KINNEWS and KIRNEWS: Benchmarking Cross-Lingual Text Classification for Kinyarwanda and Kirundi

Rubungo Andre Niyongabo[1,2] and **Hong Qu**[1,*] and **Julia Kreutzer**[2] and **Li Huang**[1]

[1]School of Computer Science and Engineering, Computational Intelligence Lab,
University of Electronic Science and Technology of China, China
[2]Masakhane, Africa
niyongabor.andre@std.uestc.edu.cn, hongqu@uestc.edu.cn,
juliakreutzer.jul@gmail.com, lihuang@std.uestc.edu.cn

Abstract

Recent progress in text classification has been focused on high-resource languages such as English and Chinese. For low-resource languages, amongst them most African languages, the lack of well-annotated data and effective preprocessing, is hindering the progress and the transfer of successful methods. In this paper, we introduce two news datasets (KINNEWS and KIRNEWS) for multi-class classification of *news* articles in *Kin*yarwanda and *Kir*undi, two low-resource African languages. The two languages are mutually intelligible, but while Kinyarwanda has been studied in Natural Language Processing (NLP) to some extent, this work constitutes the first study on Kirundi. Along with the datasets, we provide statistics, guidelines for preprocessing, and monolingual and cross-lingual baseline models. Our experiments show that training embeddings on the relatively higher-resourced Kinyarwanda yields successful cross-lingual transfer to Kirundi. In addition, the design of the created datasets allows for a wider use in NLP beyond text classification in future studies, such as representation learning, cross-lingual learning with more distant languages, or as base for new annotations for tasks such as parsing, POS tagging, and NER. The datasets, stopwords, and pre-trained embeddings are publicly available at https://github.com/Andrews2017/KINNEWS-and-KIRNEWS-Corpus.

1 Introduction

The availability of large monolingual and labeled corpora, paired with innovations in neural text processing have led to a rapid improvement of the quality of text classification over the last years.[1] However, the effectiveness of deep-learning-based text classification models depends on the amount of monolingual and labeled data. Low-resource languages are traditionally left behind because of the few available prepared resources for these languages to extract the data from (Joshi et al., 2020). However, nowadays, the increase in internet use in many African developing countries has made access to information easier. This in turn has strengthened the news agencies of those countries to cover many stories in their native languages. For example, *BBC News* now provides online news in Arabic, Amharic, Hausa, Kiswahili, Somali, Oromo, Igbo, Nigerian Pidgin, Tigrigna, Kinyarwanda and Kirundi.[2] This development makes news the most reliable source of data for low-resource languages. We explore this opportunity for the example of Kinyarwanda and Kirundi, two African low-resource Bantu languages, and build news classification benchmarks from online news articles. This has the goal to enable NLP researchers to include Kinyarwanda and Kirundi in the evaluation of novel text classification approaches, and diversify the current NLP landscape.

Kinyarwanda is one of the official languages of Rwanda[3] and belongs to the Niger-Congo language family. According to *The New Times*,[4] it is spoken by approximately 30 million people from four dif-

---

[1]See e.g. https://nlpprogress.com/english/text_classification.html or https://paperswithcode.com/task/text-classification

[2]https://www.bbc.co.uk/ws/languages

[3]https://en.wikipedia.org/wiki/Kinyarwanda

[4]https://www.newtimes.co.rw/section/read/24728

ferent African countries: Rwanda, Uganda, DR Congo, and Tanzania. *Kirundi* is an official language of Burundi, a neighboring country of Rwanda, and it is spoken by at least 9 million people.[5] Kinyarwanda and Kirundi are mutually intelligible. Table 1 shows two example sentences to illustrate the similarity of the languages. The two languages are part of the wider dialect continuum known as Rwanda-Rundi.[6] In this family, there are other four indigenous low-resource languages: *Shubi*,[7] *Hangaza*,[8] *Ha*,[9] and *Vinza*.[10] Among these four languages, *Ha* is also mutually intelligible for Kirundi and Kinyarwanda speakers, while other three are partially mutually intelligible. Developing NLP models for these languages is the goal for future work, because we could not retrieve any written news data for them.

Language	Sentence
Kirundi	*Turafise* **ubwoba ko inyuma y'ikiza** *ca* **coronavirus** *bazohava bakena cane* **kuruta mbere.**
Kinyarwanda	*Dufite* **ubwoba ko inyuma y'ikiza** *cya* **coronavirus** *bashobora gukena cyane* **kuruta mbere.**
English	*We fear that after the coronavirus epidemic, they may become poor than before.*
Kinyarwanda	***Ashushanya inyama*** *n'ibintu* **bikozwe mu masashe n'impapuro.**
Kirundi	***Ashushanya inyama*** *n'ivyamwa* **bikozwe mu masashe n'impapuro.**
English	*He draws meat and things made of sacks and paper.*

Table 1: Two examples of news titles from our datasets that show the similarity level between Kinyarwanda and Kirundi. Same words in both languages for each sentence are shown in bold.

Joshi et al. (2020) classify the state of NLP for both Kinyarwanda and Kirundi as *"Scraping-By"*, which means that they have been mostly excluded from previous NLP research, and require the creation of dedicated resources for future inclusion in NLP research. To this aim, we introduce two datasets KIN-NEWS and KIRNEWS for multi-class text classification in this paper. They consist of the news articles written in Kinyarwanda and Kirundi collected from local news websites and newspapers. KINNEWS samples are annotated using fourteen classes while that of KIRNEWS are annotated using twelve classes based on the agreement of the two annotators for each dataset. We describe a data cleaning pipeline, and we introduce the first ever stopword list for each language for preprocessing purposes. We present word embedding techniques for these two low-resource languages, and evaluate various classic and neural machine learning models. Together with the data, these baselines and preprocessing tools are made publicly available as benchmarks for future studies. In addition, pre-trained embeddings are published to facilitate studies for other NLP tasks on Kinyarwanda and Kirundi.

In the following, we will first discuss previous work on Kinyarwanda and Kirundi and low-resource African languages in general in Section 2, and then describe the dataset creation in Section 3. We then present a range of experiments for text classification on the collected data in Section 4, concluding with an outlook to future work in Section 5.

2 Related Work

Joshi et al. (2020) introduce a taxonomy of languages in NLP that expresses to which degree they have been subject to NLP research until 2020. Many of the roughly 2000 African languages fall under the categories *"Scraping-By"* or *"Left-Behind"*, which means that they have been systematically understudied or ignored in NLP research. As Joshi et al. (2020) discuss, those languages require dedicated, often manual effort to enable NLP research, since even monolingual digital data does not exist or is hard to find on the web. In recent years, NLP research on African languages has been increasing thanks to new datasets being published, such as the *JW300* corpus (Agić and Vulić, 2019), which provides parallel data for 300

---

[5] https://en.wikipedia.org/wiki/Kirundi
[6] https://en.wikipedia.org/wiki/Rwanda-Rundi
[7] https://en.wikipedia.org/wiki/Shubi_language
[8] https://en.wikipedia.org/wiki/Hangaza_language
[9] https://en.wikipedia.org/wiki/Ha_language
[10] https://en.wikipedia.org/wiki/Vinza_language

languages and facilitated machine translation research for many low-resource languages (Tiedemann and Thottingal, 2020), amongst them many African languages (∀ et al., 2020b; ∀ et al., 2020a) which have not been subject to machine translation before.

Beyond the multilingual *JW300* corpus, there have been a few works have been done for creating new datasets for individual African languages. For example, Emezue and Dossou (2020) introduced the FFR project for creating a corpus of Fon-French (FFR) parallel sentences. Closer to our work, Marivate et al. (2020) created news classification benchmarks for Setswana and Sepedi, two Bantu languages of South Africa. Different to our work, their corpus is limited to headlines, while we provide headlines and the full articles. The size of our dataset is also several magnitudes larger since we include data from more sources and spent large efforts on expanding an initial set of news sources.

While there is practically no NLP research on Kirundi, there are a few recent studies on Kinyarwanda for the tasks of Morphological Analysis (Muhirwe, 2009), Part-of-Speech (POS) tagging (Garrette and Baldridge, 2013; Garrette et al., 2013; Fang and Cohn, 2016; Duong et al., 2014; Cardenas et al., 2019), Parsing (Sun et al., 2014; Mielens et al., 2015), Automated Speech Recognition (Dalmia et al., 2018), Language Modeling (Andreas, 2020), and Name Entity Recognition (Rijhwani et al., 2020). Most of these works are largely based on a single Kinyarwanda dataset created by Garrette and Baldridge (2013). This dataset contains transcripts of testimonies by survivors of the Rwandan genocide, was provided by the Kigali Genocide Memorial Center, and contains 90 annotated sentences with fourteen distinct POS tags. However, it is not suitable for text classification, and to the best of our knowledge, there are no publicly available datasets for Kinyarwanda and Kirundi text classification, which is the gap that this paper is addressing. These works have also focused on either word alignment or monolingual approaches, and did not explore a cross-lingual approach.

We hope that the publication of our benchmarks will inspire the creation of similar datasets. As a result, this would allow the inclusion of more African low-resource languages in cross-lingual studies and benchmarks such as *XTREME* (Hu et al., 2020), a multi-task benchmark for the evaluation of cross-lingual generalization of multilingual representations across 40 languages, which already include higher-resourced African languages like Afrikaans and Swahili. For past efforts of multi-lingual studies like *XTREME*, one guiding factor for language selection has been the size of the Wikipedia in the respective languages. The number of Wikipedia articles in local languages is often interpreted as a measure for digital maturity and a pragmatic estimator for the success of un- or self-supervised NLP methods, but this ignores societal and human factors that (cyclically) influence the activity of Wikipedia editor communities. In this work, we want to showcase the impact of manual collection of data sources beyond Wikipedia. The number of news articles that we could retrieve for Kirundi and Kinyarwanda exceeds the number of available Wikipedia articles by far (616 and 1828 Wikipedia articles respectively).[11]

3   Dataset Creation

In this section, we first describe the process for data collection, then the annotation, and finally our data cleaning pipeline for KINNEWS and KIRNEWS. In general, the copyright for the published news articles still remains with the original authors or publishers. Our work can be seen as an additional pre-processing and modeling pipeline, and annotation layer on top.

3.1   Collection Process

KINNEWS   KINNEWS is collected from fifteen news websites and five newspapers from Rwanda. An initial seed of news sources was retrieved from two websites which list newspapers from Rwanda.[12] These lists also include Rwandan news sources that publish their news in other languages such as French and English, so we select those which publish in Kinyarwanda only. Additionally, we expand the initial seed through Google Search by searching for manually selected Kinyarwanda key words and phrases such as *"Iterambere ry'umugore mu Rwanda"* (*"Women's development in Rwanda"*) to list all news

---

[11]`https://meta.wikimedia.org/wiki/List_of_Wikipedias` as of June 29, 2020.
[12]`https://www.w3newspapers.com/rwanda/` and `http://www.abyznewslinks.com/rwand.htm`

sources that have published the searched key phrase or related news. These were used to expand the list of news sources that publish their news in Kinyarwanda.

KIRNEWS We used the same process for KIRNEWS. KIRNEWS was collected from eight news sources in total. However, it was more challenging, since most of the news sources that were listed on the overview websites[13] publish their news either in French or English. Only one news website was found that publishes in Kirundi. Thus, to solve this problem, the same seed expansion technique as described above was used to find three more news websites and four newspapers that publish in Kirundi, which was very time-consuming. We hope that this kind of seed expansion can in future be automated with the help of NLP technology for Kirundi.

Document Structure Each data sample for both KINNEWS and KIRNEWS consists of the news headline and the article's content. We separate the title from the content to make the annotation process easier, since the annotator may sometimes simply annotate the news based on its title without reading the whole article. In addition, the original source URLs are recorded with the extracted article, such that meta-information about dates or authors, or multi-modal content such as embedded images and captions can be retrieved if needed. Future NLP studies may also exploit this structure, e.g. for headline prediction or automatic summarization.

3.2 Annotation Process

The news we collected were initially given different categories by the publishers. The articles in KIN-NEWS and KIRNEWS were categorized in 48 and 26 different categories, respectively. However, many of these categories were related and to reduce the noisy samples, annotators agreed in grouping the related categories into one category and finally resulted into fourteen categories for KINNEWS and twelve categories for KIRNEWS in total (details in Appendix A).

In both datasets, each category was assigned with its own numerical labels (`label`) that range from 1 to 14 and English labels (`en_label`) to help those who do not understand Kinyarwanda and Kirundi to know what the article is related to. Moreover, Kinyarwanda labels (`kin_label`) for KINNEWS and Kirundi labels (`kir_label`) for KIRNEWS were also provided.

Then, based on these agreed categories, two annotators for each dataset who are all linguistic graduates and native speakers of each language, attentively revised each news article and annotate it based on its title and content. If they encountered one article that they would hesitate on its category, they annotated it as *neutral* and it receives a numerical label of 0, to be later removed from the final dataset to focus on clearly classifiable data.

3.3 Dataset Cleaning

For each language, we provide a cleaned version and a raw version. The cleaning is done in two stages: (1) removal of special characters, and (2) stopword removal. Nowadays, low-resource languages lack of language processing tools and resources (Baumann and Pierrehumbert, 2014; Muis et al., 2018). Kinyarwanda and Kirundi are also among those languages which do not have any language-specific processing tools (tokenizers, lemmatizers, stemmers, stopword filters, normalizers etc.).

Special Characters Removal The retrieved documents from the internet are often too noisy. To obtain high-quality and cleaned datasets, we remove the following non-alphanumerical characters ;.?/\|_#$%−<>()[]{}&*~'+-=^\n\r\t and URLs from the text. Note that removing punctuation might lead to losing sentence boundary information within the article. However, this does not hurt the performance of our models, because they were trained based on word-based features within each article. And since the raw data is provided, punctuation might be restored for other applications, for example for developing language-specific tokenizers.

---

[13]https://www.w3newspapers.com/burundi/ and http://www.abyznewslinks.com/burun.htm

Stopwords Since stopwords play an important role in semantic text preprocessing, we create first stopword lists for both Kinyarwanda and Kirundi languages. Using additional data from the Kinyarwanda Bible[14] and based on the sufficient knowledge the annotators have on both languages, we found that the words with two letters such as *"mu":"in"* and *"ku":"on"/"at"*, words with three letters such as *"uyu":"this"* and *"iyo":"that"*, and words with four letters such as *"muri":"in"* have high frequency, see Figure 1, but do not carry a significant role in training text classification models. Thus, these words and other similar words were used to create a list of 80 stopwords for Kinyarwanda and 59 stopwords for Kirundi, listed in Table 2. The words found in the stopword lists were then removed from the respective cleaned news datasets.
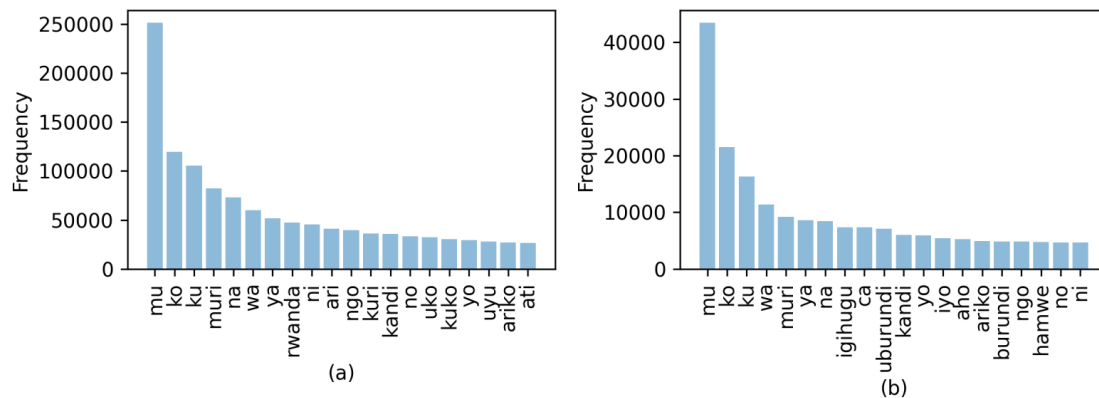

Figure 1: Top 20 tokens in (a) Kinyarwanda and (b) Kirundi data.

Kinyarwanda Stopwords	Kirundi Stopwords
aba, abo, aha, aho, ari, ati, aya, ayo, ba, baba, babo, bari, be, bo, bose, bw, bwa, bwo, by, bya, byo, cy, cya, cyo, hafi, ibi, ibyo, icyo, iki, imwe, iri, iyi, iyo, izi, izo, ka, ko, ku, kuri, kuva, kwa, maze, mu, muri, na, naho, nawe, ngo, ni, niba, nk, nka, no, nta, nuko, rero, rw, rwa, rwo, ry, rya, ubu, ubwo, uko, undi, uri, uwo, uyu, wa, wari, we, wo, ya, yabo, yari, ye, yo, yose, za, zo	aba, abo aho, ari,ata, ati, ayo, ba, bari, bo, bose, bw, bwa, bwo, ca, cane, co, de, ico, iryo, ivyo, iyo, izo,ko, ku, kuri, kuva, kw, maze, mu, muri, mw, na, naho, nayo, ngo, ni, nk, no, rero, rw, ry, rya, ubu, uko, uri, uwo, uyu, vy, vya, vyo, wa, wo, ya, yari, yo, yose, za, zo

Table 2: Stopword list for Kinyarwanda and Kirundi.

3.4 Dataset Statistics

The datasets contain a total of 21,268 and 4,612 news articles which are distributed across 14 and 12 categories for KINNEWS and KIRNEWS, respectively. As shown in Table 3 *Politics* related articles are the majority in both datasets, while *education* and *history* related articles are the minority in KINNEWS and KIRNEWS, respectively.

Table 4 shows that KINNEWS is divided in 17,014 articles for training and 4,254 articles for the test set, with an average of 302.9 words per article and approximately 370K unique words (when lowercased) for the raw version, and average of 246.5 words per article and approximately 300K unique words (when lowercased) for the cleaned version. KIRNEWS which is quite small compared to KINNEWS, is divided in 3,690 articles for train set and 922 articles for test set, with an average of 264.5 words per article

---

[14]https://bibiliya.com

KINNEWS		KIRNEWS	
en_label	# Articles	en_label	# Articles
Politics	4,908	Politics	1,642
Economy	4,418	Sport	1,400
Entertainment	3,229	Economy	388
Sport	3,104	Relationship	328
Health	1,914	Health	264
Relationship	800	Education	206
Religion	782	Entertainment	136
Technology	439	Technology	70
Culture	436	Environment	58
Tourism	378	Culture	52
History	256	Religion	48
Environment	213	History	20
Fashion	205		
Education	186		
Total	**21,268**	**Total**	**4,612**

Table 3: Dataset label distribution, ranked by frequency.

and approximately 86K unique words (when lowercased) for the raw version, and average of 210.2 words per article and approximately 63K unique words (when lowercased) for the cleaned version. The comparison with other high-resource news classification datasets for English, namely *AG News*[15] and *Reuters* (ApteMod) (Lewis, 1997) and low-resource news datasets (Setswana and Sepedi (Marivate et al., 2020)) shows that our datasets are composed of the longest news articles and have the largest vocabulary size, which indicates their suitability for training large deep learning models and NLP studies in general.

Resources	Dataset	Classes	Train Samples	Test Samples	Total	Average Words/Article	Unique Words
High-resource	AG's News	4	120,000	7,600	127,600	37.8	188,110
	Reuters	90	7,769	3,019	10,788	-	-
Low-resource	Setswana	10	-	-	219	-	1,561
	Sepedi	10	-	-	491	-	3,018
Low-resource (Ours)	KIRNEWS	12	3,690	922	4,612	264.5	86,023
	KIRNEWS*	12	3,690	922	4,612	210.2	63,143
	KINNEWS	14	17,014	4,254	21,268	302.9	370,036
	KINNEWS*	14	17,014	4,254	21,268	246.5	301,016

Table 4: Comparison with other news classification datasets. KIRNEWS* and KINNEWS* denote cleaned versions of KIRNEWS and KINNEWS, respectively.

An in-depth evaluation of the similarity between Kinyarwanda and Kirundi using the created datasets shows that they share 27,489 words of the vocabularies which stands for 32% of all unique words from KIRNEWS, using the raw version datasets and 22,841 vocabularies which stands for 36.2% of all unique words from KIRNEWS*, using the cleaned versions of the datasets.

---

[15]http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

4 Experiments

4.1 Word Embedding Training

Many African low-resource languages, including Kinyarwanda and Kirundi, do not enjoy the success of recent word embeddings available as pre-trained models such as GloVe (Pennington et al., 2014), BERT (Devlin et al., 2018), XLNet (Yang et al., 2019), or Fasttext (Grave et al., 2018) because these models were trained on higher-resource languages exclusively. Recent text classification approaches for low-resource languages rely on transfer learning approach that uses the features of resource-rich languages learned by pre-trained word embeddings to train low resource models. However, this technique might not be effective enough or even not be applicable when there is no parallel corpus of that resource-rich and resource-poor languages.

Since our datasets contain a reasonable amount of sentences, the features to train our neural-network-based models are obtained by training Word2Vec embeddings (Mikolov et al., 2013) from scratch. Word2Vec is trained using the `gensim` framework[16] with a window size of 5, ignoring all words with total frequency lower than 5, removing stopwords, special characters and URLs, and using skip-gram training algorithm with hierarchical softmax. We train two versions with different dimensions on each language, one with 50 dimensions (`W2V-Kin-50`) and other with 100 dimensions (`W2V-Kin-100`) for Kinyarwanda, and `W2V-Kir-50` and `W2V-Kir-100` for Kirundi.

4.2 Text Classification Task

Monolingual For a monolingual approach, we train and evaluate our baseline models using KINNEWS for Kinyarwanda and KIRNEWS for Kirundi separately. This means that we are using exclusively the data available for each task, ignoring the similarity of both languages.

Cross-lingual Cross-lingual transfer has been leveraged for many low-resource applications (Agić et al., 2015; Buys and Botha, 2016; Adams et al., 2017; Fang and Cohn, 2017; Cotterell and Duh, 2017). Most commonly, these approaches rely on machine translation and word alignments between resource-rich and low-resource languages. In this paper, however, we follow a simpler approach that exploits the fact that both languages are mutually intelligible, and does not require parallel or aligned resources. We train the baseline models using KINNEWS and embeddings learned from Kinyarwanda, and test them on KIRNEWS. This simulates the scenario if we did not have any training data for Kirundi. Alternatively, we train and test embedding-based models on KIRNEWS using the Kinyarwanda embeddings. This models a scenario where embeddings in a higher-resourced related language are available, and a small training set in the target language. We only investigate the transfer from Kinyarwanda to Kirundi and not the reverse, since our Kinyarwanda data is much larger than Kirundi data.[17]

4.3 Baseline Models

We perform benchmark experiments on the datasets using several different classic and neural approaches. In all experiments, we use the pre-processed (cleaned) version datasets, because the raw versions contain too much noise. The training set and validation set are split with a ratio of 9:1.

4.3.1 Classic Models

For all classic machine learning approaches, we use Term Frequency Inverse Document Frequency (TF-IDF) to get the values of unigram input features. We define the maximum number of features to be used depending on different train set and method. All of the below models are implemented with the help of the `scikit-learn` framework and use its default hyperparameters:

- **Multinomial Naive Bayes (MNB)**

- **Logistic Regression (LR)**

- **Support Vector Machine (SVM) with SGD**

---

[16]`https://radimrehurek.com/gensim/models/word2vec.html`

[17]We remove *tourism* and *fashion* related samples from KINNEWS to get a compatible training set for the KIRNEWS test set which does not contain articles from these two categories.

4.3.2 Neural Models

For neural models, we use the pre-trained embeddings as input and fine-tune them on the task (except for character-based models). This is to mimic approaching an arbitrary NLP task with little training data but with available word embeddings. The following neural models are implemented:

- **Character-level Convolutional Neural Networks (Char-CNN):** We use a small size Char-CNN model for text classification as proposed in (Zhang et al., 2015) with default hyperparameters, except that we removed the letters 'q' and 'x' from the alphabet list which are not in both Kinyarwanda and Kirundi languages. Thus, the alphabet used in our model consists of 68 characters instead of 70 characters from the original paper. The input feature length is also changed from 1,014 to 1,500 to capture most of the texts of interest, since our datasets have relatively long news articles. The special properties of Char-CNN that makes it a good choice for low-resource languages are that (1) it does not require any data preprocessing nor (2) the use of word embeddings which makes it more effective when processing very noisy data.

- **Convolutional Neural Network (CNN):** We use the CNN for sentence classification model proposed in (Kim, 2014) with default hyperparameters, except that we change the original feature maps of 100 to 150 and min-batch size of 50 to 32. The model is trained on two Word2Vec embeddings with different dimensions of 50 and 100, and using different epochs and number of features based on different train sets and embedding dimensions.

- **Bidirectional Gated Recurrent Unit (BiGRU):** We design a model that consists of 2-layer bidirectional GRU (Cho et al., 2014) followed by a softmax linear layer. It uses the dropout of 0.5 and batch size of 32. The dimension of hidden layers were set to either 256 or 128, it is trained on two Word2Vec embeddings with different dimensions of 50 and 100 similar to CNN, and different epochs and number of features were used according to different train sets and embedding dimensions similar to the previous models.

4.4 Results

4.4.1 Monolingual Text Classification

The experimental results for monolingual text classification are shown in Tables 5 and 6, respectively. In each table the benchmark results of the classic TFIDF-based models and the neural embedding-based models are grouped separately, and we highlight the result of the best model in each group.

Model	Embeddings	Features ($\times$1000)	Epochs	Accuracy (%)
MNB	TF-IDF	6	-	82.70
LR	TF-IDF	20	-	87.14
SVM	TF-IDF	90	-	**88.53**
Char-CNN	-	-	20	71.70
CNN	`W2V-Kin-50`	15	8	87.55
	`W2V-Kin-100`	40	4	87.54
BiGRU	`W2V-Kin-50`*	10	10	**88.65**
	`W2V-Kin-100`	10	6	88.29

Table 5: Benchmark results on the KINNEWS dataset. The * on `W2V-Kin-50` for BiGRU denotes that the dimension of hidden layers was set to 128 instead of 256.

As shown in Table 5 and 6, in the group of classic TFIDF-based models, SVM yields the best accuracy on both datasets compared to LR and MNB. It has high predictive power thanks to the hyperplane which can avoid the overfitting and separates the classes in very effective way. Another good property of SVM is that by using its supporting vectors, it can use relatively small amount of data to get a good prediction, which makes it to perform well on both KINNEWS and KIRNEWS. In this group, MNB performs the

Model	Embeddings	Features (×1000)	Epochs	Accuracy (%)
MNB	TF-IDF	4	-	82.67
LR	TF-IDF	6	-	86.13
SVM	TF-IDF	63	-	**90.14**
Char-CNN	-	-	20	69.23
CNN	W2V-Kir-50	40	8	85.75
	W2V-Kir-100	35	4	**88.01**
BiGRU	W2V-Kir-50	10	12	85.86
	W2V-Kir-100	10	6	86.61

Table 6: Benchmark results on the KIRNEWS dataset.

worst on both datasets, however, it was able to give relatively good results by requiring fewer features compared to other methods.

In the group of neural embedding-based models, the performance is based on the type of data, where BiGRU perform the best on KINNEWS which is larger than KIRNEWS dataset, while CNN perform the best on KIRNEWS. A possible reason might be that BiGRU needs a larger amount of data to perform better than the CNN. The Char-CNN performs the worst on both datasets, likely because of the limited computation power of the compute resources used in the experiments. Because of this we had to limit the input feature length to 1500 while the average length of the news article in each dataset was far greater than that. Thus it is an open challenge to further work that can be able to use higher length of input features to achieve better results. It might also be interpreted as a pointer towards the general "pre-train and fine-tune" regime, which places classification models in an initial representation space that reflects word relations in the input.

Train set	Test Set	Model	Embeddings	Features (×1000)	Epochs	Accuracy (%)
KINNEWS	KIRNEWS	MNB	TF-IDF	6	-	**73.46**
		LR	TF-IDF	6	-	68.26
		SVM	TF-IDF	35	-	72.70
KINNEWS	KIRNEWS	Char-CNN	-	-	6	49.60
		CNN	W2V-Kin-50	30	5	60.64
			W2V-Kin-100	20	4	61.72
		BiGRU	W2V-Kin-50	10	7	**67.54**
			W2V-Kin-100*	10	7	65.06
KIRNEWS	KIRNEWS	CNN	W2V-Kin-50	40	12	85.75
			W2V-Kin-100	35	8	**88.01**
		BiGRU	W2V-Kin-50	10	12	83.38
			W2V-Kin-100	10	10	86.61

Table 7: Benchmark results for cross-lingual approaches. The * on W2V-Kin-100 for BiGRU denotes that the dimension of hidden layers was set to 128 instead of 256.

4.4.2 Cross-lingual Text Classification

The results on cross-lingual approaches in Table 7 show that in the group of machine learning models, MNB surprisingly performs relatively better compared to SVM and LR. The reason might be that MNB is a generative model while the rest are discriminative models.

Similar to the monolingual experiments, BiGRU performs the best when trained on KINNEWS while CNN performs the best when trained on KIRNEWS. Interestingly, the Char-CNN suffers much more from the transfer, which illustrates that the embeddings reflect the similarities of both languages on an

abstract level that allow transfer much better than low-level features trained from scratch.

When trained neural models on KINNEWS and testing them on KIRNEWS, they do not give satisfactory results compared to when trained and tested on only KIRNEWS. This is not surprising, since they were both trained on Kinyarwanda word embeddings from the same domain which includes the word vectors of many similar words from Kirundi. Nevertheless, this shows that with pretrained Kinyarwanda word embeddings, which are easier to obtain in high quality since the data retrieval for Kinyarwanda is much easier, can be effectively used in training Kirundi text classification models in a zero-shot scenario without any labeled or unlabeled data for Kirundi.

Based on the performed experiments, the results for our examplary languages Kinywarwanda and Kirundi show that the text classification based on cross-lingual transfer between mutually intelligible low-resource languages is possible, without creating any words alignments or any parallel translation dataset between those languages. What is merely required is to get sufficient data of one language to train the word embeddings.

Analysing the classification errors of highest-scoring cross-lingual models, we find that the MNB model characteristically places many articles about "education" in the category of "politics", while the neural models are more accurate in this distinction. All models tend to confuse "relationship" articles with the "politics" category, which might be due to an overlap in common vocabulary focused on interactions of people. The most accurate classification is generally obtained for sports articles. Complete confusion matrices are displayed in Appendix B.

5 Conclusion and Future Work

In this paper, we built the first news text classification benchmark for Kinyarwanda and Kirundi, two low-resource Bantu languages. We described the data collection process, provided guidelines for data cleaning, and evaluated classic text classification models as initial baselines. We found fairly strong cross-lingual generation of embedding models trained on the resource-richer Kinyarwanda to the lower-resource Kirundi due to their mutual intelligibility. This gives hope for future studies of languages from the Rwanda-Rundi language family, which have not been studied in NLP at all, and would otherwise classify as "Left-Behind" according to (Joshi et al., 2020), or analogously of other extremely low-resource languages with a slightly higher-resource "sibling".

Future studies on the new dataset will investigate (1) contextualized embeddings, e.g. BERT, (2) subword modeling. Furthermore, the dataset may get enriched with other linguistic annotations, such as named entities, and serve as resource for other NLP tasks than text classification.

Acknowledgements

References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947.

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, July. Association for Computational Linguistics.

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the bible: Learning pos taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272.

Jacob Andreas. 2020. Good-enough compositional data augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7556–7566, Online, July. Association for Computational Linguistics.

Peter Baumann and Janet B Pierrehumbert. 2014. Using resource-rich languages to improve morphological analysis of under-resourced languages. In *LREC*, pages 3355–3359.

Jan Buys and Jan A Botha. 2016. Cross-lingual morphological tagging for low-resource languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1954–1964.

Ronald Cardenas, Ying Lin, Heng Ji, and Jonathan May. 2019. A grounded unsupervised universal part-of-speech tagger for low-resource languages. 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 2428–2439, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 91–96.

Siddharth Dalmia, Xinjian Li, Florian Metze, and Alan W Black. 2018. Domain robust feature extraction for rapid low resource asr development. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 258–265. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Long Duong, Trevor Cohn, Karin Verspoor, Steven Bird, and Paul Cook. 2014. What can we get from 1000 tokens? a case study of multilingual pos tagging for resource-poor languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 886–897.

Chris Chinenye Emezue and Femi Pancrace Bonaventure Dossou. 2020. FFR v1.1: Fon-French neural machine translation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 83–87, Seattle, USA, July. Association for Computational Linguistics.

Meng Fang and Trevor Cohn. 2016. Learning when to trust distant supervision: An application to low-resource POS tagging using cross-lingual projection. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 178–186, Berlin, Germany, August. Association for Computational Linguistics.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 587–593.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddee Hassan Muhammad, Salomon Kabongo, Salomey Osei, Sackey Freshia, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa, Mofe Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Jane Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkabir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Emezue, Bonaventure Dossou, Blessing Sibanda, Blessing Itoro Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020a. Participatory research for low-resourced machine translation: A case study in african languages. In *EMNLP Findings*.

∀, Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020b. Masakhane – machine translation for africa. In *"AfricaNLP" Workshop at the 8th International Conference on Learning Representations (ICLR)*, Virtual.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147.

Dan Garrette, Jason Mielens, and Jason Baldridge. 2013. Real-world semi-supervised learning of pos-taggers for low-resource languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 583–592.

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

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

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, July. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

David D Lewis. 1997. Reuters-21578 text categorization collection data set.

Vukosi Marivate, Tshephisho Sefara, and Abiodun Modupe. 2020. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. In *Proceedings of the first workshop on Resources for African Indigenous Languages*, pages 15–20, Marseille, France, May. European Language Resources Association (ELRA).

Jason Mielens, Liang Sun, and Jason Baldridge. 2015. Parse imputation for dependency annotations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1385–1394.

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

Jackson Muhirwe. 2009. Morphological analysis of tone marked kinyarwanda text. In *International Workshop on Finite-State Methods and Natural Language Processing*, pages 48–55. Springer.

Aldrian Obaja Muis, Naoki Otani, Nidhi Vyas, Ruochen Xu, Yiming Yang, Teruko Mitamura, and Eduard Hovy. 2018. Low-resource cross-lingual event type detection via distant supervision with minimal effort. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 70–82.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Shruti Rijhwani, Shuyan Zhou, Graham Neubig, and Jaime Carbonell. 2020. Soft gazetteers for low-resource named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8118–8123, Online, July. Association for Computational Linguistics.

Liang Sun, Jason Mielens, and Jason Baldridge. 2014. Parsing low-resource languages using gibbs sampling for pcfgs with latent annotations. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 290–300.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Appendices

A Annotation Rules

Shared Labels			KINNEWS	KIRNEWS	
Label	En_Label	Kin_Label	Initial_Category	Kir_Label	Initial_Category
1	politics	politiki	politiki (politics) imiyoborere (governance) diyasipora (diaspora) amatora (elections) ubutabera (justice) umutekano (security) kwibuka (genocide memorial)	poritike	politike (politics) imibano (int'l relations) diyasipora (diaspora) imigenderanire(relations) ubutungane (justice) umutekano (security) impunzi (refugee news)
2	sport	imikino	imikino (sport) siporo (sport) football basketball volleyball amagare (biking) handball ngororangingo (stretching) karate	inkino	inkino (sport)
3	economy	ubukungu	ubukungu(economy) ubucuruzi (business) ubuhinzi (agriculture) iterambere (development) ubworozi (farming) ubwiteganyirize (insurance) imigabane (stock market) ivunjisha (exchange) ishoramari (investment)	ubutunzi	ubutunzi (economy) ubudandaji (business) uburimyi(agriculture) iterambere (development) ubworozi (farming) indanga (business)
4	health	ubuzima	ubuzima (health) indwara (disease)	amagara	amagara (health) ubuzima (health)
5	entertainment	imyidagaduro	imyidagaduro (entertainment) showbiz indirimbo (songs) muzika (music) sinema (cinema) ibirori (party) urwenya (comedy)	kwidagadura	kwidagadura (entertainment)

6	history	amateka	amateka (history) mu mateka (in history)	akahise	akahise (history)
7	techn-ology	ikorana-buhanga	ikoranabuhanga (technology)	ubuhinga	ubuhinga (technology)
8	tourism	ubuke-rarugendo	ubukerarugendo (tourism)	-	-
9	culture	umuco	umuco (culture)	akaranga	imicokama (culture)
10	fashion	imideli	imideli (fashion)	-	-
11	religion	iyobo-kamana	iyobokamana (religion)	ukwemera	ukwemera (religion) ivy'Imana (religion)
12	envir-onment	ibidukikije	ibidukikije (environment)	ibidukikije	ibidukikije (environment)
13	education	uburezi	uburezi (education) ubumenyi (knowl-edge) mu mashuli (schools)	indero	indero (education) inyigisho (lessons)
14	relat-ionship	urukundo	urukundo (relation-ship) gushaka umukunzi (dating) ubukwe no gusaba (weddings)	urukundo	urukundo (relationship)

Table 8: Annotation category combinations based rules.

B   Confusion Matrices of our best Cross-Lingual Models

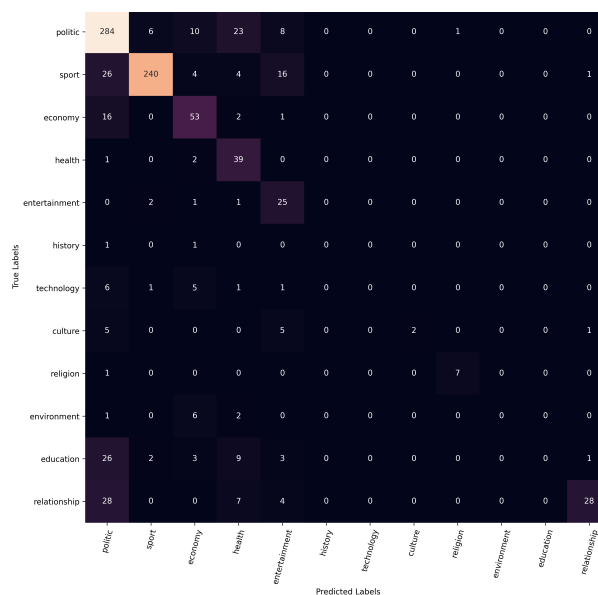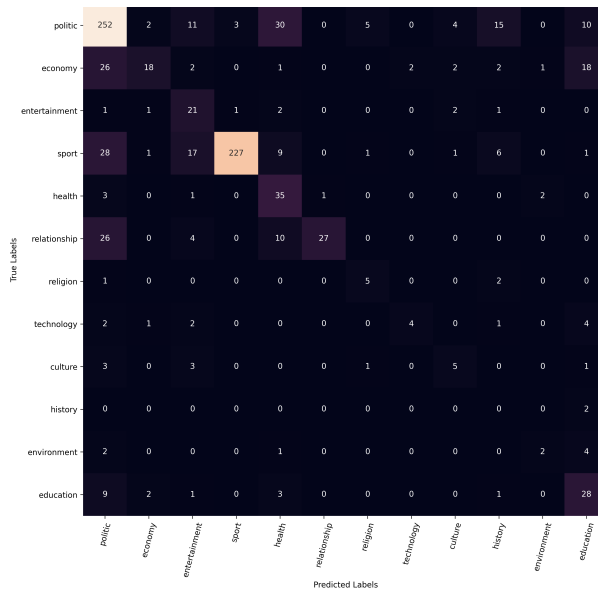Figure 2: Confusion matrices of TF-IDF+MNB.

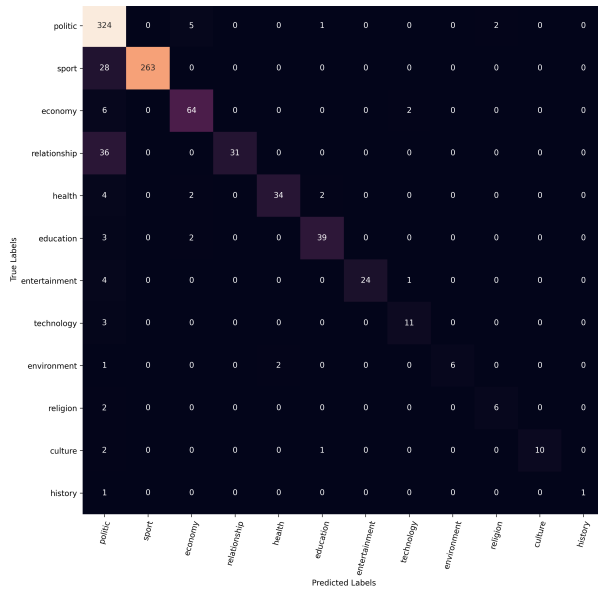Figure 3: Confusion matrices of `W2V-Kin-50`+BiGRU.

Figure 4: Confusion matrices of `W2V-Kin-50`+CNN.