Low Resource Language News Classification

Praneeth Chandra, Saurav Chennuri, Sherry Courington, Keanu Nichols, Leo Seoane, Zilu Tang

(Links to our code, code artifacts, and presentation)

Introduction:

Natural language processing (NLP) models for low resource languages is a difficult but important task that has caught the attention of many researchers in the NLP community. Many languages with tens of millions of speakers are heavily underrepresented in databases, despite the fact that the associate communities of these languages could benefit greatly from increased access to automated classification, translation, and other tools. Chichewa, or Nyanja, is a language spoken mainly in Zambia and Malawi with 12 million speakers. However, there are almost no existing language models and very little online documentation for this language. To combat this inequality, a Zindi Data Science Competition was started to explore methods to build a news frame classification model for Chichewa (AI4D-Africa, 2021). One possible use case for this model could be to classify popular news sites in Malawi. We attempted to build a robust news classification model by creating embeddings using the pretrained mT5 machine translated model developed by Google (Xue et. al. 2020), and aligned the data using parallel news articles with contrastive learning. We also tried a variety of data augmentation and feature augmentation techniques along with fitting a multitude of different classifiers to our embedding data.

Datasets:

Our dataset consists of 1435 news articles written in Chichewa with 20 mutually exclusive classes. Each news article contains around 300 white space separated words. It should be noted that the classes are extremely imbalanced, with some classes such as transportation, witchcraft, chiefs, and flooding containing less than 10 articles each in the original training data set. The large majority of the training data belongs to popular news topics such as politics, law, social issues, and religion.

Augmenting Data:

In order to increase the amount of data for classes containing relatively low numbers of articles, we looked into performing web scraping on popular Malawi news sites. Through this data collection technique, we gathered articles for five categories (music, health, flood, wildlife and witchcraft). However, due to the lack of scalability of web scraping, and lack of resources for some categories of news articles (i.e ‘Chiefs’), we stopped our web scraping data early on.
We also utilized NLPAug to augment our dataset (Ma 2019), upsampling lower populated classes. The goal was to follow Fadaee et. al. (2017) which employed a method similar to that used in visual computing called “translation data augmentation”. This method augments the training data by altering existing sentences in the parallel corpus, similar in spirit to the data augmentation approaches in computer vision. This study used a weaker notion of label

preservation that alters both source and target sentences at the same time as long as they remain translations of each other. The paper augmented only low-frequency words; the code used in this study included the entire set. We used Google Translation API to translate the original training data from Chichewa to English. Each English sentence was then passed through NLPAug contextual word embeddings which leverages contextual word embeddings to find similar words for augmentation. After the augmentation, the resulting sentences were translated back to Chichewa and used as additional training data. Lastly, we experimented with mixup (Zhang et al. 2017), a popular data augmentation technique to create new data points by linearly interpolating data points and labels from different classes weighted by beta distribution. This creates additional examples that fill the dataset space, allowing models to draw better decision boundaries.

Augmenting Input:

In addition to augmenting the dataset, we also tried augmenting the features for each news data input. In order to take advantage of online translation services and pretrained multilingual language models, we used Chichewa input, translated English input, and concatenated Chichewa and English embeddings. To take advantage of the length of news articles, we experimented with breaking up each news frame into three-sentence splits with the same labels. During inference time, we would ensemble the prediction from each split.

Methods:

Featurizers:

In order to establish baselines, we used count vectorizer and TFIDFvectorizer to convert sentences into sparse vector embeddings. To explore modern deep learning model performance, we also chose mT5 (small) (Xue et. al. 2020) as our deep learning baseline because it was pretrained on Chichewa already. Despite having been trained in unsupervised Chichewa text, pre-trained multilingual language models such as mT5 are pretrained with disproportionately higher amounts of English text, the Chichewa language embedding space may not be as developed. To improve the Chichewa embedding space by bootstrapping English context, we finetuned an mT5 model by aligning English and Chichewa embedding spaces using the parallel news corpus we created. To obtain the corpus, we subsample 0.15% of the Realnews dataset (Zellers et. al. 2019) and create a parallel dataset through Google. The corpus contains around 50K parallel news articles, which we split into 3-sentence segments. During training, we use the sentence-transformer library (Reimers and Gurevych, 2019), and fine-tune mT5 with symmetrical softmax between Chichewa and English translations embeddings. After 8 epochs of finetuning with Adam 3e-6, 1 downsampling projection layer, we obtained an embedding space that is much more aligned (Figure 1). After alignment, classes seem better separated as well. Given more time, we would also like to investigate whether the embeddings before or after the dense layer gives better feature space.

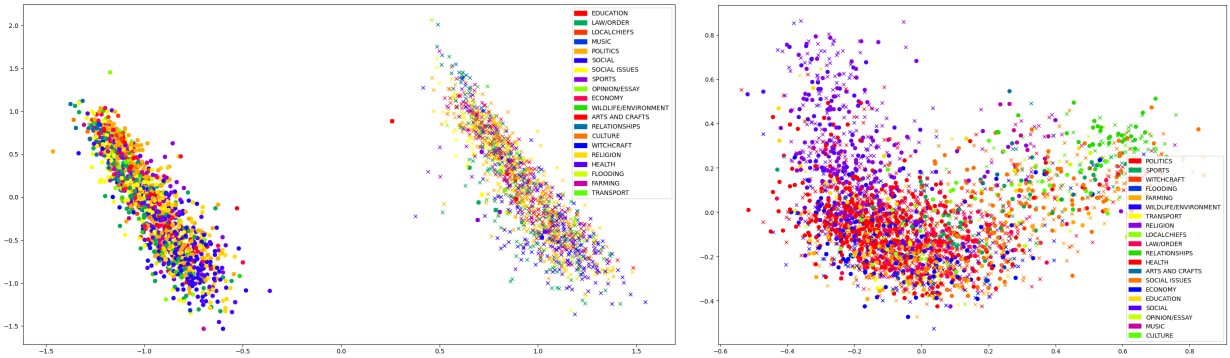

Figure 1: English (marked by cross “X”) and Chichewa (marked by dots “.”) news embedding in training data before (left) and after (right) alignment. Colors represent different news classes.

Classifiers:

For classifiers heads, we tried a representative sample of classifiers from different families. We included gaussian naive bayes, logistic regression (LR), random forest, XGBoost, and a 2-layer MLP.

Results (Experiments):

	Logistic Regression	XGBoost	Random Forest	Naive Bayes	MLP-2
CountVectorizer	**0.47 / 0.65**	0.43 / 0.61	0.32 / 0.55	0.31 / 0.45	0.47 / 0.62
TFIDFVectorizer	0.34 / 0.61	0.44 / 0.61	0.34 / 0.54	0.31 / 0.44	**0.46 / 0.63**
MT5-small embeddings	0.27 / 0.51	0.32 / 0.51	0.29 / 0.49	0.31 / 0.35	**0.45 / 0.54**
Aligned MT-5 small	0.32 / 0.53	0.37 / 0.53	0.36 / 0.55	0.36 / 0.42	**0.43 / 0.57**

Table 1: Cross validation results on original training data only with different featurizers and classifiers. Each cell contains macro f1 scores (left) and accuracy (right).

	mT5 + LR	Aligned mT5 + LR
Chichewa	0.28 / 0.44	0.28 / 0.45
+ Ensemble	0.38 / 0.55	0.34 / 0.52
English	0.33 / 0.50	0.33 / 0.51
+ Ensemble	0.41 / **0.61**	**0.38 / 0.58**
Chichewa + English	0.28 / 0.44	0.28 / 0.45
+ Ensemble	**0.45** / 0.60	0.34 / 0.52

	Aligned MT5 + MLP
Original Data	0.43 / 0.57
+ Contextual Word Substitution*	0.50 / 0.59
+ web scraped*	0.54 / 0.77
+ Mixup	**0.80 / 0.80**

Table 2: Cross validation results on testing different feature inputs to multilingual models (left) and cross validation results on dataset augmentations (right). Each cell contains macro f1 scores (left) and accuracy (right). *Cross validation splits contain augmented data points as well.

Discussion:

MLP is the best classifier. In almost all cases, MLP proves to be the best classifier (Table 1). Being the most expressive model, MLP is able to adapt to dense vectors much better than shallow machine learning classifiers, which puts constraints on the type of boundaries they can draw. However, it does not benefit from ensemble techniques. We hypothesize that the MLP is able to get enough contextual information from its embeddings without ensembling, so the addition of grouping the sentence chunks does not significantly affect model performance.

High-resource languages have better embedding space. This is perhaps not a surprising result but is still an interesting finding. Simply translating Chichewa news into English increases classification accuracy by 6 points (in the same model) (Table 2, Left). This is perhaps due to the much broader context provided, as well as better representation structure from the English embedding space. Linguistically and culturally, machine translation loses so much crucial contextual information that is captured in small syntactic nuances . Even though current NLP research has been improving general model performance, there are still gaps between high and low resource language performance. Our alignment results (Table 2, left) also indicate that simply pulling parallel translations from two languages is not enough to accurately represent the target language. Embedding spaces may be inherently different due to linguistic and cultural differences. Future directions should look into how to balance diversity vs performance, and how to keep linguistically unique elements independent of other languages, and only aligning knowledge that are common to all human experiences.

Model Combination	Test	Train
Mixup Chichewa LR ensemble	9.3%	78%
Chichewa+English concat LR ensemble	8.0%	66%

Table 3: Train and Test accuracies in Ensemble Models

Ensembling can help? Our simple methods of predicting on mutually exclusive splits of news articles improves performance reliably across trials in the cross validation set. However, ensembling techniques gave dismal results in the test set. The Chichewa-English concatenated mT5 small embeddings with ensembling had the highest accuracy on our training set, but did a very poor job of predicting the test set (Table 3). We hypothesize that this is due to the embedding mismatch between English and Chichewa, and the concatenated embeddings only generate noisy, informationless representations that become overfit in training data, and not generalizable to the test set.

Best performing "baseline" model. Our best performing model on the test set is a MLP on top of TFIDF features (61%) **(Table 4**). We hypothesize that the mT5 model is too heavily reliant on English training data, and is unable to generalize to a semantically different language such as Chichewa. Additionally, aligning Chichewa and English destroys local context crucial for

accurate predictions. Although, given more time, we would like to standardize our testing pipeline for more consistent and repeatable results, as well as drawing more conclusions from actual test set performance. In addition, we would like to experiment more with standard data science pipeline, from preprocessing text to different methods in upsampling and downsampling.

Datasource	Test Accuracy
TFIDFVectorizer	**61%**
MT5-small embeddings	56%
Mixup-unaligned	38%

Table 4: Results of the MLP trained on different forms of the text data submitted to the competition leaderboard.

Conclusion:

Although we were not able to produce state of the art results, we observed several important insights into the task of text classification in low resource language. First, we learned that we cannot reliably depend on using web scraping tools or other outside data gathering techniques to bolster our dataset, especially in areas that might not be often written about in the target language. Instead, we should focus on word embeddings and representations until more labeled data can be manually generated. Second, we learned that in areas, where a large amount of text per news frame is present, that ensembling predictions on chunks of data can reliably improve performance. However, more investigations are needed to see how we can improve test performance. Furthermore, the classifier choice is not as important as other aspects of training the model. If the data passed into the classifier is not representative of the true meaning of words and sentences, then the classifiers will have a hard time accurately predicting the labels. We did learn that MLP, with its ability to draw decision boundaries around complex feature spaces, is all around the best model compared to standard ML techniques. However, it does involve a highly manual-intense hyperparameter tuning procedure. Finally, we learned that standard data augmentation techniques might not be enough to improve results, and that only truly generating more data would see improvements. The task of low resource text classification remains a large issue, and native speakers of chichewa would greatly benefit from having native tools that could accurately describe the peculiarities and complexities of the language for different NLP tasks. In the meantime, given the trend of NLP Research, translating Chichewa to English is the best way forward.

Citations:

AI4D-Africa. (2021). *AI4D Malawi News Classification Challenge*. Zindi. Retrieved May 5, 2022, from https://zindi.africa/competitions/ai4d-malawi-news-classification-challenge

Chichewa to English translator. Chichewa to English Translator online. (n.d.). Retrieved May 5, 2022, from https://www.webtran.eu/chichewa_english_translator/

Fadaee, M., Bisazza, A., & Monz, C. (2017). Data augmentation for low-resource neural machine translation. arXiv preprint arXiv:1705.00440.

Hugging Face. (2020, January 16). *Helsinki-NLP/Opus-mt-ny-en · hugging face*. Helsinki-NLP/opus-mt-ny-en. Retrieved May 5, 2022, from https://huggingface.co/Helsinki-NLP/opus-mt-ny-en

Ma, E. (2019). NLP Augmentation. Retrieved May 5, 2022, from https://github.com/makcedward/nlpaug

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *Advances in neural information processing systems*, *32*.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Appendix:

	Google Translate	Helsinki-NLP (Hugging Face, 2020)	Online Translator (Chichewa to English Translator)
DistilGPT2 Perplexity (Radford et. al. 2019)	44 ± 20	78 ± 88	153 ± 85

Table 5: Translation quality between different online servies. In order to determine the most reliable translation to create parallel data for alignment, we used perplexity of Chichewa-English translation target sentence as an approximation of the quality. Google translate won the race by far.