BERT monolingual lebih baik dari BERT multilingual untuk inferensi bahasa alami di SWAHILI = monolingual BERT is better than multilingual BERT for natural language inference in SWAHILI

Hajra Faki Ali, author

Abstrak

Penelitian ini mengusulkan pengembangan model monolingual untuk Natural Language Inference (NLI) dalam bahasa Swahili untuk mengatasi keterbatasan model multibahasa saat ini. Studi ini melakukan fine-tuning pada model SwahBERT yang sudah dilatih sebelumnya untuk menangkap hubungan semantik dan nuansa kontekstual unik dalam bahasa Swahili. Komponen penting dari penelitian ini adalah pembuatan dataset SwahiliNLI, yang dirancang untuk mencerminkan kompleksitas bahasa Swahili, sehingga menghindari ketergantungan pada teks bahasa Inggris yang diterjemahkan. Selain itu, kinerja model SwahBERT yang telah di-fine-tune dievaluasi menggunakan dataset SwahiliNLI dan XNLI, dan dibandingkan dengan model multibahasa mBERT. Hasilnya menunjukkan bahwa model SwahBERT mengungguli model multibahasa, mencapai tingkat akurasi sebesar 78,78% pada dataset SwahiliNLI dan 73,51% pada dataset XNLI. Model monolingual juga menunjukkan presisi, recall, dan skor F1 yang lebih baik, terutama dalam mengenali pola linguistik dan memprediksi pasangan kalimat. Penelitian ini menekankan pentingnya menggunakan dataset yang dihasilkan secara manual dan model monolingual dalam bahasa dengan sumber daya rendah, memberikan wawasan berharga untuk pengembangan sistem NLI yang lebih efisien dan relevan secara kontekstual, sehingga memajukan pemrosesan bahasa alami untuk bahasa Swahili dan berpotensi menguntungkan bahasa lain yang menghadapi keterbatasan sumber daya serupa. .....This research proposes the development of a monolingual model for Natural Language Inference (NLI) in Swahili to overcome the limitations of current multilingual models. The study fine-tunes the pre-trained SwahBERT model to capture Swahili's unique semantic relationships and contextual nuances. A critical component of this research is the creation of a SwahiliNLI dataset, crafted to reflect the intricacies of the language, thereby avoiding reliance on translated English text. Furthermore, the performance of the fine-tuned SwahBERT model is evaluated using both SwahiliNLI and the XNLI dataset, and compared with the multilingual mBERT model. The results reveal that the SwahBERT model outperforms the multilingual model, achieving an accuracy rate of 78.78% on the SwahiliNLI dataset and 73.51% on the XNLI dataset. The monolingual model also exhibits superior precision, recall, and F1 scores, particularly in recognizing linguistic patterns and predicting sentence pairings. This research underscores the importance of using manually generated datasets and monolingual models in low-resource languages, providing valuable insights for the development of more efficient and contextually relevant NLI systems, thereby advancing natural language processing for Swahili and potentially benefiting other languages facing similar resource constraints.