Abstract/Executive Summary

A lot of work has been done into researching the need for integrating under-resourced languages [1, 2, 3]. Without emphasising again on the need for the development of models that can accurately and efficiently perform sentiment analysis on under-resourced African languages, this paper is our contribution to the discourse and body of knowledge. In this paper, we describe how we would develop a model that performs sentiment analysis in Twi, using task adaptive pretraining (TAPT) and language adaptive pre-training (LAPT), with a focus on the Afrisent SemEval Twi dataset.

Literature Review of Work Done in the Area

In thinking about how to go about building our model, we realised that there were two broad routes we could take when attempting to build a model on the Afrisent Twi dataset: the supervised route and the unsupervised route. The supervised route would involve manually labelling or annotating the train, dev, and test data and then building and training a supervised sentiment analysis machine learning model on the labelled data. Realistically speaking, although it could take some time, this approach was possible, especially given that even large language models and systems such as those developed by OpenAI needed manual labelling of data in the pre-training process [4, 5]. However, as we conducted our literature review, we stumbled across the Afrisenti paper, in which the authors and curators of the Afrisenti SemEval datasets (which included Twi) described some of the challenges faced with data annotation. They stated that they were unable to annotate the Twi dataset because, amongst other issues, a significant portion of them were too ambiguous, making it difficult to accurately categorise sentiment [6].

On further research, we found that some work had been done for other resource-scarce languages (Indonesian and Chinese) which attempted to circumvent the need for labels–or a lot of labels. These semi-supervised learning methods showed a lot of promise and could be applied to African languages [7, 8]. Supervised machine learning is currently the backbone of most machine learning processes. However, as stated earlier, although the dataset could still be labelled, we decided to go down the unsupervised machine learning route.

Having decided on the unsupervised machine learning route, we conducted further research into what methods of unsupervised machine learning would be most appropriate for sentence classification or sentiment analysis. We discovered two approaches that stood out to us. One was unsupervised pre-training models, which use autoencoders or language models to learn a latent representation of the sentence that captures its semantic meaning. It does this by learning to recognize which aspects of observable data are relevant and limit noise in data that can be discarded [9]. Once the sentence is encoded into a latent representation, clustering or nearest neighbour techniques are used to group similar sentences together. This can help identify clusters of sentences that are associated with positive or negative sentiment. After the clusters of positive and negative sentences are identified, a supervised learning model can

then be trained to perform sentence classification using the labelled sentences in the identified clusters as the training data.

The second approach was transfer learning. This method involved fine-tuning a pre-trained model on a (usually large) corpus of text in a related language. This fine-tuned model is then used to perform the sentiment analysis.

Further research led us to a second AfriSenti paper [12] in which the authors described a competition that was held which involved developing sentiment analysis models for 14 African languages. Based on the data provided, the winning team for the subtask of developing a monolingual classification system used two unsupervised learning methods that are types of transfer learning called language adaptive pre-training (LAPT) and task adaptive pre-training (TAPT). We finally settled on using the LAPT and TAPT model for our approach because, based on the literature we found, it seemed to be the best unsupervised route, given our use case. TAPT is a type of transfer learning.

Our Approach

To build a model with TAPT that performs sentiment analysis in Twi, our process would be as follows. Please note that most of our understanding of how TAPT works and other terminologies are gotten from [10]:

- Gather the unlabeled text data from our target language, Twi.
- Pretrain a language model on the unlabelled data using LAPT using self-supervised learning techniques such as masked language modelling or next sentence prediction. In this case, we could use one of the models developed by [11], such as ABENA or BARKO.
- Fine-tune the pre-trained model on a labelled dataset in a related language that is annotated with positive or negative sentiment. The pre-trained model is fine-tuned on the related language dataset to learn how to perform the sentiment analysis task in that language. This step is known as cross-lingual transfer learning. This step is useful because as stated in [11] about mBERT containing other “Niger-Congo” languages, it is reasonable to expect that this model contains some knowledge useful for constructing a Twi embedding.
- Fine-tune the model on the specific sentiment analysis task using the unlabelled dataset in the target language, Twi. The model is fine-tuned on the unlabeled dataset to adapt to the specific characteristics of the target language. This step is known as unsupervised domain adaptation.
- Afterward, we would evaluate the model on a separate test dataset and tune the hyperparameters as needed.
- Finally, we would deploy the model to classify the sentiment of new sentences in the target language.

Analysis of Our Approach

After examining the available literature, we believe our approach has the potential to yield good results, especially for under-resourced language.

Strengths

- This approach is effective when labelled data is unavailable in a target language, but labelled data is available in a related language. Given our Afrisenti dataset, it would be much easier to get data for languages related to Twi for our pre-training process. This could also explain why the winning team in [12] was able to produce an F1 score that was 4 points above the AfriSenti baseline.

- Additionally, using unsupervised machine learning can help improve optimization, especially the unsupervised pre-training models. This is as a result of the network starting near a global minimum, implying that there will be a lower training error than if the data was supervised, where the network would most likely start near a local minimum thereby increasing the training error.

- Research from [10] has shown the TAPT → Finetuning →Self-training process can effectively utilise unlabeled data to achieve strong combined gains consistently in sentiment analysis and other NLP areas.

Limitations

- A major limitation of transfer learning and the approach used above is that it requires large data for the pretraining. By modern standards, the Afrisent isn't large. But if the competition winners used it, it is possible for us to use it too. However, we think that, rather than attempting to build a model from scratch, it would be better to use an already existing Twi model, such as those described above.

- Another limitation is bias. Our final model could inherit the biases of the model used for pre-training. This could be a potential cause for concern if not handled properly. For example, as [11] notes, their model, which was trained on the JW1000 dataset, showed significant religious bias. It would be important to try to mitigate such bias.

- Limited transferability: Although our approach has been shown to deliver good results, due to the under-resourced nature of both Twi and its related languages, there might be limited transferability, especially if the pre-trained model is not designed well and overfits the data

Conclusion

In this paper we have described an approach to developing a machine learning model for performing sentiment analysis on the AfriSent Twi dataset. We described our approach in

detail and showed evidence of its effectiveness. We also described possible strengths and limitations of the approach. We believe that, although African languages like Twi are under-resourced, by standing on the shoulders of giants (work that has already been done) and applying their results to our use case, we can practically develop models that accurately perform sentiment analysis and other tasks for the African region.

References

[1] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. *Proceedings of the 1st Workshop on Multilingual Representation Learning* 1, 1 (2021), 116-126. DOI:https://doi.org/10.18653/v1/2021.mrl-1.11

[2] Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)* 60, 1 (March 2022). DOI:https://doi.org/https://doi.org/10.48550/arXiv.2203.08351

[3] Esma Wali, Yan Chen, Christopher Mahoney, Thomas Middleton, Marzieh Babaeianjelodar, Mariama Njie, and Jeanna Matthews. 2020. Is Machine Learning Speaking my Language? A Critical Look at the . *ArXiv*. Retrieved March 28, 2023 from https://arxiv.org/abs/2007.05872

[4] OpenAI. 2022. DALL.E 2 Pre-training Mitigations. (June 2022). Retrieved April 24, 2023 from https://openai.com/research/dall-e-2-pre-training-mitigations

[5] Billy Perrigo. 2023. Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. (January 2023). Retrieved April 24, 2023 from https://time.com/6247678/openai-chatgpt-kenya-workers/

[6] Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa id Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dario Mario Antonio Ali, Davis Davis, 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 Steven Arthur. 2023. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages. Retrieved April 24, 2023 from https://arxiv.org/abs/2302.08956

[7] Yong Ren, Nobuhiro Kaji, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. 2011. Sentiment Classification in Resource-Scarce Languages by using Label Propagation. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, Institute of Digital Enhancement of Cognitive Processing, Waseda University, Singapore, 420 429. Retrieved April 24, 2023 from https://aclanthology.org/Y11-1044

[8] Tuan Anh Le, David Moeljadi, Yasuhide Miura, and Tomoko Ohkuma. 2016. Sentiment Analysis for Low Resource Languages: A Study on Informal Indonesian Tweets. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, The COLING 2016 Organizing Committee, Osaka, Japan, 123 131. Retrieved April 24, 2023 from https://aclanthology.org/W16-5415

[9] 2022. *What are Autoencoders?* Retrieved April 25, 2023 from https://www.youtube.com/watch?v=qiUEgSCyY5o

[10] Shiyang Li, Semih Yavuz, Wenhu Chen, and Xifeng Yan. 2021. Task-adaptive Pre-training and Self-training are Complementary for Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Association for Computational Linguistics, Punta Cana, Dominican Republic, 1006 1015. DOI:https://doi.org/10.18653/v1/2021.findings-emnlp.86

[11] Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah. 2021. Contextual Text Embeddings for Twi. DOI:https://doi.org/10.48550/arXiv.2103.15963

[12] Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa id Ahmad, Nedjma Ousidhoum, Abinew Ayele, Saif M. Mohammad, and Meriem Beloucif. 2023. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). Retrieved April 26, 2023 from http://arxiv.org/abs/2304.06845