniz et al., 2024), where larger models generally outperform smaller ones, a case not usually well-studied in standard scaling laws literature (Kaplan et al., 2020; Hoffmann et al., 2022). While OkaLM models lag behind stronger multilingual models, their performance confirms the feasibility of language modeling and fine-tuning in highly constrained settings. However, OkaLM models still outperform the prompted LLMs, which we discuss next.

Prompted LLMs yield poor performance. **Gemini 2.0** and **Claude 3** achieve the best performance in this group, with accuracies of **64.33%** and **61.67%**, respectively. However, other models underperform significantly: **GPT-4o**, **DeepSeek-V3**, and **GPT-3.5** all fall below **47%**, and **Mistral-Large** performs at chance level (**33.33%**).

We analyzed model predictions using a confusion matrix, which revealed that **GPT-3.5** and **Mistral-Large** are biased toward negative labels. In contrast, Gemini and Claude produce more balanced predictions and appear to exhibit partial grounding in Kwanyama lexical semantics, an observation supported by further analysis in §4.

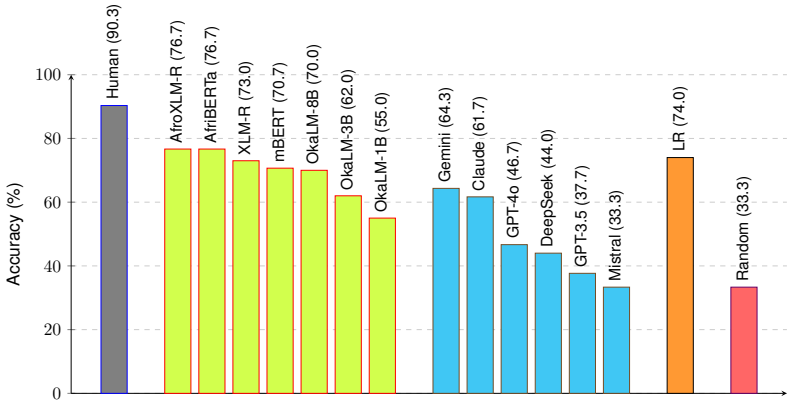


Figure 6: Sentiment classification accuracy on OKASENTIMENT. LR is a logistic regression baseline trained on n-grams. Random is a chance-level baseline. Human performance is from an expert linguist.

Positive Sentiment	**Negative Sentiment**	**Neutral Sentiment**
omulineekelwa (*a trustworthy person*)	keshi omulineekelwa (*not a trustworthy person*)	haye naana (*not really*)
omunyakukwi (*a happy person*)	kaka (*dirty*)	efimbo limwe (*sometimes*)
omunambili (*a peaceful person*)	iha dulu (*incapable*)	ashike (*but*)
okushi (*knows*)	kena oshili (*economical with the truth*)	ngoo (*so-so*)
	onyanya (*harsh*)	
	kena (*has no*)	

Table 5: Selected top lexical features for each sentiment class based on logistic regression weights, shown in their original Kwanyama form (*English translations in parentheses*).

A gap to human performance remains. Even the strongest models fall well short of the **90.33%** accuracy achieved by a human expert on the same test set (§4). This is an opportunity. Just as benchmarks like SuperGLUE (Wang et al., 2019), MMLU (Hendrycks et al., 2021), and BIG Bench (bench authors, 2023) have spurred progress in general-purpose NLP, we see OKASENTIMENT playing a parallel role for Kwanyama: a challenge that drives both model development and evaluation innovation.

5.2. Analysis of Model Performance

Error Analysis. We examine errors from one of the top models (AfroXLM-R). The *positive* class had the highest error rate (27%), mostly confused with *negative*. The *neutral* class followed (23%), skewing toward *negative*. Negative examples had the lowest error rate (20%).

Common sources of error include dialectal variation coming from the Ndonga dialect, and misclassification of clearly expressed sentiment. Many errors occurred on unambiguous examples (e.g., those that a human expert got right), suggesting models still struggle with basic polarity cues and require stronger lexical grounding in Kwanyama.

Discrete Feature Analysis. We examine the top 20 predictive n-grams from the logistic regression model. Terms like “untrustworthy,” “dirty,” and “harsh” are strong indicators of negative sentiment. Features for positive and neutral classes were also semantically aligned as shown in (Table 5), further reinforcing the dataset’s validity.

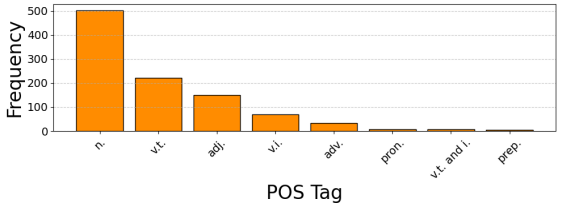


Figure 7: The POS distribution of the words in our bilingual English-Kwanyama lexicon.

Lexical Grounding in Prompted LLMs. To test whether prompted LLMs genuinely understand Kwanyama or are simply guessing sentiment, we asked them to translate 1,000 English words such as “book", "man", and “milk" into Kwanyama. We curated this bilingual lexicon to cover a range of parts of speech (POS) and semantic domains, including common nouns, verbs, adjectives, and function words. The POS distribution of the words in our English-Kwanyama bilingual lexicon is shown in Figure 7.

We collected these words from existing offline resources, specifically, a missionary-era dictio-

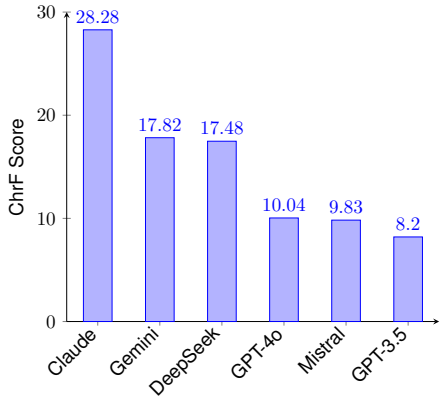


Figure 8: ChrF scores of prompted LLMs on the English–Kwanyama OKALEXICON word translation task.

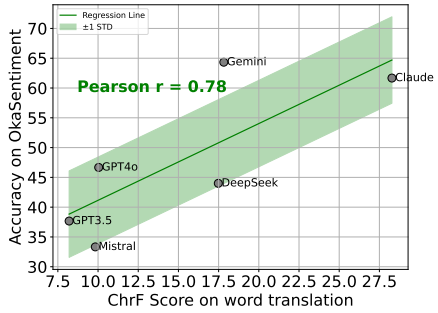


Figure 9: Correlation between ChrF scores from word translation task and sentiment classification accuracy on OKASENTIMENT.

nary (Tobias and Turvey, 1954).

With the exception of **Claude** and **Gemini**, most LLMs responded with what appeared to be a guess, a hallucination, or a word from a random African language. We evaluated the outputs using **chrF** (Popović, 2015), a character-level metric that captures partial word overlap as Kwanyama is an agglutinative language with complex morphology, and thus ChrF is more forgiving than exact match.

As shown in Figure 8, the top-performing LLMs on sentiment, **Claude** and **Gemini**, also achieve the highest chrF scores. We observe a strong correlation (**Pearson** $r = 0.78$) between chrF and sentiment accuracy (Figure 9), suggesting that lexical competence is a bottleneck. This aligns with early sentiment analysis approaches, which often relied on lexicons of positive and negative words (Liu, 2015). Frontier LLMs may similarly benefit from stronger grounding in Kwanyama vocabulary.

6. Related Work

Our work builds sentiment analysis resources and language models, extending these advances to alanguage that is largely absent from NLP, Kwanyama. We discuss related work in two areas: sentiment analysis in African languages, and adapting pretrained models with small corpora.

Sentiment Analysis in African Languages. Most prior sentiment analysis datasets for African languages are drawn from social media, primarily Twitter, and lack domain grounding (El Abdouli et al., 2017; Moudjari et al., 2020; Yimam et al., 2020; Martin et al., 2021). Larger multilingual efforts such as NaijaSenti (Muhammad et al., 2022) and AfriSenti (Muhammad et al., 2023) are similarly based on social media, and though broader in coverage, these efforts still favor higher-resource African languages. In contrast, OKASENTIMENT is built for a previously unrepresented, zero-resource language, with reviews centered on domestic labor, a socioeconomically rooted, offline domain of high local relevance.

Although modest in size, OKASENTIMENT compares favorably with early English sentiment benchmarks, such as the movie review dataset by Pang et al. (2002), which contains 752 negative and 1,301 positive reviews. As OKALM continues to evolve, we anticipate it supporting few-shot learning (Brown et al., 2020) for sentiment and beyond.

Adapting Pretrained Models with Small Corpora. We perform continued pretraining of LLaMA-3 models using full parameter updates on a small-scale corpus. While future work may explore parameter-efficient fine-tuning techniques—such as adapter-based methods (Houlsby et al., 2019; Pfeiffer et al., 2020, 2021) or LoRA (Hu et al., 2022)—recent findings suggest that continued pretraining remains the better option under extreme data constraints, as demonstrated by Ebrahimi and Kann (2021) who showed strong improvements over XLM-R (Conneau et al., 2020) using only New Testament text across 1600+ languages. These insights directly inform our approach.

7. Conclusion

We introduced Kwanyama's first: 1) sentiment analysis dataset, and 2) open language models. These resources begin to address the absence of NLP tools for a community of over one million speakers. Beyond technical results, we emphasized ethical and transparent practices, including fair contributor compensation and carbon reporting. We hope our work not only catalyzes research in Kwanyama NLP, but also inspires work on languages in a similar predicament.

8. Bibliographical References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Anthropic. 2025. Claude opus 4-5. https://www.anthropic.com/.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Namibia Statistics Agency Census. 2023. 2023 population and housing census main report.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232.

DeepSeek-AI. 2024. Deepseek-v3 technical report.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language

understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. 2024. The llama 3 herd of models.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics.

Abdeljalil El Abdouli, Larbi Hassouni, and Houda Anoun. 2017. Sentiment analysis of moroccan tweets using naive bayes algorithm. *International Journal of Computer Science and Information Security (IJCSIS)*, 15(12):191–200.

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for Basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972, Bangkok, Thailand. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Malcolm Guthrie. 1935. Lingala grammar and dictionary. *(Lingala)*.

Malcolm Guthrie. 1967. Comparative bantu. an introduction to comparative linguistics and the prehistory of the bantu languages.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023.

Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Jouni Filip Maho. 1999. *A Comparative Study of Bantu Noun Classes*, volume 13 of *Orientalia et Africana Gothoburgensia*. Acta Universitatis Gothoburgensis, Göteborg. Doctoral dissertation, University of Gothenburg, November 1999.

Gati L Martin, Medard E Mswahili, and Young-Seob Jeong. 2021. Sentiment classification in swahili language using multilingual bert. *arXiv preprint arXiv:2104.09006*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mistral-AI. 2024. Mistral large. https://mistral.ai/news/mistral-large.

Leila Moudjari, Karima Akli-Astouati, and Farah Benamara. 2020. An Algerian corpus and an annotation platform for opinion and emotion analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1202–1210, Marseille, France. European Language Resources Association.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2023. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alipio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Stephen Arthur. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language

models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 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*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 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)*, pages 7654–7673, Online. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

George Wolfe Robert Tobias and Basil HC Turvey. 1954. *English-kwanyama dictionary*. Witswatersrand University Press.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,

Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. 2020. Exploring Amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1048–1060, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shiyue Zhang, Vishrav Chaudhary, Naman Goyal, James Cross, Guillaume Wenzek, Mohit Bansal, and Francisco Guzman. 2022. How robust is neural machine translation to language imbalance in multilingual tokenizer training? In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–116, Orlando, USA. Association for Machine Translation in the Americas.