A Languages

In this work, we consider three languages: Hausa, isiXhosa and Yorùbá. These languages are from two language families: Niger-Congo and Afro-Asiatic, according to Ethnologue (Eberhard et al., 2019), where the Niger-Congo family has over 20% of the world languages.

The Hausa language is native to the northern part of Nigeria and the southern part of the Republic of Niger with more than 45 million native speakers (Eberhard et al., 2019). It is the second most spoken language in Africa after Swahili. Hausa is a tonal language, but this is not marked in written text. The language is written in a modified Latin alphabet.

Yorùbá, on the other hand, is native to south-western Nigeria and the Republic of Benin. It has over 35 million native speakers (Eberhard et al., 2019) and is the third most spoken language in Africa. Yorùbá is a tonal language with three tones: low, middle and high. These tones are represented by the grave ("\"), optional macron ("−") and acute ("/") accents respectively. The tones are represented in written texts along with a modified Latin alphabet.

Lastly, we consider isiXhosa, a Bantu language that is native to South Africa and also recognized as one of the official languages in South Africa and Zimbabwe. It is spoken by over 8 million native speakers (Eberhard et al., 2019). isiXhosa is a tonal language, but the tones are not marked in written text. The text is written with the Latin alphabet.

Kann et al. (2020) used as an indicator for a low-resource language the availability of data in the Universal Dependency project (Nivre et al., 2020). The languages we study suit their indicator having less than 10k (Yorùbá) or no data (Hausa, isiXhosa) at the time of writing.

B Datasets

B.1 Existing Datasets

The WikiAnn corpus (Pan et al., 2017) provides NER datasets for 282 languages available on Wikipedia. These are, however, only silver-standard annotations and for Hausa and isiXhosa less than 4k and 1k tokens respectively are provided. The LORELEI project announced the release of NER datasets for several African languages via LDC (Strassel and Tracey, 2016; Tracey et al., 2019) but have not yet done so for Hausa and Yorùbá at the time of writing.

Eiselen and Puttkammer (2014) and Eiselen (2016) created NLP datasets for South African languages. We use the latter's NER dataset for isiXhosa. For the Yorùbá NER dataset (Alabi et al., 2020), we use the authors' split into training, dev and test set of the cased version of their data.[2] For the isiXhosa dataset[3], we use an 80%/10%/10% split following the instructions in (Loubser and Puttkammer, 2020b). The split is based on token-count, splitting only after the end of the sentence (information obtained through personal conversation with the authors). For the fine-tuning of the zero- and few-shot models, the standard CoNLL03 NER (Tjong Kim Sang and De Meulder, 2003) and AG News (Zhang et al., 2015) datasets are used with their existing splits.

B.2 New Datasets

B.2.1 Hausa NER

For the Hausa NER annotation, we collected 250 articles from VOA Hausa[4], 50 articles each from the five pre-defined categories of the news website. The categories are Najeriya (Nigeria), Afirka (Africa), Amurka (USA), Sauran Duniya (the rest of the world) and Kiwon Lafiya (Health). We removed articles with less than 50 tokens which results in 188 news articles (over 37K tokens). We asked two volunteers who are native Hausa speakers to annotate the corpus separately. Each volunteer was supervised by someone with experience in NER annotation. Following the named entity annotation in Yorùbá by Alabi et al. (2020), we annotated PER, ORG, LOC and DATE (dates and times) for Hausa. The annotation was based on the MUC-6 Named Entity Task Definition guide.[5] Comparing the annotations of the volunteers, we observe a conflict for 1302 tokens (out of 4838 tokens) excluding the non-entity words (i.e. words with 'O' labels). One of the annotators was better in annotating DATE, while the other was better in annotating ORG especially for multi-word expressions of entities. We resolved all the conflicts after discussion with one of the volunteers. The split of annotated data of the Yoruba and Hausa NER data

[2] https://github.com/ajesujoba/YorubaTwi-Embedding/tree/master/Yoruba/Yor%C3%B9b%C3%A1-NER

[3] https://repo.sadilar.org/handle/20.500.12185/312

[4] https://www.voahausa.com

[5] https://cs.nyu.edu/faculty/grishman/NEtask20.book_1.html

is 70%/10%/20% for training, validation and test sentences.

B.2.2 Hausa and Yorùbá Text classification

For the topic classification datasets, news titles were collected from VOA Hausa and the BBC Yoruba news website[6]. Two native speakers of the language annotated each dataset. We categorized the Yorùbá news headlines into 7 categories, namely "Nigeria", "Africa", "World", "Entertainment", "Health", "Sport", "Politics". Similarly, we annotated 5 (of the 7) categories for Hausa news headlines, excluding "Sport" and "Entertainment" as there was only a limited number of examples. The "Politics" category in the annotation is only for Nigerian political news headlines. Comparing the two annotators, there was a conflict rate of 7.5% for Hausa and 5.8% for Yorùbá. The total number of news titles after resolving conflicts was 2,917 for Hausa and 1,908 for Yorùbá.

C Word Embeddings

For the RNN models, we make use of word features obtained from Word2Vec embeddings for the Hausa language and FastText embeddings for Yorùbá and isiXhosa languages. We utilize the better quality embeddings recently released by Abdulmumin and Galadanci (2019) and Alabi et al. (2020) for Hausa and Yorùbá respectively instead of the pre-trained embeddings by Facebook that were trained on a smaller and lower quality dataset from Wikipedia. For isiXhosa, we did not find any existing word embeddings, therefore, we trained FastText embeddings from data collected from the I'solezwe[7] news website and the *OPUS*[8] parallel translation website. The corpus size for isiXhosa is 1.4M sentences (around 15M tokens). We trained FastText embeddings for isiXhosa using *Gensim*[9] with the following hyper-parameters: embedding size of 300, context window size of 5, minimum word count of 3, number of negative samples ten and number of iterations 10.

D Distant Supervision

We provide the distantly supervised data for both the existing and new datasets along with the other data.

D.1 Distant supervision for Personal names, Organisation and Locations

We make use of lists of entities to annotate PER, ORG and LOC automatically. In this paper, we extract personal names, organizations and locations from Wikidata as entity lists and assign a corresponding named entity label if tokens from an unlabeled text match an entry in an entity list.

For NER, we use manual heuristics to improve matching. For Yorùbá, a minimum token length of 3 was set for extraction of LOC and PER, while the minimum length for ORG was set to 2. This reduces the false positive rate, e.g. preventing matches with function words like "of". Applying this on the test set, we obtained a precision of 80%, 38% and 28% for LOC, ORG and PER respectively; a recall of 73%, 52% and 14% for LOC, ORG and PER respectively; and an F1-score of 76%, 44% and 19% for LOC, ORG and PER respectively.

For Hausa NER, a minimum token length of 4 was set for extraction of LOC, ORG and PER. Based on these manual heuristics, on the test set, we obtained a precision of 67%, 12% and 47% for LOC, ORG and PER respectively; a recall of 63%, 37% and 56% for LOC, ORG and PER respectively; and an F1-score of 65%, 18% and 51% for LOC, ORG and PER respectively.

D.2 DATE rules for NER

Rules allow us to apply the knowledge of domain experts without the manual effort of labeling each instance. We asked native speakers with knowledge of NLP to write DATE rules for Hausa and Yorùbá. In both languages, date expressions are preceded by date keywords, like "*ranar*" / "*ọjọ́*" (day), "*watan*" / "*oṣù*" (month), and "*shekarar*" / "*ọdụ́*n" (year) in Hausa/Yorùbá. For example, *"18th of December, 2019"* in Hausa / Yorùbá translates to " *ranar 18 ga watan Disamba, shekarar 2019*" / "*ọjọ́ 18 oṣù Ọpẹ̀, ọdún 2019*". The annotation rules are based on these three criteria: (1) A token is a date keyword or follows a date keyword in a sequence. (2) A token is a digit, and (3) other heuristics to capture relative dates and date periods connected by conjunctions e.g between July 2019 and March 2020. Applying these rules result in a precision of 49.30%/51.35%, a recall of 60.61%/79.17% and an F1-score of 54.42%/62.30% on Hausa /Yorùbá test set respectively.

[6] https://www.bbc.com/yoruba

[7] https://www.isolezwelesixhosa.co.za/

[8] http://opus.nlpl.eu/

[9] https://radimrehurek.com/gensim/

D.3 Rules for Topic classification

For the Yorùbá topic classification task, we collected terms that correspond to the different classes into sets. For example, the set for the class Nigeria contains names of agencies and organizations, states and cities in Nigeria. The set for the World class is made up of the name of countries of the world, their capitals and major cities and world affairs related organizations. Names of sporting clubs and sportspeople across the world were used for the Sports class and list of artists and actresses and entertainment-related terms for the Entertainment class. Given a news headline to be annotated, we get the union set of 1- and 2-grams and obtain the intersection with the class dictionaries we have. The class with the highest number of intersecting elements is selected. In the case of a tie, we randomly pick a class out the classes with a tie. Just as we did for Yorùbá, we collected the class-related tokens for the Hausa text classification. We, however, split the classification into two steps, checking some important tokens and using the same approach as we used for Yorùbá. If a headline contains the word *cutar* (disease) , it is classified as Health, if it contains tokens such as *inec*, *zaben*, *pdp*,*apc* (which are all politics related tokens) it is classified as Politics. Furthermore, sentences with any of the following tokens *buhari*, *legas*, *kano*, *kaduna*, *sokoto* are classified as Nigeria while sentences with *afurka*, *kamaru* and *nijar* are classified as Africa. If none of the tokens highlighted above is found, we apply the same approach as we did for the Yorùbá setting, which is majority voting of the intersection set of the news headline with a keyword set for each class. Applying these rules results in a precision of $59.54\%/60.05\%$, a recall of $46.04\%/53.66\%$ and an F1-score of $48.52\%/54.93\%$ on the Hausa /Yorùbá test set respectively.

E Experimental Settings

E.1 General

All experiments were repeated ten times with varying random seeds but with the same data (subsets). We report mean F1 test score and standard error ($\sigma/\sqrt{10}$). For NER, the score was computed following the standard CoNLL approach (Tjong Kim Sang and De Meulder, 2003) using the *seqeval* implementation.[10] Labels are in the BIO2-scheme. For evaluating topic classification, the implementation by *scikit-learn* was used.[11] All models are trained for 50 epochs, and the model that performed best on the (possibly size-reduced) development set is used for evaluation.

E.2 BERT and XLM-RoBERTa

As multilingual transformer models, mBert and XLM-RoBERTa are used, both in the implementation by Wolf et al. (2019). The specific model IDs are *bert-base-multilingual-cased* and *xlm-roberta-base*.[12] For the DistilBERT experiment it is *distilbert-base-multilingual-cased*. As is standard, the last layer (language model head) is replaced with a classification layer (either for sequence or token classification). Models were trained with the Adam optimizer and a learning rate of $5e^{-5}$. Gradient clipping of value 1 is applied. The batch size is 32 for NER and 128 for topic classification. For distant supervision and XLM-RoBERTa on the Hausa topic classification data with 100 or more labeled sentences, we observed convergence issues where the trained model would just predict the majority classes. We, therefore, excluded for this task runs where *on the development set* the class-specific F1 score was 0.0 for two or more classes. The experiments were then repeated with a different seed.

E.3 Other Architectures

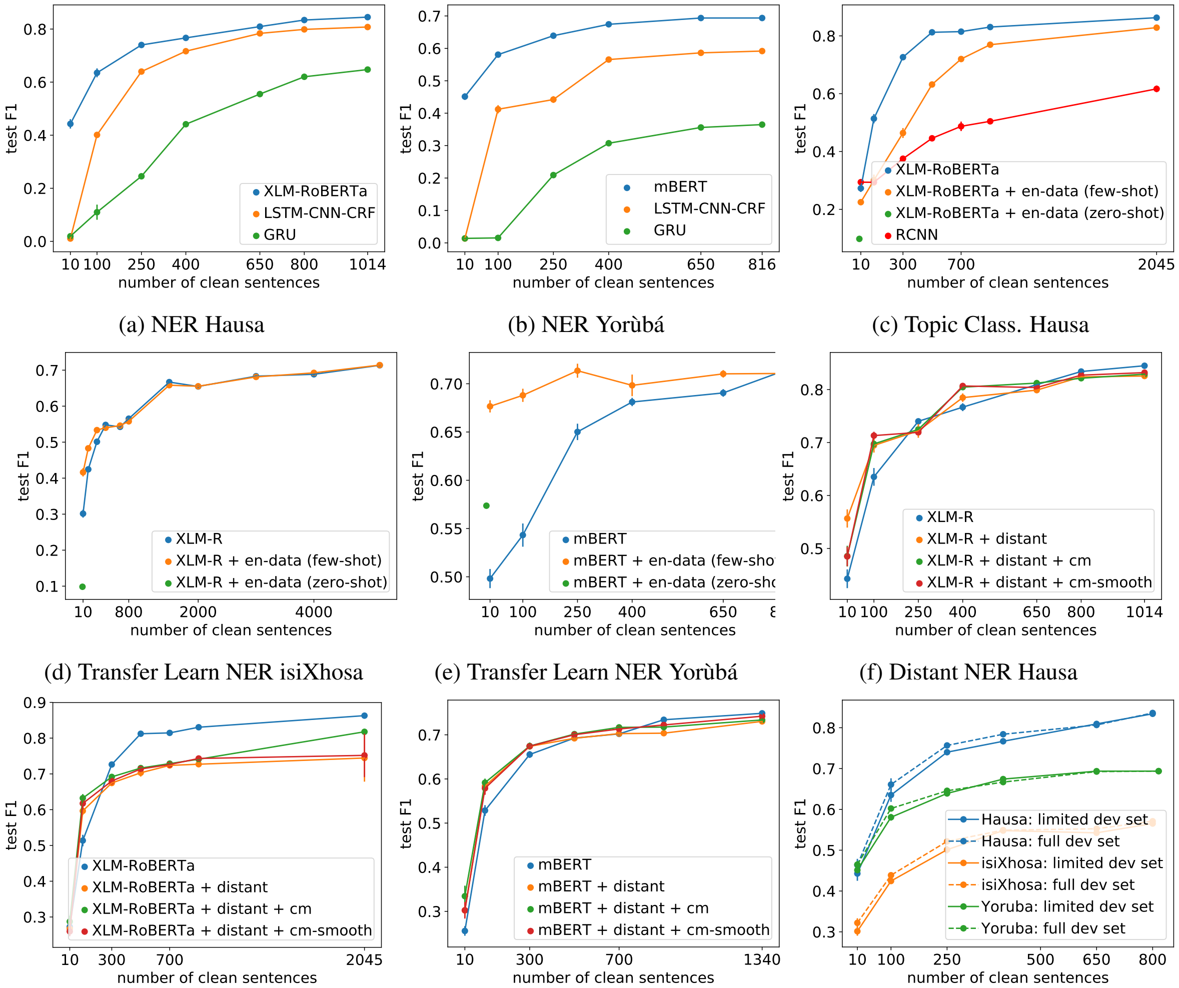
For the GRU and LSTM-CNN-CRF model, we use the implementation by Chernodub et al. (2019) with modifications to support FastText embeddings and the *seqeval* evaluation library. Both model architectures are bidirectional. Dropout of 0.5 is applied. The batch-size is 10 and SGD with a learning rate of 0.01, and a decay of 0.05 and momentum of 0.9 is used. Gradients are clipped with a value of 5. The RNN dimension is 300. For the CNN, the character embedding dimension is 25 with 30 filters and a window-size of 3.

For the topic classification task, we experiment with the RCNN model proposed by (Lai et al., 2015). The hidden size in the Bi-LSTM is 100 for each direction. The linear layer after the Bi-LSTM reduces the dimension to 64. The model is trained for 50 epochs.

[10] https://github.com/chakki-works/seqeval

[11] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

[12] https://huggingface.co/transformers/pretrained_models.html



(a) NER Hausa

(b) NER Yorùbá

(c) Topic Class. Hausa

(d) Transfer Learn NER isiXhosa

(e) Transfer Learn NER Yorùbá

(f) Distant NER Hausa

(g) Distant Topic Class. Hausa

(h) Distant Topic Class. Yorùbá

(i) Development Set NER

Figure 3: Additional plots.

E.4 Transfer Learning

For transfer learning, the model is first fine-tuned on labeled data from a high-resource language. Following (Hu et al., 2020), we use the English CoNLL03 NER dataset (Tjong Kim Sang and De Meulder, 2003) for NER. It consists of ca. 8k training sentences. The model is trained for 50 epochs and the weights of the best epoch according to the development set are taken. The training parameters are the same as before. On the English CoNLL03 test set, the model achieves a performance of 0.90 F1-score. As the Hausa and Yorùbá datasets have slightly different label sets, we only use their intersection, resulting in the labels PER, LOC and ORG and excluding MISC from CoNLL03 and the DATE label from Hausa/Yorùbá. For isiXhosa, the label sets are identical (i.e. also including MISC). After fine-tuning the model on the high-resource data, the model is directly evaluated on the African test set (for zero-shot) or fine-tuned and then evaluated on the African data (for few-shot).

For topic classification, the AG News corpus is used (Zhang et al., 2015). It consists of 120k training sentences. The model is trained for 20 epochs and the weights of the best epoch according to the test set are used. On this set, an F1-score of 0.93 is achieved. The training procedure is the same as above. For the labels, we use the union of the labels of the AG News corpus (Sports, World, Business and Sci/Tech) and the African datasets.

E.5 Label Noise Handling

We use a confusion matrix which is a common approach for handling noisy labels (see, e.g. (Fang and Cohn, 2016; Luo et al., 2017; Lange et al., 2019; Wang et al., 2019)). The confusion matrix models the relationship between the true, clean label of an instance and its corresponding noisy label. When training on noisy instances, the confusion matrix is added to the output of the main model (that usually predicts clean labels) changing the output label distribution from the clean to the noisy one. This allows to then train on noisily labeled instances without a detrimental loss obtained by predicting the true, clean label but having noisy, incorrect labels as targets.

We use the specific approach by Hedderich and Klakow (2018) that was developed to work with small amounts of manually labeled, clean data and a large amount of automatically annotated, noisy labels obtained through distant supervision. To get the confusion matrix of the noise, the distant supervision is applied on the small set of clean training instances. From the resulting pairs of clean and noisy labels, the confusion matrix can be estimated.

In a setting where only a few instances are available, the estimated confusion matrix might not be close to the actual change in the noise distribution. We, therefore, combine it with the smoothing approach by Lv et al. (2020). Each entry of the probabilistic confusion matrix is raised to the power of β and then row-wise normalized.

As studied by Hedderich and Klakow (2018), we do not use the full amount of available, distantly supervised instances in each epoch. Instead, in each epoch, only a randomly selected subset of the size of the clean, manually labeled training data is used to lessen the negative effects of the noisy labels additionally. For smoothing, $\beta = 0.8$ is used as this performed best for Lv et al. (2020).