Sefamerve at SemEval-2023 Task 12: Semantic Evaluation of Rarely Studied Languages

Selman Delil
Sefamerve R&D Center
Istanbul, Turkey
selman.delil@sefamerve.com
selmandelil@gmail.com

Birol Kuyumcu
Sefamerve R&D Center
Istanbul, Turkey
birol.kuyumcu@sefamerve.com
birolkuyumcu@gmail.com

Abstract

This paper describes our contribution to SemEval-23 Shared Task 12: ArfiSenti. The task consists of several sentiment classification subtasks for rarely studied African languages to predict positive, negative, or neutral classes of a given Twitter dataset. In our system we utilized three different models; FastText, MultiLang Transformers, and Language-Specific Transformers to find the best working model for the classification challenge. We experimented with mentioned models and mostly reached the best prediction scores using the Language Specific Transformers. Our best-submitted result was ranked 3rd among submissions for the Amharic language, obtaining an F1 score of 0.702 behind the second-ranked system.

1 Introduction

Sentiment analysis is a task in natural language processing (NLP) that involves identifying and extracting the sentiment expressed in a text. It has many practical applications, such as in business and marketing analysis, product review analysis, and social media monitoring (Hutto and Gilbert, 2014). While there has been considerable research in sentiment analysis for widely spoken languages such as English, there has been relatively little work done for African languages. This is mainly due to the scarcity of labeled data and language resources for African languages, which hinders the development of effective sentiment analysis models for these languages (Muhammad et al., 2023a).

The AfriSenti task (Muhammad et al., 2023b,a) motivated the need for curated datasets for African languages to enable various AI applications such as sentiment analysis, machine translation, and hateful content detection. Due to the lack of such datasets for most African languages, the importance of research to determine the suitability of current NLP techniques and the development of novel techniques for these languages are critical.

The task aims to evaluate sentiment analysis methods for African languages, specifically for Twitter data. The AfriSenti-SemEval Shared Task 12 involves sentiment classification of tweets in 14 African languages, with three sub-tasks including monolingual sentiment classification, multilingual sentiment classification, and zero-shot sentiment classification. Participants can select one or more sub-tasks and participate in any number of languages. Task A has 15 tracks with training data in different languages, while Task B has only one track with 12 languages. Task C involves determining the sentiment of unlabelled tweets in two African languages using any available training datasets (Muhammad et al., 2023a, 2022).

In this paper, we describe our approach to the AfriSenti-SemEval Shared Task 12, where we experimented with three different models to find the best working model for sentiment classification in African languages. We adopted the FastText embedding model, the AfriBERTa transformer-based language model, and language-specific transformer models. Through our experimentation, we aimed to determine the most effective model for sentiment classification on the provided dataset. Our best results were obtained using the language-specific transformer models, which achieved an F1 score of 0.702 for the Amharic language, ranking 3rd among all submissions.

The paper is organized in the following way. In section 2, system overview, we explore the best models for the classification tasks on the provided dataset by experimenting with different model approaches. In section 2.1, we start experimenting FastText word embedding model and adapt this model for all three sub-tasks in the competition dataset. In 2.2, we adapt the transformers-based language model AfriBERTa, an improvement over the original BERT model, trained on a large corpus of multilingual text. Finally, in section 2.3, we use the language-specific models that were developed

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 512–516
July 13-14, 2023 ©2023 Association for Computational Linguistics

specifically for African languages. In the results and conclusion sections, we provide details about the results of each model used, including their optimization, and evaluation metrics as well as our conclusions regarding the study.

2 System Overview

In this study, we explore the best models for the classification tasks on the provided dataset by experimenting with different model approaches. First, we used the word embedding model FastText which takes into account the internal structure of words while learning word representations that is particularly beneficial for morphologically rich languages (Bojanowski et al., 2016; Joulin et al., 2016).

Second, we adapt transformers-based language model AfriBERTa to the problem which is an improvement over the original BERT model and trained on a large corpus of multilingual text. AfriBERTa is also based on the RoBERTa architecture and was trained on a large corpus of text in African languages (Ogueji et al., 2021).

Finally, we used the language-specific models that were developed specifically for African languages. We experimented with mentioned models and mostly reached the best prediction scores using the Language Specific Transformers. System details are provided in the upcoming sections.

2.1 FastText

The FastText embedding model utilizes a distinct approach to word representation as compared to other models, such as word2vec. Unlike word2vec, which considers each word as the smallest unit, FastText regards a word as composed of n-grams of characters, where n can vary from one to the word's length. The advantage of this method is that by retaining the word vectors as character n-grams, it can generate vector representations for words that are not present in the dictionary (Bojanowski et al., 2016; Joulin et al., 2016).

We adapt FastText for all 3 subtasks in the competition dataset. In the first stage of the competition, training data with label and development data non-labeled provided to build the best model. We trained only with training data by splitting it into training and testing sets with a ratio of 80% and 20%, respectively. In the second stage where development dataset labels are provided, we add newly available training data to the model to cover new sample space.

Optimizing hyperparameters is crucial for effective model building, but manually searching for them can be difficult and time-consuming due to parameter interdependence and variable dataset effects. FastText's Autotune feature can automatically identify the best hyperparameters for your specific dataset (Bojanowski et al., 2016; Joulin et al., 2016). We utilized FastText autotune automatic adjustment feature to optimize the hyperparameters of our models. The main parameters subjected to the optimization are defined below. Table-1 presents the parameters utilized for each language.

- dim : size of word vectors
- epoch : number of epoch
- wordNgrams : max length of word ngram
- bucket : number of buckets
- minn : min length of char ngram
- maxn : max length of char ngram

2.2 Multilanguage Transformers

Fine-tuning work was performed based on some models based on xlm-roberta, but the loss did not decrease during training, and therefore good results could not be achieved. A study based on "AfriBERTa", which is directly trained for African languages, was more successful. AfriBERTa is a BERT-based language model that has been specifically trained on a total of 11 African languages. Furthermore, AfriBERTa underwent evaluation for Named Entity Recognition (NER) and text classification across most of these languages. Notably, it outperformed mBERT and XLM-R on several languages and demonstrated overall high competitiveness (Ogueji et al., 2021).

The dataset was divided into 80% training and 20% validation for each language. The AfriBERTa model was used for fine-tuning with 5 epochs, and the best model was selected based on the F1 score. Model parameters are selected as follows:

LR = 2e-5
EPOCHS = 5
BATCH_SIZE = 8
MODEL = "castorini/afriberta_base"

For Task B, a model based on AfriBERTa was fine-tuned and evaluated on a multilingual dataset. Interestingly, it performed only slightly worse than

Language	dim	epoch	wordNgrams	bucket	minn	maxn
Amharic	189	49	3	6410340	0	0
Algerian Arabic	43	11	1	10000000	2	5
Hausa	316	100	5	2905495	3	6
Igbo	58	100	5	10000000	3	6
Kinyarwanda	3	64	3	236257	3	6
Moroccan Arabic/Darija	136	56	3	4403363	2	5
multilingual	187	100	2	196144	3	6
Nigerian Pidgin	181	100	4	1731320	0	0
Mozambican Portuguese	107	79	5	2466614	2	5
Swahili	51	100	5	184083	0	0
Xitsonga	273	11	5	513644	3	6
Twi	462	40	5	352988	3	6
Yoruba	14	100	1	148489	3	6

Table 1: FastText Autotuned Parameters

Language	Model Used
Amharic (Adelani, 2022d)	Davlan/xlm-roberta-base-finetuned-amharic
Moroccan Arabic/Darija (Gaanoun et al., 2023)	SI2M-Lab/DarijaBERT
Hausa (Adelani, 2022e)	Davlan/xlm-roberta-base-finetuned-hausa
Igbo (Adelani, 2022f)	Davlan/xlm-roberta-base-finetuned-igbo
Yoruba (Adelani, 2022c)	Davlan/bert-base-multilingual-cased-finetuned-yoruba
Nigerian Pidgin (Adelani, 2022b)	Davlan/bert-base-multilingual-cased-finetuned-naija
Algerian Arabic (Abdaoui et al., 2021)	alger-ia/dziribert
Mozambican Portuguese (Souza et al., 2020)	neuralmind/bert-base-portuguese-cased
Swahili (Adelani, 2022g)	Davlan/xlm-roberta-base-finetuned-swahili
Kinyarwanda (Adelani, 2022a)	Davlan/bert-base-multilingual-cased-finetuned-kinyarwanda
Twi (Adelani, 2021)	Ghana-NLP/abena-base-akuapem-twi-cased

Table 2: Language specific models used in the study

Language	F1:Fasttext	F1:AfriBERTa	F1:Lang. Specific	Ranking
Amharic	0.442	**0.702**		**3**
Moroccan Arabic / Darija	**0.573**	0.409	0.525	24
Hausa	0.762	0.779	**0.786**	20
Igbo	**0.786**	0.780	0.732	27
Yoruba	0.713	**0.741**	0.253	32
Nigerian Pidgin	0.618	0.637	**0.637**	27
Algerian Arabic	0.626	0.455	**0.717**	**6**
Mozambican Portuguese	0.617	0.589	**0.679**	16
Swahili	0.525	0.629	**0.639**	**5**
Kinyarwanda	0.604	0.627	**0.631**	21
Twi	0.641	0.612	**0.657**	10
Xitsonga (Mozambique Dialect)	**0.521**	0.483		25

Table 3: Submission results and F1 scores for Task-A

the model trained with Fasttext. For Task C, we utilize the same multi-language model for the zero-shot learning challenge. The model that fell behind FastText's results in Task B has surpassed FastText in zero-shot predictions.

2.3 Language Specific Transformers

Since the languages we work on have few resources, if available, the language-specific fine-tuned versions have been used for fine-tuning on each language dataset provided by the task and then retrained for Task A.

The dataset was divided into 80% training and 20% validation for each language. The model was used for fine-tuning with 5 epochs, and the best model was selected based on the F1 score. Table-2 provides a list of language-specific models utilized in this study for African languages.

LR = 2e-5
EPOCHS = 5
BATCH_SIZE = 8
OPTIMIZER = AdamW

3 Results

In Table-3, you can see the F1 scores achieved for each model during the final testing phase. Accordingly, FastText gave the best results for a total of 3 languages in the Moroccan Arabic / Darija, Igbo, and Xitsonga (Mozambique Dialect) languages. AfriBERTa provided the best results for a total of 2 languages in the Amharic and Yoruba languages, and language-specific transformers models achieved the best results for a total of 7 languages in the Hausa, Nigerian Pidgin, Algerian Arabic, Mozambican Portuguese, Swahili, Kinyarwanda, and Twi languages. Language-specific transformers models naturally gave the best results.

Method	Result
FastText	0.603
AfriBERTa	0.634

Table 4: Results for Task-B

Method	Tigrinya	Oromo
FastText	0.304	0.345
AfriBERTa	0.342	0.394

Table 5: Results for Task-C

For Task B, AfriBERTa outperformed FastText with an F1 score of 0.63 (Table-4, while for Task C, both FastText and AfriBERTa models yielded very low scores (Table-5).

The code used in this study is available on GitHub [1]. Researchers interested in reproducing or building upon our work can access the code and associated resources at this location.

4 Conclusion

We experimented with FastText, transformers-based multilanguage models (AfriBert), and language-specific models and mostly reached the best prediction scores using the language-specific transformers. Our models do not use any of the external datasets or automatic linguistic annotations, such as named entity tags. Overall, we showed that language-based systems can produce better re-

[1] https://github.com/birolkuyumcu/AfriSenti-SemEval-2023

sults to address rarely studied language tasks. Our best-submitted result was ranked 3rd among submissions, obtaining an F1 score of 0.702 behind the second-ranked system.

References

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. Dziribert: a pre-trained language model for the algerian dialect. *arXiv preprint arXiv:2109.12346*.

David Adelani. 2021. abena-base-akuapem-twi-cased. https://huggingface.co/Ghana-NLP/abena-base-akuapem-twi-cased. Accessed: February 26, 2023.

David Adelani. 2022a. bert-base-multilingual-cased-finetuned-kinyarwanda. https://huggingface.co/Davlan/bert-base-multilingual-cased-finetuned-kinyarwanda. Accessed: February 26, 2023.

David Adelani. 2022b. bert-base-multilingual-cased-finetuned-naija. https://huggingface.co/Davlan/bert-base-multilingual-cased-finetuned-naija. Accessed: February 26, 2023.

David Adelani. 2022c. bert-base-multilingual-cased-finetuned-yoruba. https://huggingface.co/Davlan/bert-base-multilingual-cased-finetuned-yoruba. Accessed: February 26, 2023.

David Adelani. 2022d. xlm-roberta-base-finetuned-amharic. https://huggingface.co/Davlan/xlm-roberta-base-finetuned-amharic. Accessed: February 26, 2023.

David Adelani. 2022e. xlm-roberta-base-finetuned-hausa. https://huggingface.co/Davlan/xlm-roberta-base-finetuned-hausa. Accessed: February 26, 2023.

David Adelani. 2022f. xlm-roberta-base-finetuned-igbo. https://huggingface.co/Davlan/xlm-roberta-base-finetuned-igbo. Accessed: February 26, 2023.

David Adelani. 2022g. xlm-roberta-base-finetuned-swahili. https://huggingface.co/Davlan/xlm-roberta-base-finetuned-swahili. Accessed: February 26, 2023.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2023. Darijabert: a step forward in nlp for the written moroccan dialect.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, 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. 2023a. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif M. Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023b. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, David Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Shehu Bello Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahuddeen Abdullahi, Anuoluwapo Aremu, Alipio Jeorge, and Pavel Brazdil. 2022. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.